Author's response to Reviewer/Editor critique

Overview of changes
I have uploaded the revisions for my paper "The acoustic characteristics of Swedish vowels". I thank both the associate editor and the reviewers for insightful comments on my work. I have revised the manuscript following reviewers' suggestions. I have also followed the associate editor's suggestion to exclude Study 2 from the article. This means that the entire section labelled 'Study 2: Vowel category movements over time – from SweDia to SwehVd' has been removed, and additional edits have been made to Introduction and General discussion.

I have further refitted the GAMMs with a difference smooth instead of separate smooths. This makes visualizations and conclusions clearer. Importantly, this change has not altered the qualitative interpretation of results. I have also removed the GAMM modeled on duration since duration is a one-point measure which would counter the purpose of a GAMM and increase the risk of over-fitting.

While under review, I managed to collect data from an additional 4 male talkers, thus completing the final release of the SwehVd database. Data from these additional talkers is now included in the paper, and a clarification is made in the Methods section.

Finally, I discovered a coding mistake in the calculations of the separability index including F1-F2-F3. This has now been adjusted, resulting in an increase of the category separability in the already established direction, leaving all contrasts in both the rounding and the long-short vowel pairs evaluations with statistically significant improvements. I also decided to include duration as a cue in the calculations of the category separability index for the long-short vowel pairs evaluation. To increase comparability across the two types of separability evaluations (rounding and quantity), I have added panels showing the increase in separability by the inclusion of additional cues relative to F1-F2 baseline in the category separability plots. For visualization purposes, I have moved the F1-F3 and F2-F3 evaluations to the Supplementary information along with new combinations with duration – F1-duration, F2-duration (SI section Category separability indices for additional cue combinations). Adjustments to this end were made to the Methods section as well.

All changes are unlined in the pdf and docx files.

Response to reviewers' comments
I respond in blue below.

Associate editor
[summary omitted]
The combination of static and dynamic measures is very welcome and the analyses carried out were meticulously described and presented. The paper should therefore be publishable following the minor revisions suggested by the reviewers.

I thank the associate editor for the encouragement and helpful feedback below.

Below I add two main points that are comments rather than requests for action:

- While I appreciate how you acknowledged the lack of systematic listening in the limitations of the study towards the end of the paper, I do feel it is a shame that this was not carried out, and no reason was offered for why this was not done. Even impressionistic listening by one trained phonetician would have been good, but presented more systematically than the single comment in lines 753 to 760 regarding potential differences between [iː] and [yː] which support "the presence of a buzzing sound, similar to [ɨː]": I'm not entirely sure what 'buzzing' refers to, but would have liked to see a more extended presentation of impressionistic analyses, to parallel the acoustic findings. While one does not of course expect a one-to-one correspondence between auditory judgement and acoustic realisation, it would be really helpful to the reader to find out if the rich system of vowels are all distinguishable auditorily, and/or how their auditory impression compares with the IPA category (symbol) that is assigned to them. It would also enrich the conversation around the possible centralization of [iː] - [yː], the potential loss of lip-rounding in [yː], and lowering of [ɛː]. I do NOT wish to make this a required revision, but find it frustrating that so few studies making conclusions about positions and movements in the vowel space (e.g. fronting/backing, raising/lowering, etc.) actually include impressionistic judgements of this vowel space. I very much agree, and including systematic listening is something that I plan to do in future studies. As my overarching research goal is to understand listeners' vowel perception, the approach I am planning for is to ask naïve listeners rather than trained phoneticians to categorize vowels. While any form of phonetic training is in its own right valuable, it can, however, create unwanted biases and be different from how naïve listeners perceive vowels. To conduct a perception study, however, means that I will need to develop a task that is intuitive enough for naïve listeners, and further recruit and collect data from a sufficient number of listeners. In other work, I have done this on a much simpler vowel system, the US English vowel space (Persson, Barreda & Jaeger, 2024), and it took me about a year to plan and conduct the study, and another year to write it up. Hence, the primary reason I have not done this here, is time and funding. I have, nevertheless, made a clarification as to the 'buzzing' sound, in both Introduction and General discussion. Since I now have removed Study 2 from the paper, the discussion around possible movements in the space is substantially reduced, however, I have developed the comment on systematic listening in the General discussion to include the points above.

- I am not totally convinced that study 2 is required as you have enough very rich data and discussion from study 1, or that the comparison being made can actually tell us something meaningful about the change in the Swedish vowel system between 1999 and the present day given that: 1) the analyses are based only on 8 speakers and 2) SweDia data elicitation was of real words in various phonetic contexts while only hVd was used for SwehVd. This necessitated more static measures and constrained conclusions. Do keep it if you feel strongly about it, but only future work that is also sociolinguistic in nature will be able to

corroborate these initial findings and build on them. Thanks for this comment. I have decided to follow your advice and remove Study 2 from the article. My aim is to publish it as a separate article after additional corrections and more stringent analysis. This means that I have made changes to the Introduction and the General discussion, besides the removal of the entire section 3.

Minor comments:

- Fig 1 caption: "Vowels that mismatched intended label are excluded (1.33% of all recordings)": I am not sure what you mean by this I agree that this is a vague statement, I have adjusted the caption to make this point clearer.

- Fig 3 is hard to process due to there being too many mini figures within it, making it hard to discern patterns. The same applies to other figures throughout. I appreciate there is little you can do about it due to the rich nature of the data being presented, but if you can either increase the resolution of the figures and/or the size/font wherever possible that would be helpful. I have adjusted the font size in all figures throughout the manuscript. Unfortunately, I was not able to increase the resolution for submission of revised manuscript due to file size limit.

- [Y] is curious: it seems to be more peripheral and more rounded than its tense counterpart. Some comment on that would be useful. The literature is full of work suggesting that short/lax vowels are more centralised than their tense/long counterparts. Thank you for commenting this. I have briefly expanded on this in the Results section as well as the General discussion.

- Note 10: "An effect of gender was found for the high back vowels predicting F1 for the long and short vowels (both $ps < .022$), which suggests that the normalization approach likely reduced some talker-specific differences related to anatomical differences, but not all": well the obvious thing here is possible sociolinguistic differences, not just the fact that normalisation might not reduce anatomical differences Thanks. I have now added a comment about possible sociolinguistic differences and moved this into the main text (removing the footnote).

- 521: as statistically indistinguishable performance was found for the F1-F2 model: I think you mean distinguishable, no? Thanks for catching this. What I meant was that statistically indistinguishable performance was found for the F1-F2-F3 model relative to the F1-F2 model (so, no statistically significant changes when F3 was included). This is now adjusted given the updated separability index.

Reviewer 2 also meant to add the following reference as an important addition to Table 1, but had submitted before they realised it was missing so emailed it instead:
Gross J., Boyd S., Leinonen T., Walker J.A.: A tale of two cities (and one vowel): Sociolinguistic variation in Swedish. Language Variation and Change. 2016;28(2):225-

247. doi:10.1017/S0954394516000065 Thanks. It is now added to the table and referenced in text.

Reviewer #1
[summary omitted]
They thus make a valuable contribution to research on Swedish-language phonetics and phonology as well as the application of acoustic analysis. The study is well conceived, the arguments are well presented, and the topic is very suitable for Phonetica. I believe that with the appropriate revisions, this manuscript could become a point of reference for those investigating Swedish vowels. However, I have identified a number of minor revisions that need to be addressed prior to the manuscript's publication.

I am glad reviewer 1 feels that there is merit to this work, and I thank reviewer 1 for helpful feedback and encouragement.

- In line 16, the result in the abstract was too general. It should be more specific by highlighting the role of each acoustic analysis and which one presented the most exhaustive description and captured the spectral acoustics of Swedish vowels (see line 666-669 as an example). Additionally, another line should be added regarding the role of the spectral (formant frequencies) and temporal cues (duration) in vowel distinctions in Swedish vowels. Abstracts in Phonetica are limited to 200 words; therefore, the author still has space to include more information and details. I agree, the abstract is now revised.

- In line 27, the author should define/explain briefly the meaning of "static point," how it is usually measured, and why it is widely used. Thanks. This is now added.

- In lines 28-29, it is better to order the citations according to their publishing year (e.g., Joos, 1948; Ladefoged & Broadbent, 1957; Peterson, 1961; Nearey & Assmann, 1986). I have set references to follow in alphabetical order, in line with De Gruyter Mouton Journal Style Sheet.

- In line 44, please correct the text citation of "Leinonen." It is written as "T. Leinonen" throughout the manuscript. The same can be said regarding "K. Leinonen" in line 48. Done.

- In lines 37-38, "recorded by 44 male and female talkers of Swedish, the SwehVd database"' is unnecessary in the introduction because such details are already mentioned in the methodology section. I agree, I have removed the "recorded by…" but kept the reference.

- In lines 67-68, the author stated, "The use of an hVd database minimizes coarticulatory effects from the surrounding phonetic context." It is better to explain briefly how. Thanks. I have revised the sentence with a brief explanation.

- In line 97, the author cited R Core Team and RStudio Team, but forgot to add them in the reference list. Please do so. Done
- In line 159, the author needs to define F0 when it is first mentioned. This is now done in the Introduction, as suggested below.

- In lines 200-201, the author cited PRAAT as (Boersma & Weenink, 2022). In fact, the best way to cite PRAAT is to refer to PRAAT software from the first day of its creation until the day of the last update. For more details, open PRAAT, then click on the "Help" button, then the "About PRAAT" button to see the proper citation (e.g., Boersma & Weenink, 1992-2024). Please do not forget to correct this in the reference list. Done

- In line 201, the author stated that the first three formants were extracted automatically using the Burg algorithm. Using this automatic method might result in some errors in extracting formants, so the author should write another line about the additional method used to ensure the extracted data's accuracy and that no error had occurred. Thanks. The procedure for identifying potential measurement errors was described only in the second study. I have now moved this part to the methods section and revised the wording.

- In lines 201-202, the author stated that the first three formants (F1, F2, and F3) were extracted from five points (20, 35, 50, 65, 80%). It would be better to justify why they extracted only these five points, not more or less. Thanks. I have added a motivation to this end.

- Additionally, in lines 202-203, the author mentioned that F0 was extracted across the entire vowel segment. It would be better to justify why they extracted F0 across the entire vowel segment, not from five points (20, 35, 50, 65, 80%), as done for F1, F2, and F3. Thanks. I have edited the introduction (in accordance with the following bullet point). I have also expanded somewhat on this question in the General discussion (please see my comment to the first bullet point under 'Other Comments' below).

- In lines 232-235, "While F0…and is therefore reported." The role of F0 should be mentioned in the introduction; therefore, this sentence should be moved up in the introduction to line 33. If the author does so, please ignore my comment regarding line 159. I have moved this up to the Introduction.

- In line 395, the author should state in the footnote or in the methodology section the alpha level used (e.g., if the p value is lower than 0.05, the result is considered statistically significant). Done

- I appreciate the way the author listed the references, particularly the inclusion of the DOI when possible, but a few references need to be double checked. For instance, in line 856, add "*the"* before "*Journal of the Acoustical Society of*

America." In line 867, the ICPhS should be first defined as the "International Congress of Phonetic Sciences." Then it can abbreviated to "ICPhS." Also, please add the page numbers (e.g., 1957-1960). In line 994, the publisher's name is missed. Etc. Thanks for catching this. I have updated the reference list.

Other Comments:

- The author only reported the F0 results in the static but not the dynamic section. Also, the author did not examine them in the category separability index. Please state if there is any rationale. Thanks. This is a very valid point. F0 is potentially of interest. For instance, we know from research on vowel normalization that F0 have an *indirect* effect on vowel identity by providing the listener with an overall formant template of the talker. However, previous work on vowel acoustics have generally not included F0; if included, it is most often reported as mean F0 across the segment. It can further be noted in Figure 3 (densities along the diagonal) that F0 does not seem to carry much information for vowel identity in this material. There seems to be a very small but general difference between the long and short vowels, which might suggest an interaction with duration. In addition, there is also a feasibility aspect to including F0 in the dynamic analysis. Adjusting the Praat script to extract F0 at the same time-points as the other formants is not time-consuming in itself, however, the process of measurement correction that follows is (for reference, the correction of F1-F2-F3 took me 2-3 full work weeks). Finally, we know from separate work (auxiliary study in SI and discussed in General discussion in Persson & Jaeger, 2023) that a non-informative cue can add noise and drown out other cues in a category separability index. For these reasons, I decided to not go through the additional effort of including F0 in the analyses. However, all data is shared freely along with an updated version of the Praat script that now allows for extraction of F0 at the same time-points as the formants. I have further expanded on this question in the General discussion.

- Finally, please revise the manuscript using the Phonetica style sheet. Thanks. I have edited the reference list according to the De Gruyter Mouton Journal Style Sheet.

Reviewer #2

This is a very well written and systematically structured paper that presents new findings concerning Swedish vowels. Although several researchers have presented acoustic measurements of Swedish vowels before, as summarized by the author in the Introduction, carrying out this kind of studies on a regular basis is extremely important, as it allows linguists to draw conclusions about the mechanisms of language change both for the specific language and from a typological point of view.

I thank reviewer 2 for encouraging words and helpful feedback.

The statistical analyses are supported by a number of well chosen data visualizations. Figure 3 gives a very nice summary of the data. The resolution of the graphics was, however, a bit too low in the review version of the paper making the figures partly too blurry for thorough inspection. Please fix this problem. I have adjusted the font size in all figures throughout the manuscript. Unfortunately, I was not able to increase the resolution for submission of revised manuscript due to file size limit.

The vowels were elicited with hVd words, keeping the phonetic context constant. On the other hand, this means that the words were partly real words and partly non-words, which could have been problematized by the author. Thanks. I have added a couple of sentences to the footnote describing the words elicited.

A separability index was used for comparing the effect of different cues on vowel contrasts. While this method has some clear disadvantages as mentioned by the author, the motivation for choosing this method is not clearly stated in the paper. Why was an index calculated, instead of using a method for directly predicting vowel categories (e.g. logistic regression or linear discriminant analysis)? I decided on doing a separability index for feasibility and comparability reasons – it is a simple but frequently used index in phonetic research. Even if I share your reservations, the downside of logistic regression and linear discriminant analysis is that these approaches are prone to over-fitting and add researchers' degrees of freedom in the fitting from production to predicting perception. For what it is worth, in work conducted in parallel with this (Persson & Jaeger, 2024), I used a general model of speech perception (Bayesian ideal observers) to assess the effect of F3-inclusion for high-front vowel distinctions. The results from that work qualitatively replicated the results from the category separability index. I have now added a comment to this end in the Methodological considerations and future directions section (in General discussion), and also tried to be more clear on the motivation in the Methods section.

The discussion of the results has a very strong focus on language intrinsic factors on vowel shifts, while sociolinguistic factors are barely mentioned. This is especially striking, since one of the main results of the paper has to do with the centralization of [iː] and [yː], which is hypothesized to be explained by a "damped" pronunciation, a feature which is known to carry strong social meaning in Swedish (see e.g. Nilsson et al. 2021). A significant gender difference is also interpreted as anatomical differences not removed by normalization (footnote 10) without even considering the fact that there might be a real pronunciation difference; sociolinguists repeatedly find young women leading language change, so finding a gender difference would not be surprising. Thanks for bringing this up. As I have followed the advice of the associate editor to remove Study 2 from the paper, the General discussion is now re-organized and the section on vowel shifts substantially shortened. I have, however, included a comment about sociolinguistic factors and moved this into the main text (removing the footnote), and further added the reference to the description of the damped [ɨː].

Systematic differences between speakers are mentioned in the general discussion. There is the somewhat laconic comment (r.759-760): "Further insights into these individual differences can be gained by studying the SwehVd materials on a talker-

specific level". From the point of view of language change, an analysis of the amount of inter-speaker variability within each vowel category would have been an interesting addition to the main analysis, rather than only being assessed by impressionistic listening. However, I realize that due to space limitations it is probably not possible to add more quantitative results to this paper. As the SwehVd corpus has kindly been made publicly available, we can hopefully look forward to upcoming papers with further analyses of the data. Thanks. I agree that an inter-talker variability measure would have enriched the vowel change analysis. As Study 2 is now removed from the paper (see comment above), I hope to be able expand more on that analysis in a separate paper. I have nevertheless extended that comment somewhat in the General discussion.

References:
Nilsson, J. & Wenner, L & Leinonen, T. & Thorselius, E. (2021). New and old social meanings in urban and rural Sweden: The changing indexicalities of damped /i/. In Urban Matters: Current approaches in variationist sociolinguistics. John Benjamins. https://doi.org/10.1075/silv.27.08nil