

Sebastian Silas, MSc, BSc, BA

Doctoral Student, Musicology

[ssilas@stud.hmtm-hannover.de](mailto:ssilas@stud.hmtm-hannover.de)

Hochschule für Musik, Theater und Medien Hannover

Neues Haus 1

30175 Hanover

Hanover, 15 May 2023

Dr. Marc Brysbaert

Editor-in-Chief

Behavior Research Methods

**Submission of the manuscript: “Singing Ability Assessment: Development and validation of an open-source environment for singing data collection.”**

Dear Behavior Research Methods Editorial Team,

We would like to thank you for your careful consideration of our manuscript, *“Singing Ability Assessment: Development and validation of an open-source environment for singing data collection”.*

We have been working on the development of our singing test environment, the Singing Ability Assessment (SAA), for a number of years now. It has been an extremely positive experience to see it come to fruition in the context of submitting to *BRM*. We were delighted to receive the constructive reviews from experts in the field, who clearly have an excellent knowledge in the domain of singing accuracy measurement. As a result of their responses, we have updated the manuscript significantly, as requested, and feel this has substantially improved the paper. We are extremely grateful for the particularly engaging and thoughtful responses.  
  
Below, we document the changes we have made by responding to the reviewers’ requests. We have responded to nearly all requests by updating something in the manuscript. Where we have not done so, we note why. Our responses to the reviewer are highlighted in yellow. Updates to the manuscript are highlighted in green and indented. In view of our changes, we hope that find the manuscript suitable for publication.  
  
Yours faithfully,

Sebastian Silas, Daniel Müllensiefen and Reinhard Kopiez

**Response to Reviewers**

Brief summary   
In this paper, Silas et al. present Singing Ability Assessment (SAA): a new, open-source, R-based platform for vocal data collection and analysis. SAA will be a valuable resource to researchers in the fields of psychology (particularly music cognition and psycholinguistics), and pedagogy (particularly music education, sciences of teaching and learning). The paper compares outcomes from two versions of the SAA (offline and online analysis) with various perception and production measures and presents several PCA models to demonstrate the value of SAA. I do not recommend any further data collection for this paper. Ultimately, this paper has the potential to become an exceptional contribution to the field of vocal research, but major revision is warranted.  
  
General comments  
The paper represents a huge effort on the part of the authors, and I applaud them for their thoughtful approach to developing a system that can be deployed for real-time pitch production accuracy and singing research. There are many strengths of the paper, for which I include SAA itself. The platform is intriguing and I am very excited about its potential applications. The paper presents a well-written Intro that presents appropriate support in favor of the development of SAA. The purposes of Experiments 1 and 2 are easy to understand. However, there are several areas of the paper that must be addressed before publication.  
  
The method and results of each section seem to be incomplete or ambiguous. Many procedural and analytical details are underspecified. And, whereas the intro is accessibly written, the Method features technical terms (e.g., “d.entropy, “ngrukkon”) that are not always effectively operationalized.

Thank you for all your excellent points! We have studied them carefully and give answers below.

Several of the dependent variables such as *ngrukkon* are operationalised at least at a conceptual level e.g., in the following passage: “The similarity in interval content is captured by the ngrukkon measure that measures the difference of the occurrence frequencies of short pitch sequences (e.g., length 3-8) contained within two melodies (Uitdenbogerd, 2002).”

with a table in Appendix C to provide more information. However, we added a second note to the reader in-text in Experiment 2:

See Appendix C for descriptions about these variables and Silas and Müllensiefen (2023) for a comprehensive assessment of melodic similarity measures applied to sung recall data.

In addition, we’ve added definitions and operationalisation for those melodic features and dependent variables (e.g., d.entropy) that were not properly operationalised.

We have added the following in-text:

“d.entropy, an estimate of the amount of "surprise" in rhythmic information”

as well as a table for the melodic features in Appendix B.

One important point missing from the Intro (e.g., “Dirty” Musical Data section; and/or Discussion) is that there is a fundamental link between participant effort and performance that no singing assessment tool effectively measures. This is a classic problem in performance research, but it is especially relevant for singing because a singer’s performance depends on their skills \*and\* their effort. Obviously, participants vary in their exerted effort. Few researchers write about the influence of effort; instead, they implore their participants to “sing their best” and then assume that the recorded performance is representative as such. This has consequences, especially for automated and online/remote assessment platforms such as SAA. For example, I loved the idea of “transposing stimuli into the computed singing range of the participant” (p. 10) but that process depends almost wholly on each participant’s self-performed range. Obviously, this depends on effort, motivation, ability to follow instructions, and other factors that are difficult to control in the lab, and impossible to control remotely. The fact that few participants bothered to record second or third attempts in Experiment 2 (p. 41, lines 50-53) suggests that effort may not have been prioritized. Of course, these are not new problems, and they are reminiscent of self-report limitations from other areas of psychology. Thus, there is no need for Silas et al. to solve these problems, but I think they should be clearly recognized within the paper, especially as a caution to future researchers to avoid overextrapolation.

Thanks for this insightful observation. We added:

1. the following to the Dirty Musical Data section:

Such imperfect singing is also surely related to the amount of effort expended by a participant, a perennial issue for performance research in general (Silm, Pedaste, & Täht, 2020). The issue of effort and motivation affecting performance outcomes is very difficult, or impossible, to entirely mitigate, especially in the context of online research.

1. The following to the discussion of Experiment 2:

Use of the one-shot paradigm allowed us to separate multiple attempts at the same item into distinct audio files. However, it was observed that only a small proportion of participants were willing to optionally expend the extra effort to take multiple attempts. This effect of effort is a problem for all performance research (Silm et al., 2020), but is particularly difficult or impossible to control in the context of an online experiment. This suggests that researchers should be careful overextrapolating from results collected online, but also demonstrates the need to minimise test lengths where possible (e.g., through adaptive testing).

Additionally, we added the following to Future Directions:

In addition, the participant’s ability to sing useable long notes could be tested more thoroughly at the beginning of the SAA test protocol in order to triage participants early in the test which will further maximise data quality and save participant’s time. Since, as documented commonly in performance research (Silm et al., 2020), and suggested in Experiment 2 of our paper, whereby not many participants optionally took more than one attempt at the same melody, any measures to improve participant effort will be valuable to the SAA. This is a main purpose of adaptive tests: shorter tests can maximise effort (e.g., with fewer trials, participants may be more likely to have more attempts at each trial). However, in parallel to the SAA development, we have also been exploring how we can make our tests more aesthetically engaging and maximise motivation (Silas, 2023), which we plan to extend to the SAA for future data collections.

I suspect that audiophiles and sound engineers will be excited by SAA if you explain that the pitch data to be analyzed can be effectively captured and extracted regardless of the type of recording equipment that is used (e.g., laptop mic). The paper describes “triaging” but that is based on SNR. A broader question is “will the pitch data be useable if the singer uses a poor mic?” If you have such comparison data available, it would be useful to include.

We do have such data available, and present this in Appendix D. We also add the following to the General Discussion:

For instance, our analysis pipeline suggests that there is no difference between a user using an internal vs. an external microphone (see Appendix D), which suggests our audio transcription is relatively robust, once certain constraints have been fulfilled (e.g., a certain SNR).

There are several methodological issues I consider below, but one important, general issue, is relevant to the use of adaptive tests in SAA. I wonder whether having such low trial counts can provide meaningful results (15 or fewer items in the tests from Experiment 1: MDT, PIAT, PDCT, MPT, JaJ). I understand that adaptive tests are efficient, but I think also that they are heavily influenced by early performance, and early trials are vulnerable to each participant’s trial and error. One way to address this is to report test-retest reliability.

In the overall context of our online research, the use of adaptive tests vs. not can be considered a trade-off. For instance, without using adaptive, shorter tests, we would only be able to validate the SAA against a very small number of tests (i.e., it would be unreasonable to use several full-length tests in an online setting). By using more, but shorter tests, we can not only maximise participant attention, but provide more different previously validated tests to benchmark the SAA against. As we have generally shown, the correlations between the SAA test scores and these other tests are in line with expectations, and hence provide validity to the SAA, which would otherwise have been missing if we did not take this approach and would create a different kind of limitation.

Yes, the low number of trials for adaptive tests can be an issue but due to the constraints of online testing we have chosen the minimum number of trials that still appear acceptable for each test, based on their psychometric properties under different testing conditions (this is the reason why each adaptive test has its own specific number of trials, rather than us choosing e.g., 10 items for each adaptive test). We have also been careful to make the number of trials dependent on chance level of the specific test paradigm (i.e. the number of response options to choose from) which explains why tests with lower or very low guessing probability (e.g. JAJ, MDT) have fewer trials than the 2AFC tests (e.g. MPT, PIAT).

We now provide a table in Appendix A which gives the reliability values at the different test lengths we used. Also, we link to a [comprehensive study](https://www.cell.com/current-biology/fulltext/S0960-9822(23)00387-1?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0960982223003871%3Fshowall%3Dtrue) which has used several of the tests at similar lengths:

We refer the reader to Appendix A to see the reliability for the adaptive tests at the respective length we chose. Several have been comprehensively validated with respect to the item lengths we use (see Liu, Hilton, Bergelson, and Mehr, 2023).

Generally, we’d say that individual measurements might contain large measurement errors and shouldn’t be used for individual diagnostics. But these errors will be largely random and therefore the inference based on the full sample data should be more reliable than individual measurements. In any case, measurement error would lead to a downward-bias in terms of effect sizes  (‘attentuation’) and should guard us from over-stating results.

The results, especially for Experiment 2, are incomplete. One striking omission, especially within the context of the paper’s attempt to establish validity and reliability, is that there is no comparison of (even a subset of) the automated analyses to manual analyses or expert, subjective evaluations of singing accuracy.

As it happens, we have actually conducted such an experiment. This was originally part of the present manuscript, as a separate experiment, in addition to another experiment about the development of a screening tool to test whether participants are singing long notes at the beginning of the battery (and potentially disqualify them if not). However, we removed these experiments from the present manuscript to combine them with some other similar, more technical experiments around improving transcription accuracy. To foreshadow this publication, and respond to your point, we have added the following to Experiment 2 discussion:

A next step for obtaining reliability and validity of our analysis procedure is to compare the automated pYIN transcription of sung recall and subsequent opti3 scoring results to those produced when using transcriptions by a professional human rater, on the same data. We have conducted such an experiment, but it is beyond the scope of the present paper and will instead be presented in a forthcoming publication. However, preliminary results show that the mean edit distance accuracy between the pYIN output with default parameters settings and professional human transcription was 65%, but improved to an edit distance accuracy of 73% after optimizing the pYIN parameters (see Müllensiefen and Frieler (2007) for a description of edit distance applied to musical data). This suggests the automated transcription procedure is not perfect, but also corresponds largely to human professional transcription.

Given the open source nature of this project, the authors should consider publicly posting all of the nonidentifiable data for these experiments (not the raw vocal data, just the extracted pitch traces).

These, and the corresponding models, data and code to reproduce the analyses/manuscript are all released as part of the SAA package and corresponding manuscript repository. We added a note:

The SAA is currently available for use here: <https://saa.musicassessr.com>. The code to produce this manuscript and analyses can be found here: <https://github.com/sebsilas/SAA_Paper_2022>. An online demo can be found here: <https://adaptiveeartraining.com/SAA-demo>.

Finally, I wonder about data privacy, especially for the real-time system as described in Experiment 2. Where are these data stored (locally, externally, etc.) and how can users manage their data after using SAA?

There are two ways main ways the SAA can be used: locally on a desktop/laptop computer or deployed on a server. The results will be stored on the corresponding hard drive (in the folder where the application is deployed).

There is currently no feature for participants to manage their own data: it would be the responsibility of the experimenter to delete any data produced by the SAA, at the request of a participant. Anonymous participant IDs are stored in the audio file names, so it is relatively easy for an experimenter to delete participant data at their request.

The ability for a participant to delete their own data is an interesting idea. We do not have current plans to implement this but will definitely consider this for the future.  
  
Specific comments  
p. 8 lines 11-24: Here are two interesting ideas that deserve more development. For example, does melodic singing comprise the same skills as single note pitch imitation? Is melodic singing more informative? This is an opportunity for the authors because, as they describe earlier, automated computation has made singing assessment much easier. There is also an opportunity to reflect on this in the Discussion because, as the Exp. 2 results show, long note and melodic singing appear to be differentiated in the PCA models.

Thank you very much for this observation. We have updated the following passage from the introduction:

An important point to note is the fact that singing accuracy research is more concerned with fine-grained pitch control compared to melodic memory research, which is about understanding high-level melodic mental representations. But even singing accuracy appears to comprise two slightly disparate skills: accuracy (proximity to a target) and precision (consistency of reproduction) (Pfordresher et al., 2010). This highlights the need to not only measure singing accuracy and melodic memory via sung recall simultaneously, but also several constructs related to singing accuracy simultaneously.

And the following to the discussion of Experiment 2:

Broadly speaking, the results in Experiment 2 suggest that long note singing and melodic singing are somewhat differentiated, as indicated by the PCA models, suggesting they are relatively distinct tasks. This is most likely because long note singing does not involve sophisticated mental templates of melodic structure and is more about fine-grained pitch production monitoring. In other words, long note singing depends more on simple low-level perceptual processes and less on high-level learned representations.

pp. 11-12 Cognitive Modeling section: The intro argues that melodic memory and singing ability must be assesses simultaneously because they influence each other. However, it’s not clear why SAA is particularly well-suited to measure these capacities. Item response theory is very briefly described in this section, but I think this concept deserves more explanation.

We have rewritten and expanded this section to make things clearer:

Performance on an ability test can vary as a function of individual differences (i.e., some participants have a higher ability than others), but also as a function of items themselves (i.e., some items may be more difficult than others). In our study, there are two broad trial types: single long note singing and melodic singing. If long notes are presented in the vocal range of a participant (as we do here), the “item” effect of long notes are not expected to be important. That is: certain single pitches do not have properties which make them more or less difficult to sing than others.

Conversely, for melodic items with multiple notes, musical features emerge (e.g., tonality, contour, rhythm). Such emergent features clearly rely on high-level mental representations and templates (i.e., musical knowledge). Consequently, there can be significant variance in complexity when a melody is the item of testing, and these kind of item difficulties are important to model. Important melodic representations can be quantified for each melodic item across important dimensions (Müllensiefen, 2009). As suggested by previous literature (Baker, 2019; Dreyfus, Crawford, Müllensiefen, & Baker, 2016; Harrison, Musil, & Müllensiefen, 2016), there are several melodic features which could indicate an item’s complexity and predict singing performance (e.g., tonality, interval contour, a melody’s frequency in occurrence).

In order to formally relate structural features of melodies to the cognitive difficulty of melody processing, the main methodological approach we utilise here is *explanatory item-response theory* (*IRT*; De Boeck et al. (2016)). In this paper, *IRT* can be considered our first level of modelling, where melodic features become predictors of the *opti3* similarity score, which we take as representing variance in both singing accuracy and melodic memory. *IRT* is useful for our enquiry since it allows the simultaneous modelling of item difficulties and individual differences together via mixed effects modelling, whilst compartmentalising the variance into fixed item effects (melodic features), random item effects (unexplained effects due to melodic items) and participant effects (effects due to individual participants’ abilities). Additionally, an *IRT* model can be the basis of creating an adaptive test, which is highly efficient and can be variable in test length, since encoding relationships between item features and performance can be used to generate or select items based on modeled difficulties (for similar approaches see Gelding et al., 2021; Harrison, Collins, & Müllensiefen, 2017; Harrison & Müllensiefen, 2018; Harrison et al., 2016; Tsigeman et al., 2022). Such an adaptive test can hence be employed flexibly, with potential applications in education.

In this paper, our strategy to relate singing accuracy to melodic memory is to extract participant and item level scores from our *IRT* mixed effects models and use these outputs in further modelling. For instance, we use participant-level scores to represent individual differences in overall melodic memory and singing ability, and participant level indicators of singing accuracy alone (comprising e.g., single long note singing, singing accuracy, precision), to predict such outputs. This allows us to evaluate the potential extent that low-level singing abilities are responsible for the overall variance in singing performance, leaving the rest to do with variance in melodic memory, or being unexplained.

pp. 16-17 (Exp. 1) and p. 37 (Exp 2): Can you report any other demographic information about your participants? Most importantly, race/ethnicity, native language, music experience (beyond what you know from the Gold-MSI).

We were only able to add the following to the manuscript (Experiment 2):

67% were from the US, 25% UK, 5% Canada, and the remaining other countries.

p. 17: The paper explains that data are sent to AWS servers and downloaded by the authors. What about in Experiment 2? What kind of information is transmitted from the local user (raw data, aggregate results, etc.). See also my general comments.

We added the following clarification to Experiment 2:

The task was again deployed on an *AWS EC2* server instance, where the scoring was now done in real-time. All scores were downloaded post-hoc for statistical analyses.  
  
pp. 17-18: Can you validate whether participants chose the correct vocal range?

Yes – or we can make an educated guess. We added the following in-text:

Post-hoc, we estimated that at least 60.59% of participants selected an appropriate range, based on matching the mean note they sang across all trials to the closest mean note of the different vocal ranges. This estimate is likely a lower bound, since vocal ranges somewhat overlap, and the mean singing note computed from trials is also dependent on the randomly selected melody a participant heard.

p. 18 lines 24-25: Melodic memory span for \*known\* melodies easily exceeds 15 notes. Thus, I think you mean to say that 15 notes encompasses the memory span for \*novel\* melodies.

We updated the phrasing to say “short-term memory span of unknown melodies”.  
  
p. 19-23 (MDT, PIAT, PDCT, MPT, JaJ): Please provide more task details (e.g., MDT: how long are the melodies; how is one melody different; is the sequential position of the different melody randomized, etc.). Given that these are adaptive tests, I suspect that important parameters will change as each test progresses. Importantly, for each test, please also explain how it is scored/what the outcome represents.

These adaptive tasks rely on explanatory IRT models and make use of complex mechanism for choosing items and scoring participants on the fly. Thus, in order to not make the length of the present paper unwieldly, we suggest not to provide all details of each additional test. We have added the following note that points readers to the corresponding publication which are all open access:

We now list the other tests and questionnaires utilised. To save space in the present manuscript, we keep the descriptions relatively brief, and encourage the reader to refer to the corresponding publications for more details. Some of the tests are adaptive. We refer the reader to Appendix A to see the reliability for the adaptive tests at the respective length we chose. Several have been comprehensively validated with respect to the item lengths we use (see Liu, Hilton, Bergelson, and Mehr, 2023).

p. 20: Is the PIAT appropriate for all cultures? Dolscheid et al. (2013, 2020, 2022) demonstrated that the low-high pitch-language correspondences are related to language. For example, some cultures endorse “thick-thin” instead. It’s worth exploring the cultural ramifications of the PIAT and how performance influences overall assessment. Also, please provide additional task details concerning the adaptive qualities of the task including how the test is scored/what the outcome represents.

The PIAT is only appropriate for musical cultures that have low/high analogue pitch concepts associated with low/high frequencies. This is the case for Western music cultures and we can assume that all participants in the experiments were  very familiar with Western music styles, predominantly coming from English-speaking countries. Though, familiarity with Western classical music isn’t necessary as one reviewer has stated. The familiarity with Western music is an assumption that holds for all tests used here. The PIAT is scored using the underlying explanatory IAT model using the weighted likelihood scoring method. Scores range are given on a Z-transformed scaled and from -4 to +4 and represent low v. high abilities to imagine and manipulated pitch mentally.  
  
However, we suggest reserving these extra details from the paper, as explained above.

p. 24 lines 23-25: Perhaps “isorhythmic” (same rhythm) would be a more descriptive term than “arhythmic.”

We have debated the best way to describe such melodies in the past and have settled on arhythmic. The issue with “isorhythmic” is that there could be conflation with its other meaning in music theory: “Isorhythm is a musical technique using a repeating rhythmic pattern, called a talea, in at least one voice part throughout a composition”. Hence, we would rather leave the present term as it is.

p. 24: Please clarify the purpose/function of the rehearsal paradigm. I imagine it could be very useful to estimating participant effort (e.g., duration of rehearsal, number of attempts).

We added the following to the manuscript:

The originally intended function of the rehearsal paradigm was to observe the changes in patterns of sung recall across the temporal dimension of the trial (e.g., do N-gram chunks become more closely spaced throughout the rehearsal process?), and in particular, in a way that machine learning approaches could predict chunking patterns. However, this initial use case was discontinued, and the measure can here be thought of as utilising a basis for basic accuracy measures and effort expended by a participant (i.e., more notes, on the whole, = more effort).

p. 26 line 40: Why were all 247 participants included in the model if 19 were excluded earlier (p. 25 lines 32-33)? 

This was a typo: not all 247 participants were entered to the model; only those who passed the SNR test. Updated in-text.

p. 26-27: The predictors would make more sense to a naïve reader if some examples were provided (e.g., does entropy/surprise constitute a interval size, tonality expectations, something else?).

We think our descriptions are appropriate to provide a high-level understaning. For example, the i.entropy description we gave seems to directly answer your question: “as estimate of the average level of “surprise” or randomness in musical interval information”. Moreover, since the terms are defined fully in the corresponding publication we cite, we feel that it would contribute a substantial amount of additional text to the manuscript that would distract from the main narrative of the manuscript and it duplicate the detailed information in open access publications that we reference .

p. 27 line 23: I don’t understand why this term (and others throughout) are written in snake case. 

These correspond to the variable names in the package (which for syntactical reasons cannot have spaces) We have made this clear in the manuscript now:

Variable names use snake case, corresponding to their naming inside the SAA software.

p. 27 lines 27-29: I don’t understand how a simple proportion measure is a valid measure of melodic singing accuracy. It seems, based on the description, that it represents the proportion of correct notes compared to total notes sung. But this means that orderliness means nothing to the measure (whereas orderliness is critical to melody). Furthermore, it seems that the participant could sing one note on a trial and be scored at 100% correct (because failing to sing other notes does not reduce the score). I appreciate more clarity. Also, why did you measure accuracy of rehearsal paradigm? I assumed that was for practice. Why isn’t the accuracy measure applied to the 36 vocal trials in Procedures A and B? Or, is “rehearsal” what you are calling performance on each of the 36 trials? In that case, how do you determine when to stop/start counting the number of correct notes? A rehearsal block surely includes some incidental vocal content.

You are quite correct, proportion of correct note events is not an optimum measure of melodic singing, which is why we replaced it in Experiment 2 with something more principled. This DV is effectively deprecated now, but we nonetheless document the original research, as it unfolded. It was chosen at the time simply due to the extreme difference in length between the target melody and a sequence of rehearsed events produced by the rehearsal paradigm. Yes, in theory, if a participant sang a single correct note, they could score perfectly on such a measure. But since generally participants produced a pattern of rehearsed notes as instructed, this does not appear to be an issue. Broadly speaking, the measure appears to hold enough validity, as evidenced by its correlations with other tests in expected ways. But yes, it was not ideal, and is hence deprecated by Experiment 2 and onwards in our development.

Since we discontinued this variable by Experiment 2, we have decided not to add more content to the paper to justify its usage and hope you find that acceptable.  
  
By the way, we have a separate paper in review which profiles accuracy and similarity variables alongside each other on sung recall data and will be able to answer such questions comprehensively (Silas & Müllensiefen, 2023).

p. 28 Results: The results are not contextualized. Please plainly describe the influence of each effect on the outcome variable (SAA ability should be identified here).

Added:

In the mixed effects model, all seven melodic feature fixed effect predictors were significant predictors of proportion\_of\_correct\_note\_events. See Table 1 for this model’s parameter estimates. As suggested, more local variation in a melody’s contour, tonalness and whether a melody is rhythmic, are factors associated with a decrease the score. Conversely, a melody being more frequent in occurrence and having more surprise in musical interval or rhythmic information is associated with an increase in the score. The model mixed effects R2 values (Nakagawa & Schielzeth, 2013) were: conditional R2c = .52 and marginal R2m = .20.

p. 29 Table 1: Recommend to include confidence intervals for each estimate. Please use a note to identify the measures (JaJ, MDT, etc.) and the units of each measure where applicable.

Done. But the measures you talk about are not to do with that table.  
  
p. 31 line 30: Please define “nomothetic span.”

Added.  
  
pp. 31-32: The nonsignificant correlation between vocal pitch accuracy and pitch discrimination has been observed several times (e.g., Pfordresher & Brown, 2007). My interpretation of this null relationship is that most people have similar (good) pitch perception abilities and therefore present little variability in pitch discrimination. Reduction in range/variability substantially diminishes the likelihood of revealing a significant correlation. On the other hand, if you have amusic participants in your sample (and you probably don’t, given the recruitment strategy), then the correlation should be stronger.

Thank you, we were not aware about this, and have added a reference in-text:

The null correlation between singing and pitch discrimination abilities has actually been observed in previous research (e.g., Pfordresher & Brown, 2007), which can also provide validity to the SAA, in that it replicates previous research results.

Your explanation makes a lot of sense to us!

p. 32 line 15-21: I don’t understand the significance of this statement; are you suggesting that “sung recall” (imitative singing as in the SAA) depends on training? I think everyone agrees that singing performance depends on training. Or, are you saying that it is stable, like working memory capacity? Also, the paragraph is one sentence, so there is room to expand.

You’re right, this was overstated. We rephrased this:

Lastly, the SAA score was related to musical training, which suggests that singing abilities may be improved by musical training. However, the reverse causal explanation could also be true: those with already good singing abilities may be more likely to undertake more musical training (see Silas et al. (2022) for a discussion of such issues of causality in musical training).

p. 33 para 2: Please carefully explain how the measures work with a meaningful example. The authors write, “ngrukkon… measures the difference of the occurrence of frequencies of short pitch sequences,” but what does that mean? It would be very useful to the reader (who may be familiar with other measures of singing accuracy that you reviewed in the intro) to understand how this measure works on an example melody. For example, does the difference score capture note by note differences, sample by sample differences, or something else?

Exactly how this is done is fully transparent due to the open source nature of the project. We have added the following footnote:

For the implementation of this scoring methods, see https://github.com/sebsilas/musicassessr/blob/master/R/scoring\_simile.R.

Since this is the clearest way of disambiguating the process and enumerating an explanation would add substantial word count, we have decided not to write more in words. We hope that is okay.

p. 34 lines 17-19: Please describe how each measure is weighted.

Added equation.  
  
pp. 35-36 real-time SNR: Does the participant see their real-time SNR? Do they know that they have to keep SNR above 14? If so, do they sing louder? It is worth noting that vocal intensity influences pitch production.

There are various parameters to control this. We have explained these in a footnote:

Whether to use an *SNR* test is controlled via the *SNR\_test* argument to *SAA* functions; whether to allow multiple attempts, or disqualify on the first failure is controlled by the *allow\_repeat\_SNR\_tests* argument; whether to display the captured *SNR* as feedback to the participant is controlled via the *report\_SNR* argument.

p. 36 line 29-32: Please explain the transposition process. For example, are the stimuli normalized to the median of each participant’s estimated vocal range?

We added the following explanation:

After the individual vocal range has been captured, each stimulus will be transposed into the range of the participant such that its mean note is matched to the mean note of the user's range.

p. 38 lines 6-11: I’m not sure what you mean here. Is it (a) each melody is heard once and performed once, (b) each melody might be heard more than once, depending on randomization, or (c) the participant can attempt each melody as many times as they want to, but only the last recording is saved? The next paragraph adds to the confusion with terms like “one-attempt” and “multi-attempt.” 

We have enumerated this explanation, and hope this clarifies:

In Experiment 1, participants were encouraged to rehearse learning a melody aloud, and could hear a target melody up to three times during their rehearsal process. Consequently, each audio file might represent up to 3 distinct attempts (i.e., after each playback), as well as rehearsal within each attempt.

Conversely, the melody singing paradigm in Experiment 2 required participants to sing back a melody in ‘one shot’ after hearing it. The meaning of one-shot here means “without rehearsal” and that, after hearing a melody, the participant must try sing it back as best they can immediately (once). This produces a clear one-to-one correspondence between a heard melody, a sung recall and an audio file. However, as in Sloboda and Parker (1985), there can still be multiple attempts per item (by default, up to 4, for statistical reasons). The difference is that the one-shot paradigm produces one audio file per attempt, unlike in the rehearsal paradigm, where multiple distinct attempts might all be contained in one audio file. In both cases, attempts are nested in items; in the rehearsal paradigm, all attempts and rehearsal are nested in a single audio file; in the one-shot paradigm; each single attempt is in a single audio file.

pp. 39-41 Table 3: First, thank you for creating this very helpful table. However, there are some errors in the table. For example, look at the descriptions for interval\_precision and interval\_accuracy.

Corrected.

Also, some of these descriptions are underspecified or inappropriate. For example, why does pca\_melodic\_singing\_accuracy omit melody\_note\_accuracy?

Because this didn’t make it in the PCA, as described:

The variables note accuracy, note precision, interval accuracy and interval precision were submitted to a unidimensional PCA. In the solution, all indicators were at a communality (h2) value above .30, except for melody\_note\_accuracy. This was removed and, in the final solution (see Table 8), note precision, interval precision and melody interval accuracy had factor loadings above .50 and h2 values above .4. The single factor achieved to explain 51% of variance in the data. Components scores were extracted from this model and we called the new aggregate variable pca\_melodic\_singing\_accuracy.

p. 43 Long note singing: This paragraph seems to be incomplete. It describes three components but doesn’t provide satisfactory statistical context to support the interpretations (e.g., of the functional representation of each component).

We added a table of the PCA results.  
  
pp. 43-45 Long note singing and Melody singing: Recommend that you consider using tables to present the results of each PCA model, providing confidence intervals where appropriate. 

Here too.

p. 43-34 Models 1, 2.1, 2.2, 3.2: Another reason to include tables is because it would enhance the reader’s ability to understand what predictors are included within each model and how SAA\_Ability, SAA\_Ability\_Arrhythmic, and SAA\_Ability\_Rhythmic were derived. It is worth elaborating on the proportion of variance explained for each model, especially in comparison to previous measures (e.g., p. 45 Principal components analysis of melody singing variables).

Added.

p. 45 lines 18-20: Recommend you consider including a \*summary\* table for the difficulty scores to, for example, report descriptive statistics (M, SD) for each melody duration.

There is only one difficulty score: these are just model predictions for each melody, given its fixed effects predictor values. The model is based on a normal distribution with mean = 0, so only the SD would be relevant. We do not add this because this doesn’t appear to be what you meant. Instead, we add some clarity in-text:

These difficulty values are essentially a model prediction (where opti3 is the dependent variable), given the fixed effects values for each melody in the corpus (i.e., it is an output of the sum of the fixed effects values for each melody, weighted by the fixed effects coefficients described in this paper).

p. 45 lines 31-32: I wouldn’t have expected melody\_note\_accuracy to be excluded from the model. I wonder whether there was a problem aligning note onsets to target notes (e.g., Figure 4).

Yes, this surprised us too. We don’t think aligning note onsets to target notes could have been the cause though. Figure 4 is not showing computed onsets in relation to the sung recall frequency curve: it is showing a target melody in relation to the sung recall frequency curve. We have updated the caption for clarity.  
  
pp. 46-47 Table 4: Aside from being difficult to interpret, this table includes several spurious correlations. Look how deviceMemory is correlated with self-reported singing abilities! With a huge correlation matrix like this you must interpret the results carefully, focusing only on correlations of a priori interest. Otherwise, for an exploratory analysis, please urge caution to the reader, and consider applying alpha corrections.

The issue appears to be to do with the document rendering process/presentational because that particular correlation is not actually significant. We have tried to make the formatting better. Also, the correlations are now Holms corrected.  
  
p. 49 Table 5: Many of the estimates are 0 or very close to 0. For example, the estimate and confidence interval for Melody note accuracy is 0, and yet it is significant? That doesn’t make sense. Can you please explain?

Bear in mind that the SAA\_Ability score comes from a normal distribution with a mean of 0 and only ranges [-0.2, 0.4]. The very low beta estimates are thus difficult to interpret, with some appearing as 0 due to rounding artifacts. We have now standardised all variables before entry to this the regression model now, and the beta estimates and confidence intervals are now easier to interpret.  
  
p. 50 lines 36-48: It would be wonderful to read more about your validity analysis and interpretation. For example, please reflect on the correlations from Table 4 and the proportion of variance explained for the relevant models.

We have fleshed out this section some more.  
  
p. 57 Limitations: This section is two sentences in length. Please consider my other comments to supplement this section. I would like to also suggest that you elaborate on potential vulnerabilities within the analysis pipeline (Figure 1). Each stage presents potential for improvement (that’s not because you did a bad job, it’s simply because SAA is in continual development). For example, although the paper describes triaging base on SNR, it does not address the robustness of SAA to data of various quality, nor does it even mention the presence or removal of acoustic artifacts which I expect to be peppered throughout any natural dataset, including this one. Consider also that pYin is one of several pitch estimation algorithms, it may be neither the fastest or most accurate in this application. It may be worth noting that some music researchers and theorists have cautioned the use of “objective” aptitude tests due to theoretical, analytical, and cultural implications. Thus, a tool like SAA might be more appropriately utilized within experimental contexts (e.g., comparing singing performance before and after a vocal lesson intervention, or comparing singing across various stimulus manipulation conditions).  
  
Overall: There are a few minor wording errors (e.g., p7 line 23) that should be corrected.

Thanks, corrected, and we’ve tried to spot any other mistakes!

Thank you for your hard work! I look forward to your revisions.

Reviewer: 2  
  
Comments to the Author  
Review of the manuscript "Singing Ability Assessment: Development and Validation of an Open-source Environment for Singing Data Collection"  
  
This work aimed to develop the Singing Ability Assessment (SAA) test and open-source testing environment of sung recall. The protocol incorporates three sets of procedures and underlying statistical models that reflect (1) single long note singing ability, (2) rhythmic melodies singing ability, and (3) arhythmic melodies singing ability in five key open source R packages. The idea is to integrate singing accuracy and melodic recall into the SAA score.   
  
As the result of Experiment I where the idea was to validate the SAA via the construction of an explanatory IRT model and to correlate the derived SAA score with other previously validated ability tests. Experiment I also supported the hypothesis that features that indicate melodic complexity are relevant predictors for the SAA score offering explanatory power in accordance with the previous literature. Experiment II used a "one-shot" paradigm, and its overarching objective was to update the SAA task to be more sophisticated and prepared for adaptive testing. An important aspect of the SAA environment is the possibility to quickly assess and triage out from further testing the participant if the acoustical conditions at the side of the testee are too noisy for getting relevant results.   
  
This work is written well with good clarity and by using an appropriate style. It has a logical structure and correct formatting. The authors are informed well about the developments and literature in the corresponding realm. Also, the possible directions for further development are explained in detail.   
  
Still, there are some local problems, fixing which would strengthen the whole outcome.   
  
1. The link to the webpage referred does not open:   
p 17, 37, 38 (error message: 403 Forbidden)  
<https://www.slicethepie.com/>

This appears to work now.

2. p 21, line 11: The unit to denote frequency is Hz (not HZ).

Changed.

3. The definition of the abbreviation PCA is given only on p 40, although it was first used in the abstract without explaining its meaning.

We have unabbreviated abstract. Now the first time the term appears it is abbreviated.

4. There is inconsistency in formatting when using the same term repeatedly (e.g., SAA ability score, SAA score, SAA Ability score, (SAA) ability (see p 42-43).

We have made the term consistent.

5. Usually, a space is inserted between the numerical value and the unit (e.g., 500 ms vs 500ms). The only exception from this rule is % where the space is not used.

We have tried to update all such occurences.

6. p 35, a formula on the bottom of the page: the definition of RMS is not given. Usually, RMS denotes the Root-Mean-Square of "something". This "something" is missing here. In the present case, this is probably the RMS of the sound signal (which corresponds to the air pressure).

7. Also, the numerical value of SNR calculated by the formula on p35 must have the unit, which is dB (see p 36), as the logarithmic scale is used. (If you use the SNR without the unit "dB", it refers to the linear comparison of RMS values, which is not the case here.)

We updated the description in the manuscript to address these points:

The SNR formula consists of computing the ratio of the signal amplitude over the background noise amplitude. These amplitudes can be estimated with the root mean square, and the SNR is calculated in dB according to [formula].

7. It is not clear whether the outcome of this research is four R packages (as stated in the abstract) or five R packages (as reported on p 14).

It is five. Updated the reference to “four”.  
  
I suggest publishing this manuscript in the Behavior Research Methods after fixing these small problems mentioned above.   
  
  
Reviewer: 3  
  
Comments to the Author  
In this paper, the authors present and validate a codebase/test battery that can facilitate online data collection related to a range of singing behaviors. The paper is well-written, and the codebase will be a valuable tool for researchers. However, the paper's focus on detailed criteria and complex use cases obscures the platform's flexibility and usefulness.  
  
I want to pre-empt my comments by saying that I fall within your target audience for this paper: I would conceivably want to use an online singing paradigm in the future. Many of my comments derive from the questions I would have if I were trying to actively assess whether this platform would work for some of the most basic singing experiments I could think of. I hope this perspective is helpful.   
  
Broad comment: In several places, this paper is written so specifically that it is difficult to see this tool's more general use cases, which does both the authors and the readers a disservice. For example, in the "Present Study" section (page 13), the authors list that they do not know of a previous experiment fulfilling all eight listed criteria. My immediate reaction to this detailed list of criteria was, "This tool is over-engineered and too complicated for my needs." To reference an idiom: why would I use a sledgehammer to crack a nut? All those eight criteria may be helpful and essential in particular contexts. Still, most readers (like myself) may only need to fulfill a subset of those eight constraints during their data collection. Listing those constraints makes this platform appear much less flexible than it (hopefully) is. As a reader, I would have appreciated a more straightforward introduction to the aims of your program.

As a concrete suggestion: It may be helpful for the authors to take a step back and think of some simpler use cases for this type of online singing test, then frame at least the ‘Present Study’ section with those simpler use cases in mind. For example, what if a scientist wants a simple measure of pitch accuracy (melody\_note\_accuracy, as operationalized here) when singing back a melody? They would need to read all the way to Experiment 2 to understand that this could be possible (maybe I missed this, but could an experimenter use two-note melodies to look at single intervals? As the paper is written, this is still unclear to me!). I think the authors and readers would be better served if the paper could briefly describe the features of their setup in an additive manner. i.e., This framework is a) open source, b) can enable online data collection for several simple singing tasks such as one note holds and melody copying, c) enables the automatic filtering of participants based on background noise, d) be connected to statistical frameworks… etc. The fact that your framework is supported by specific statistical models will be much less important to many general users than the fact that it enables online singing collection and has signal-to-noise exclusion criteria. These more basic strengths of the author’s framework should be emphasized.

Thank you for this great point, we agree that we definitely over-sold the advanced features at the expense of the basic ones, and their accessibility. However, it is important to note that, other than its accessibility, without the advanced features, our test would be much like previously described tests. Especially in the context of an article in *Behaviour Research Methods*, we feel it is important to emphasise our methodological contributions, rather than simply making previous methodologies accessible.  
  
To meet these criteria, we have restructured and reframed the narrative under a heading “Motivations”, which make the two-fold nature of our contributions clear. We hope that this separation of making previous procedures accessible and our advances allows readers coming from different perspectives to see where the framework could be applied for them.

Minor comments (in no particular order):   
  
- First sentence: ‘Music is a profoundly cognitive phenomenon which is concerned with producing sequences of auditory events.’ This sentence is so vague as to be meaningless (and could just as easily apply to speech). Maybe something framing this paper as focusing on the production rather than the perception aspect of music would be better?

We have removed the first sentence and rewritten the opening gambit.  
  
- Page 11, line 45: The term ‘ability task’ may be jargon and should be clarified. I don’t understand what is meant by that.

The term “task” is often used as opposed to “test” in working memory research, which we somewhat align our thinking with. But this is confusing, and so we have replaced it with “ability test”.  
  
- A more detailed description of the melody input format would be helpful. A statement to the effect of ‘any MIDI database can be substituted in future testing’ (or whatever form is needed for this test, be it .wav, etc.) would be helpful for researchers deciding whether they can flexibly adopt this framework.

We added a footnote:

These can be created with the itembankr R package from .mid files, .musicxml files or a dataframe of melody pitches and frequencies. See: <https://github.com/sebsilas/itembankr>.

- Page 15: ‘Our first main objective’: Redundant.

Removed.  
  
- Page 16, line 32, typo, ‘Related to task performance’? Also, this sentence did not make sense to me. How could individual differences in task performance not be related to task performance?

Updated for clarity:

Secondly, in addition to structural features of melodies, we hypothesised that performance on our new test would also be related to individual difference scores on other questionnaires/tests of related musical and non-musical abilities, in line with Pfordresher et al. (2015)’s model and the other literature reviewed above.

- Page 36: I do not think you need a reference for this SNR formula. This is a standard formula. Either way, the Sueur reference did not seem appropriate.

Removed the reference.  
  
- Table 4: Are the data normal? I.e., is it valid to use Pearson’s correlations? Furthermore, given the number of correlations, I believe a discussion of multiple comparisons is warranted, and the p values may need to be corrected.

Pearson’s correlations do not assume normality, so we believe this is valid. Perhaps you mean they are vulnerable to outliers? In our methodology, this shouldn’t be an issue since our scoring is based on extracting random effects participant intercepts which mitigates extreme outliers due to partial pooling. The correlations are Holm’s corrected, we have added this to the caption.  
  
- Given that he was a Ph.D. student in the second author’s lab, I was surprised that this paper did not engage with or even cite the recently published work by Manuel Anglada-Tort (<https://www.biorxiv.org/content/10.1101/2022.05.10.491366v1.full>). His other paper with Nori Jacoby, which just came out, should also be added; I know this was published while the current paper was under review, but it is directly relevant. <https://www.cell.com/current-biology/pdf/S0960-9822(23)00243-9.pdf)>. Previous work by Nori Jacoby (<https://www.sciencedirect.com/science/article/pii/S096098221931036X>) is also germane and should be referenced. That work uses similar interval accuracy measures, for example.

Yes, we became aware about this just after we submitted our paper to BRM! We have added the following to our introduction:

Very recently, and innovately, large scale singing research has been also conducted online outside of the context of Western music (Anglada-Tort, Harrison, & Jacoby, 2022). Anglada-Tort et al. (2022)’s approach also uses automated scoring and an online testing environment, with the main task being to sing back short melodies as immediate recalls.

AND

In terms of more sophisticated usage, and advancing previous methodologies and theoretical insights, like other innovative recent research (Anglada-Tort et al., 2022; Jacoby et al., 2019), our test framework and approach also makes a number of important contributions beyond its accessibility.

- With Jacoby 2019 in mind, a limitation that the authors should mention is that their statistical modeling only applies to Western Classical music. A future direction of this type of framework would be to extend it to different musical systems.

We added to the following to the limitations section:

Second, our statistical modeling only applies to Western music. A future direction of this type of framework might be to extend it to different musical systems. We point readers to the other very innovative research in this regard (see Anglada-Tort et al. (2022; Jacoby et al. (2019).