DOI: 10.1111/nyas.15081

ORIGINAL ARTICLE

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES

How accurate are self-evaluations of singing ability?

Daniel Yeom[1,†] | Kendall S. Stead[1,2,†] | Yi Ting Tan[3] | Gary E. McPherson[3] | Sarah J. Wilson[1,4]

[1]Melbourne School of Psychological Sciences, University of Melbourne, Melbourne, Victoria, Australia

[2]School of Psychological Sciences, Macquarie University, Sydney, New South Wales, Australia

[3]Melbourne Conservatorium of Music, University of Melbourne, Melbourne, Victoria, Australia

[4]Department of Medicine, Epilepsy Research Centre, University of Melbourne, Austin Health, Heidelberg, Victoria, Australia

Correspondence
Sarah J. Wilson, Melbourne School of Psychological Sciences, University of Melbourne, Redmond Barry Building, Parkville, VIC 3010, Australia. Email: sarahw@unimelb.edu.au

[†]These authors contributed to the work equally and share first-author status.

Funding information
Australian Government Research Training Program Scholarship; Australian Research Council, Grant/Award Numbers: DP170102479, DP200100961

Abstract

Research has shown that people inaccurately assess their own abilities on self-report measures, including academic, athletic, and music ability. Evidence suggests this is also true for singing, with individuals either overestimating or underestimating their level of singing competency. In this paper, we present the Melbourne Singing Tool Questionnaire (MST-Q), a brief 16-item measure exploring people's self-perceptions of singing ability and engagement with singing. Using a large sample of Australian twins ($n = 996$), we identified three latent factors underlying MST-Q items and examined whether these factors were related to an objective phenotypic measure of singing ability. The three factors were identified as Personal Engagement, Social Engagement, and Self-Evaluation. All factors were positively associated with objective singing performance, with the Self-Evaluation factor yielding the strongest correlation ($r = 0.66$). Both the Self-Evaluation factor and a single self-report item of singing ability shared the same predictive strength. Contrary to expectations, our findings suggest that self-evaluation strongly predicts singing ability, and this self-evaluation is of higher predictive value than self-reported engagement with music and singing.

KEYWORDS
self-assessment, self-report, singing, singing engagement, singing questionnaire

INTRODUCTION

As humans, we are typically not accurate at evaluating our own abilities.[1–4] This is evident across a range of abilities, with consistently low correlations between our ability to complete a task and our self-evaluated ability to complete the same task.[1,3] In a recent meta-synthesis of literature on self-evaluated ability, Zell and Krizan[1] estimated an average correlation of $r = 0.29$ between self-report and performance. Correlations were low across skill-based domains (e.g., sports ability); cognitive domains, such as verbal and spatial reasoning[2]; and academic ability.[1] This mismatch has been observed across a range of cultures,[4] suggesting that in many circumstances, people's perceptions of themselves do not typically align with their actual level of competence. A nuance of these findings reported by Kim et al.[4] reveals that people with higher levels of competency tend to *under*estimate their performance and vice versa. This appears to be driven by both a lack of metacognitive awareness and a desire to appear more favorable to other people.[4]

Humans are also not accurate evaluators of their musical ability, with mismatches between self-report and objective testing of musical abilities. For instance, two studies suggest that a substantial proportion of people (17%–20%) self-report having difficulties with music to that extent that they consider themselves "tone deaf"[5,6] despite evidence that self-reports are weak predictors of objective performance on amusia batteries.[5] High false-positive rates of self-reported absolute pitch have also been observed,[7] suggesting that even rare musical traits are subject to this bias. To our knowledge, there are no meta-analyses of the relationship between self-reported and objective music ability. This presents a challenge for various forms of music research, with an important example being music genetics; particularly genome-wide association studies (GWAS) where whole genomes are scanned for associations with complex traits. Phenotyping based on short self-report questions (an approach known as minimal phenotyping) has become common in GWAS in order to recruit the large samples required for adequate statistical power.[8] Despite self-reported assessments of ability being less accurate than objective measures, brief questionnaires present a unique opportunity to phenotype complex traits at a very large scale. Single-item questions are particularly attractive as they maximize the efficiency of data collection at a low cost.[9] However, evidence suggests that at least in clinical populations, single items capture less genetic signal compared to multifaceted measures like multiple-question inventories and structured diagnostic interviews.[9–11] Given there are no similar sets of diagnostic criteria for musical traits, testing the utility of single and composite self-report measures is, therefore, crucial for understanding how to meaningfully phenotype traits at scale.

Self-evaluations of singing ability

Singing is pervasive within all musical cultures and develops alongside language early in life.[12–14] Unlike performance on most other musical instruments, the act of physically producing musical tones using the voice is something that can be achieved naturally and without formal training. Singing, therefore, provides a unique method for assessing how members of the general population feel about a specific aspect of their musical ability. In this paper, we define "singing ability" in the context of everyday singing, where pitch accuracy constitutes a core feature of singing performance.[15] Moreover, singing ability is heritable, and potential genetic loci have been previously identified,[15,16] supporting its genetic basis.

Most people in the general population can sing with a reasonable level of accuracy, particularly when singing familiar tunes.[13,15,17–19] However, like other domains, people appear to misestimate their singing ability.[5,19–21] While only 10%–15% of the population are considered "poor-pitch singers,"[13,15,17–19] a higher proportion may label themselves as "tone-deaf" singers.[19,20,22] For example, in a survey of undergraduate students, 59% reported an inability to reproduce sung melodies.[20] Despite this, previous studies have identified a positive relationship between self-evaluations and objective singing ability (r = 0.18–0.41, ρ = 0.44–0.49), suggesting that these self-evaluations correlate with actual performance.[5,20,23] Pfordresher and Demorest[18] found weak to moderate correlations between participants' self-evaluations of their singing and both objective pitch accuracy (r = 0.18–0.41) and externally rated evaluations by judges (r = 0.22–0.38). However, a subset of their participants underestimated their abilities, with those deemed accurate singers assessing themselves poorly. Consistent with this, Larrouy-Maestri et al.[24] found that more accurate singers tended to underestimate their ability, whereas those who were more inaccurate overestimated their ability. Importantly, however, self-reported ability in this study was limited to a single phrase of "Happy Birthday" rather than perceived general singing ability.

These differing findings may be partially explained by methodological differences in tasks. Zell and Krizan[1] note that correlations between self-evaluated ability and performance are moderated by three key factors: (1) self-evaluations, which tend to be closer to actual performance if tasks are familiar; (2) measurements with objective criteria (as opposed to subjective assessments by judges); and (3) low complexity. Notably, several studies have used a single task, such as singing a familiar song,[24,25] or multiple tasks that are treated in isolation,[18,23] to measure objective ability. The use of a single task is unlikely to fully capture general singing ability, with observed correlations pertaining to specific task effects that will inevitably vary in difficulty and familiarity.[18] Rather, an objective measure using multiple tasks to capture everyday singing under a single, unified construct (i.e., a composite measure) would be expected to show a stronger relationship with self-evaluated singing ability.[1] In our previous work, we developed such a measure,[15] namely, the Singing Ability Phenotypic Index (SPI). This is a robust and reproducible measure of everyday singing ability that has high loadings of familiar and unfamiliar singing tasks, which we use in this study to examine the relationship between self-reported and objective singing ability.

Both objective and self-evaluated singing performance are also closely tied to singing engagement.[26] Children and adults who hold negative self-evaluations of their singing are less likely to participate in singing in the future.[26,27] This, in turn, may remove opportunities for people to develop their singing ability.[26] However, the relationship between engagement and self-evaluation is likely to be bidirectional. Engagement with music early in life also influences singing ability,[28] which has downstream implications for how people self-evaluate their abilities. The various ways in which people engage with singing, therefore, may also relate to both self-evaluated and objective singing performance.

Using a large sample of Australian twins (n = 996), we report the Melbourne Singing Tool Questionnaire (MST-Q), a 16-item questionnaire designed to assess self-evaluation and engagement with singing. This questionnaire forms part of the broader Melbourne Singing Tool (MST), an online platform including the SPI that was designed for large-scale singing studies.[15,29] In this study, we aimed to examine the relationship between self-evaluated and objective singing ability in two ways. First, by exploring latent factors underlying the 16 items of the MST-Q and examining how these factors correlate with objective singing performance as measured by the SPI. Second, we examined the relationship between a single self-report item of general singing ability taken from the MST-Q with the SPI, to compare the predictive utility of the latent factors with this single item.

MATERIALS AND METHODS

Participants

This study utilized the same participant cohort as Yeom et al.[15] In brief, participants were comprised of 1189 twins between 15 and 90 years old ($M = 43.4$, $SD = 16.5$). Most of the sample were female (875 females; 73.6%) and reported prior or current music training ($n = 868$; 73%). Of these, 113 (9.5%) reported voice training as their primary instrument. Recruitment material specified that voice ability was not a prerequisite for participation. Recruitment was aided by Twins Research Australia (TRA) and other recruitment strategies, such as social media and advertisements within the University of Melbourne. The study was approved by the Human Research Ethics Committee of the University of Melbourne (ID 1750061), and all participants provided informed consent. After listwise deletion of missing data, 996 cases were used for data analyses.

The Melbourne Singing Tool

Objective singing performance data were derived from the MST, as described in Tan et al.[29] The MST consists of three everyday singing tasks, a pitch perception task, and a series of self-report questionnaires that comprise the MST-Q. The singing tasks involve singing a series of single notes (vocal pitch-matching), a familiar tune ("Happy Birthday"), and a series of unfamiliar, seven-note tonal melodies. The song "Happy Birthday" is highly familiar across the world, and notably one of only a few well-known songs commonly sung on a regular basis, making it an ideal task for measuring singing ability in everyday contexts.[30] The tasks are presented in the corresponding vocal range of the participant's reported sex. The tool has high internal consistency, test-retest reliability, and convergent validity.[29]

The SPI was derived from five measures of pitch and interval deviation calculated from the performance of the three singing tasks. The three pitch deviation measures captured the absolute distance between each sung and target note, and the two interval deviation measures captured the absolute difference between each sung and target interval for the melodic tasks. These measures loaded strongly onto one latent factor, referred to as "Singing Ability," that explained a high proportion of variance (75%) across the tasks,[15] with factor scores used as the SPI.[15,29]

The MST-Q consists of a series of questions about a participant's music and singing engagement and self-evaluations, both in childhood and currently. It has 16 self-report items requiring a response on a 5-point scale (*1 = Never*, *5 = Always*). The items were designed around three categories: six relate to an individual's childhood singing and music experience, six relate to current singing and music experience, and four relate to self-evaluation of singing and music perception abilities. The four self-evaluation items were adapted from existing questions in the Goldsmiths Musical Sophistication Index, but included humming and whistling to capture a broader range of vocal musical behaviors.[31] Only one item (*When I sing/hum/whistle, I have no idea whether I'm in tune or not*) required reverse-scoring to ensure that lower responses corresponded with higher ratings. In addition, a single self-report item (*How much singing ability do you have?*) required participants to rate their general ability on a 7-point scale (*1 = None at all*, *7 = A great deal*).

Procedure

At the start of the online tool, participants are informed that by completing the MST, they are providing informed consent. They are then required to test their audio and recording equipment and complete a brief demographic questionnaire before completing the performance tasks and the MST-Q. Participants are advised to complete the tool in a quiet space to minimize external disruptions and to take breaks if needed to minimize fatigue. In total, the MST takes between 20 and 30 min to complete, with data stored on a server for subsequent analysis.[29]

Data analysis

All analyses were conducted in R 4.2.2 (RRID: SCR_001905)[32] using the psych package (RRID: SCR_021744).[33] Bartlett's Test of Sphericity (χ^2 (120) = 7498.93, $p < 0.001$) and the Kaiser–Myer–Olkin (KMO) Measure of Sampling Adequacy (overall KMO = 0.89; individual KMO's > 0.81) demonstrated that MST-Q data were appropriate for factor analysis. We used a principal axis extraction in combination with an oblimin rotation, as principal axis factoring makes no strong assumptions about underlying distributions and so is appropriate when multivariate normality is not assumed.[34] Factor loadings below 0.40 were excluded. For interpretability, we aimed for a solution with a simple structure that avoided cross-loading factors, removed items with no factor loading, and ensured that each factor had at least three loadings.

We used parallel analysis and the eigenvalue > 1 rule alongside theoretical considerations to determine the number of factors to retain. The eigenvalue rule suggested a single-factor solution that likely underestimated the appropriate number of factors. Parallel analysis suggested an initial solution of five factors (explaining 57.9% of variance), but this resulted in one cross-loading item and two factors that related to only two items (excluding the cross-loaded item). A four-factor solution that explained 54.2% of variance was also examined but still had one cross-loading item, one item with no factor loading, and one factor with only two items (excluding the cross-loaded item). Finally, we extracted a three-factor solution that adhered to a simple structure, which we used as our final solution. Factor scores for the three factors were calculated using a Thurstone regression method.[35]

TABLE 1 Descriptive statistics for the MST-Q ($n = 996$).

Item	*Mean*	*SD*
Background		
Q1. In my childhood, I sang, hummed, or whistled with my friends.	2.83	1.09
Q2. In my childhood, I sang, hummed, or whistled to myself.	3.67	0.95
Q3. In my childhood, I sang, hummed, or whistled with my family.	2.91	1.18
Q4. In my childhood, I listened to music.	4.12	0.82
Q5. In my childhood, I was constantly surrounded by music.	3.64	1.08
Current		
Q6. How often do you currently sing, hum, or whistle by yourself?	3.58	1.00
Q7. How often do you currently sing with your family?	2.44	1.13
Q8. How often do you currently sing with your friends?	2.24	1.05
Q9. How often do you currently sing in a choir/group?	1.77	1.22
Q10. How often do you currently sing as a solo performer?	1.39	0.85
Q11. How often do you currently listen to music attentively (e.g., during a concert)?	3.32	1.12
Q12. How often do you currently listen to music as background music (e.g., while eating, commuting, or exercising)?	3.87	0.94
Self-Evaluation		
Q13. I can tell when people sing out of tune.	4.00	0.91
Q14. When I sing/hum/whistle, I have no idea whether I'm in tune or not.[a]	3.34	1.30
Q15. I find it easy to sing, hum, or whistle a familiar tune from memory.	3.37	1.04
Q16. After hearing a new song two or three times, I can usually sing/hum/whistle it by myself.	3.90	1.06

Note: All items were measured on a 5-point scale (*1 = Never*, *5 = Always*).
Abbreviations: MST-Q, Melbourne Singing Tool Questionnaire; SD, standard deviation.
[a]Reverse-scored item.

RESULTS

Exploratory factor analysis of the MST-Q

Table 1 shows descriptive statistics for each MST-Q item. Figure 1 shows the bivariate correlations between the 16 items. Cronbach's alpha (α = 0.90, 95% CI: [0.89, 0.91]) and McDonald's omega (ω_t = 0.92) indicate excellent internal consistency, with α remaining high if each item is dropped ($\alpha = 0.89$ on average; Table S1). The final solution comprised three factors and cumulatively explained 49.6% of the variance. All items corresponded to a factor and no cross-loading was present.

Factor loadings for the three-factor solution are presented in Table 2. Factor 1 comprised six items relating to engagement with singing and music when alone, such as individual singing or music listening either in childhood or currently. It was labeled "Personal Engagement." Factor 2 comprised six items relating to singing and music engagement in social contexts, such as performing for others or group singing either in childhood or currently. It was labeled "Social Engagement." Factor 3 comprised four items relating to self-evaluations of singing ability and perceptive skills, and was labeled "Self-Evaluation."

Correlations of MST-Q factors with objective singing performance

Table 3 shows the correlations between MST-Q factor scores and the SPI. The strongest correlation was observed for Self-Evaluation ($r(994) = 0.663$, $p < 0.001$), followed by Social Engagement ($r(994) = 0.462$, $p < 0.001$), and Personal Engagement ($r(994) = 0.356$, $p < 0.001$). As shown in Figure 2, an individual's evaluation of their own productive and perceptual singing skills is strongly associated with their objective singing performance. This relationship was similar between monozygotic and dizygotic twins (Table S2).

Comparing Self-Evaluation to self-reported general singing ability

The Self-Evaluation factor was strongly correlated with the single self-report item of general singing ability ($r(994) = 0.717$, $p < 0.001$; Table 3). We ran a multiple regression to compare their predictive strengths for objective singing performance (SPI). The results of the analysis are presented in Table 4. Regression diagnostics showed no violations in collinearity (VIF = 2.06) and multivariate normality

TABLE 2 Exploratory factor analysis loadings for the 16 MST-Q items.

Item	Factor 1: Personal Engagement	Factor 2: Social Engagement	Factor 3: Self-Evaluation
Q4. In my childhood, I listened to music.	0.79		
Q12. How often do you currently listen to music as background music (e.g., while eating, commuting, or exercising)?	0.66		
Q5. In my childhood, I was constantly surrounded by music.	0.60		
Q6. How often do you currently sing, hum, or whistle by yourself?	0.51		
Q2. In my childhood, I sang, hummed, or whistled to myself.	0.44		
Q11. How often do you currently listen to music attentively (e.g., during a concert)?	0.43		
Q8. How often do you currently sing with your friends?		0.79	
Q9. How often do you currently sing in a choir/group?		0.63	
Q1. In my childhood, I sang, hummed, or whistled with my friends.		0.56	
Q10. How often do you currently sing as a solo performer?		0.56	
Q7. How often do you currently sing with your family?		0.55	
Q3. In my childhood, I sang, hummed, or whistled with my family.		0.53	
Q13. I can tell when people sing out of tune.			0.79
Q14. When I sing/hum/whistle, I have no idea whether I'm in tune or not.[a]			0.61
Q15. I find it easy to sing, hum, or whistle a familiar tune from memory.			0.61
Q16. After hearing a new song two or three times, I can usually sing/hum/whistle it by myself.			0.55

Note: All items measured on a 5-point scale (*1* = *Never*, *5* = *Always*).
Abbreviation: MST-Q, Melbourne Singing Tool Questionnaire.
[a]Reverse-scored item.

TABLE 3 Correlations between the MST-Q factors, the SPI, and the single self-report item of global singing ability ($n = 996$).

Measure	SPI	Single-Item	Factor 1: Self-Evaluation	Factor 2: Social Engagement	Factor 3: Personal Engagement
SPI	–				
Single-Item	0.663*	–			
Self-Evaluation	0.663*	0.717*	–		
Social Engagement	0.462*	0.622*	0.520*	–	–
Personal Engagement	0.356*	0.376*	0.462*	0.447*	–

Note: SPI, Singing Ability Phenotypic Index factor scores; Single-Item, self-report item of global singing ability.
*$p < 0.001$.

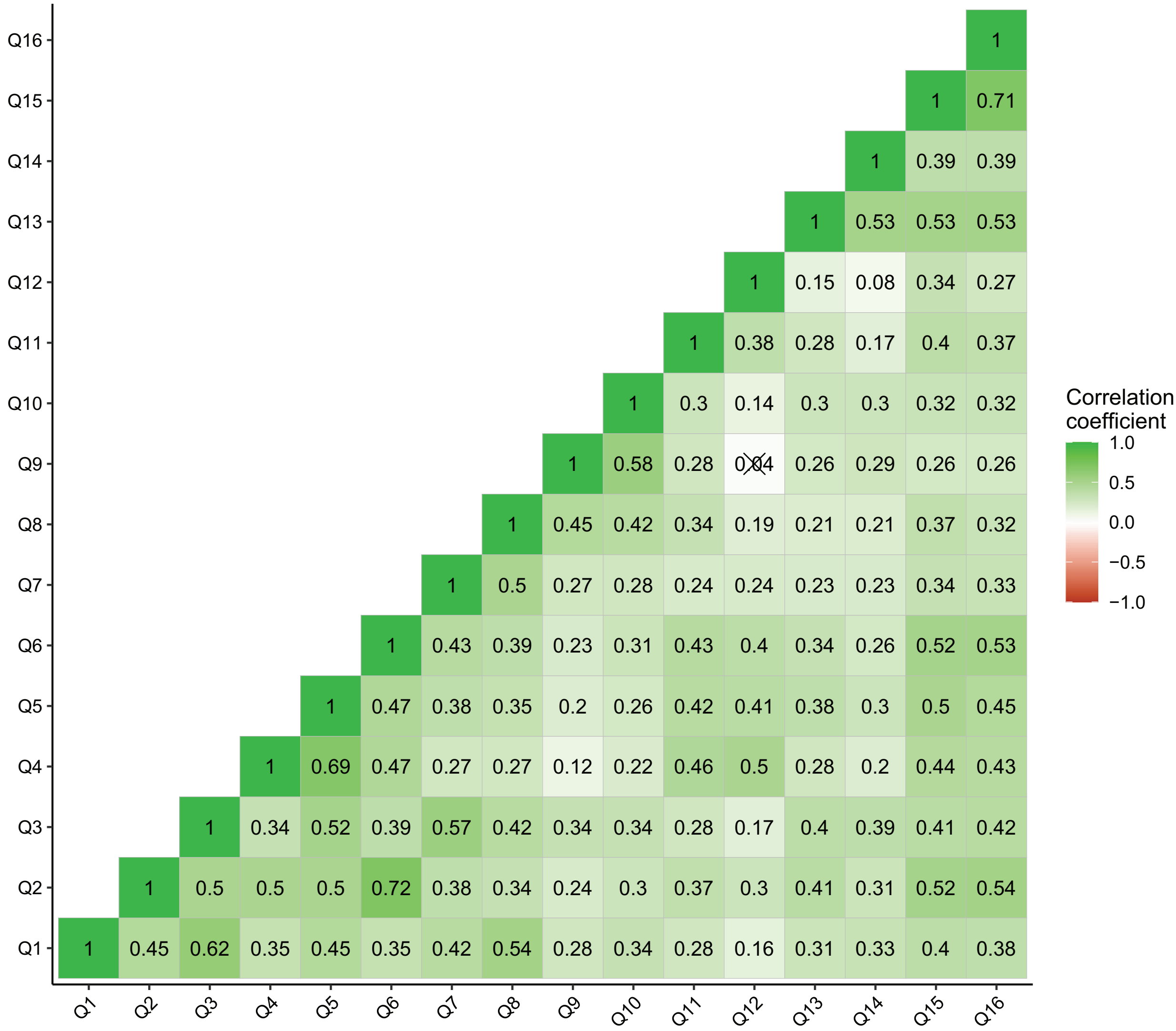

FIGURE 1 Heatmap showing item correlations in the MST-Q ($n = 996$). *Note*: "X" represents nonsignificant correlations. Abbreviation: MST-Q, Melbourne Singing Tool Questionnaire.

TABLE 4 Multiple regression analysis: Predictors of the Singing Phenotypic Index ($n = 996$).

Effect	β	B	SE	t	p
Intercept		−0.725	0.062	−11.708	< 0.001
Self-Evaluation Factor	0.386	0.376	0.031	12.135	< 0.001
Single Item	0.386	0.196	0.016	12.155	< 0.001

Abbreviation: SE, standard error.

(Shapiro–Wilks' $W = 0.998$, $p = 0.227$). Overall, the model explained 51.2% of the variance in objective singing performance (Adjusted $R^2 = 0.511$; $F(2, 993) = 520.79$, $p < 0.001$). Importantly, the Self-Evaluation factor and the single item were both significant and equally strong predictors of the SPI (see Table 4).

DISCUSSION

In this study, we showed that self-evaluation of singing ability predicts objective singing performance, using either a composite or single-item, self-report measure. We also showed that two other factors relate

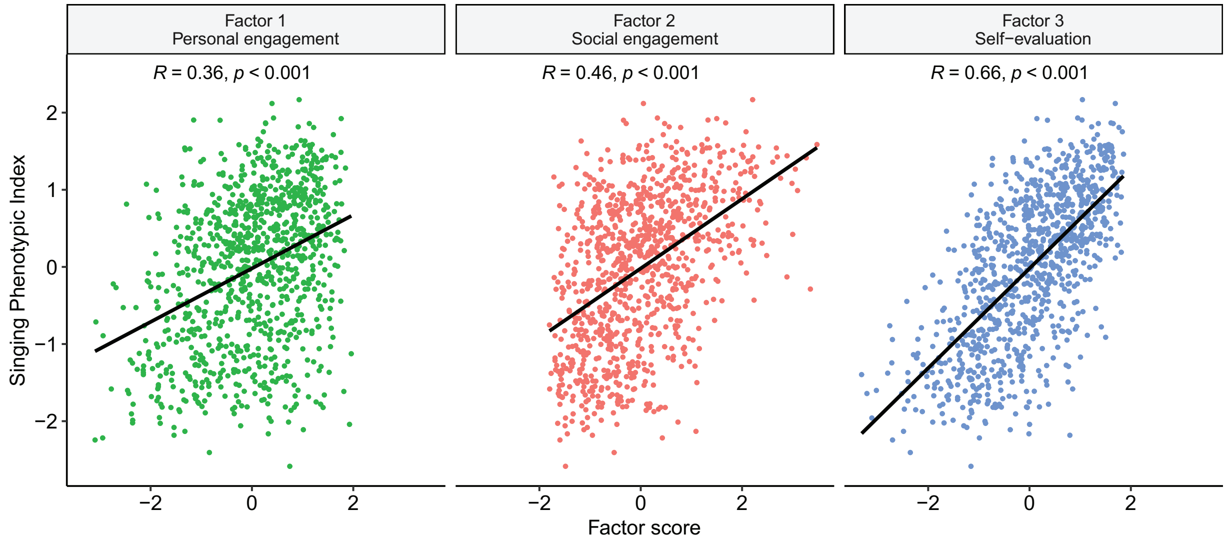

FIGURE 2 MST-Q factor score correlations with the Singing Phenotypic Index ($n = 996$). Abbreviation: MST-Q, Melbourne Singing Tool Questionnaire.

to objective singing performance, namely, personal engagement and social engagement in singing and music, both in childhood and currently. Social engagement was more strongly correlated with objective performance compared to personal engagement, suggesting that social contexts and group singing may be especially important for developing singing ability. This is consistent with our previous work showing that singing in early childhood with family predicts objective singing performance.[15]

In contrast to prior research suggesting a mismatch between people's self-evaluations and actual singing proficiency, our Self-Evaluation factor yielded a strong association with objective singing performance. This correlation ($r = 0.66$) is higher than those previously reported in music and singing,[18,24] and nonmusical abilities,[1,2,4,36] and may be explained by differences in our objective performance and self-report measures. We used a psychometrically robust objective measure (SPI) that captures singing performance across familiar and unfamiliar tasks that had demonstrated high internal and test-retest reliability and convergent validity.[29] The vocal pitch-matching and unfamiliar melody tasks were designed to be of low complexity, while the most complex task ("Happy Birthday") used a highly familiar tune that most people have repeatedly sung in social groups. Thus, the tasks capture the three factors shown to increase associations between objective performance and self-evaluation.[1] They also reflect everyday singing behaviors that people typically engage in, namely, reproducing highly familiar songs and imitating simple new tunes,[24,25] which people may commonly consider when asked to reflect upon their general singing ability.

The self-report questions we used also differ from those employed in previous studies. Our Self-Evaluation factor includes specific questions about a range of different aspects of singing that have not been previously validated against objective singing measures. Prior studies have employed items that are more broadly related to musical ability (e.g., *I am musically talented*, *I enjoy singing*, *People think I am a good singer*[18]) or only relate to the singing task participants were asked to perform.[23] The various questions used in previous studies assess constructs related to general singing ability, but do not fully capture it; for example, asking participants about their proficiency to sing in tune[25] will primarily reflect beliefs about their pitch accuracy of singing. The questions in our Self-Evaluative factor cover a range of singing behaviors that are underpinned by a general self-evaluation of ability. People likely hold intuitive beliefs about their general everyday singing ability because it is a ubiquitous skill. While correlations of a similar magnitude to ours are rare, Westbrook et al.[37] note that higher correlations have typically been observed when the ability can be evaluated relative to a reference, such as academic ability.[2,37] However, people may misestimate their abilities on individual singing tasks either due to differences in how familiar or complex these tasks are, or because the self-evaluative questions measure constructs that do not closely align with the singing task. Thus, the difference between our results and previous studies may reflect the wording and content of the self-report items combined with the use of the psychometrically robust SPI, thus potentially accounting for our stronger correlational results. While we focus on the viability of a minimal phenotype of singing in this study, future work on predictive models that incorporate other factors, such as singing experience, would be valuable to extend the findings reported here.

Unexpectedly, we found that the composite Self-Evaluation factor yielded the same predictive power as a single-item measure of self-reported general singing ability. In contrast to the composite measure, this single item asked participants to simply rate their ability on a 7-point scale. This suggests that singing ability can be minimally phenotyped using either a single-item or a composite Self-Evaluation measure in large-scale genetic studies of singing. While singing performance data can be robustly collected online, reliable and valid extraction of objective acoustic measures across various singing tasks still requires substantial manual processing.[18,29] Clearly, this is not feasible for the large sample sizes required for adequately powered GWAS. Thus, the equal predictive utility of the composite and single items suggests that both measures likely tap into the same general construct of ability but may index it in different ways (i.e., capture overlapping but distinct aspects of self-evaluation). Evidence from GWAS of other psychological traits suggests that composite measures likely capture more genetic signal than single items.[11,38] It is unclear from our results whether this would be true for singing ability given the equal predictive strength of the single item and composite measures. Future GWAS studies investigating this issue are indicated.

Our study has some limitations. First, there are more women than men in the sample. Twins Research Australia, which helped facilitate recruitment, also has a higher proportion of women in its database.[15,39] In addition, the questionnaire items rely on retrospective recall, particularly those relating to early childhood. Responses to these questions may be subject to recall bias, particularly for older participants, although the strong correlations between the three factors of the MST-Q and the SPI suggest this is unlikely to have greatly impacted the results. Finally, the order in which the MST tasks were presented may have influenced the strength of the association between self-evaluation and objective performance. Similar to previous studies,[18,23] participants completed the performance tasks immediately before responding to the self-evaluation measures, providing a recency effect that may have afforded a more accurate assessment of their singing ability. Meta-analytic evidence from other domains suggests that correlations tend to be larger when self-evaluations are completed after a task, though the recency effects are generally small.[1,3] Future use of counterbalanced designs would allow the strength of any order effects to be assessed.

CONCLUSION

Here, we show that self-evaluation predicts objective singing performance to a large extent, using either a composite or single-item measure of self-reported singing ability designed for online singing research. Our results suggest that these measures may be useful for minimal phenotyping of singing ability in large-scale research, including GWAS studies.

AUTHOR CONTRIBUTIONS

Y.T.T., G.E.M., and S.J.W. designed the original study and created the singing tool. Y.T.T. collected data. D.Y., K.S.S., and S.J.W. analyzed the data. D.Y.Y., K.S.S., G.E.M., and S.J.W. collectively wrote and approved the final manuscript.

ACKNOWLEDGMENTS

We thank the twins for their participation in the study. We thank Trisnasari Fraser for her assistance with recruitment and Dr. Valerie Yap for her help with cleaning the data. We also thank Mark Solly and Oscar Correa for their assistance in developing the Melbourne Singing Tool. This research was supported by Australian Research Council Discovery Project Grants (DP170102479 and DP200100961) and an Australian Government Research Training Program Scholarship. The research was facilitated through access to Twins Research Australia, a national resource supported by a Centre of Research Excellence Grant (ID: 1079102) from the National Health and Medical Research Council.

Open access publishing facilitated by The University of Melbourne, as part of the Wiley - The University of Melbourne agreement via the Council of Australian University Librarians.

COMPETING INTERESTS

The authors declare no competing interests.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author and with the permission of Twins Research Australia. The data are not publicly available due to ethical restrictions.

ORCID

Daniel Yeom https://orcid.org/0000-0002-1281-6299

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/nyas.15081

REFERENCES

1. Zell, E., & Krizan, Z. (2014). Do people have insight into their abilities? A metasynthesis. *Perspectives on Psychological Science*, *9*(2), 111–125. https://doi.org/10.1177/1745691613518075
2. Freund, P. A., & Kasten, N. (2012). How smart do you think you are? A meta-analysis on the validity of self-estimates of cognitive ability. *Psychological Bulletin*, *138*(2), 296–321. https://doi.org/10.1037/a0026556
3. Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology*, *67*(3), 280–296. https://doi.org/10.1037/0021-9010.67.3.280
4. Kim, Y.-H., Kwon, H., Lee, J., & Chiu, C.-Y. (2016). Why do people overestimate or underestimate their abilities? A cross-culturally valid model of cognitive and motivational processes in self-assessment biases. *Journal of Cross-Cultural Psychology*, *47*(9), 1201–1216. https://doi.org/10.1177/0022022116661243
5. Cuddy, L. L., Balkwill, L. L., Peretz, I., & Holden, R. R. (2005). Musical difficulties are rare: A study of "tone deafness" among university students. *Annals of the New York Academy of Sciences*, *1060*(1), 311–324. https://doi.org/10.1196/annals.1360.026
6. Peretz, I., & Vuvan, D. T. (2017). Prevalence of congenital amusia. *European Journal of Human Genetics*, *25*(5), 625–630. https://doi.org/10.1038/ejhg.2017.15
7. Baharloo, S., Service, S. K., Risch, N., Gitschier, J., & Freimer, N. B. (2000). Familial aggregation of absolute pitch. *American Journal of Human Genetics*, *67*(3), 755–758. https://doi.org/10.1086/303057
8. Sanchez-Roige, S., & Palmer, A. A. (2020). Emerging phenotyping strategies will advance our understanding of psychiatric genetics. *Nature Neuroscience*, *23*(4), 475–480. https://doi.org/10.1038/s41593-020-0609-7
9. Mcintosh, A. M., Sullivan, P. F., & Lewis, C. M. (2019). Uncovering the genetic architecture of major depression. *Neuron*, *102*(1), 91–103. https://doi.org/10.1016/j.neuron.2019.03.022
10. Flint, J. (2023). The genetic basis of major depressive disorder. *Molecular Psychiatry*, *28*, 2254–2265. https://doi.org/10.1038/s41380-023-01957-9
11. Cai, N., Revez, J. A., Adams, M. J., Andlauer, T. F. M., Breen, G., Byrne, E. M., Clarke, T. K., Forstner, A. J., Grabe, H. J., Hamilton, S. P., Levinson, D. F., Lewis, C. M., Lewis, G., Martin, N. G., Milaneschi, Y., Mors, O.,

Müller-Myhsok, B., Penninx, B. W. J. H., Perlis, R. H., & Flint, J. (2020). Minimal phenotyping yields genome-wide association signals of low specificity for major depression. *Nature Genetics*, *52*(4), 437–447. https://doi.org/10.1038/s41588-020-0594-5
12. Welch, G. F. (2015). Singing and vocal development. In G. E. McPherson (Ed.), *The child as musician: A handbook of musical development* (2nd ed., pp. 441–461). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198744443.001.0001
13. Dalla Bella, S., Giguère, J.-F., & Peretz, I. (2007). Singing proficiency in the general population. *Journal of the Acoustical Society of America*, *121*(2), 1182–1189. https://doi.org/10.1121/1.2427111
14. Stadler Elmer, S. (2021). Song transmission as a formal cultural practice. *Frontiers in Psychology*, *12*, 654282. https://doi.org/10.3389/fpsyg.2021.654282
15. Yeom, D., Tan, Y. T., Haslam, N., Mosing, M. A., Yap, V. M. Z., Fraser, T., Hildebrand, M. S., Berkovic, S. F., McPherson, G. E., Peretz, I., & Wilson, S. J. (2022). Genetic factors and shared environment contribute equally to objective singing ability. *iScience*, *25*(6), 104360. https://doi.org/10.1016/j.isci.2022.104360
16. Park, H., Lee, S., Kim, H. J., Ju, Y. S., Shin, J. Y., Hong, D., von Grotthuss, M., Lee, D. S., Park, C., Kim, J. H., Kim, B., Yoo, Y. J., Cho, S. I., Sung, J., Lee, C., Kim, J. I., & Seo, J. S. (2012). Comprehensive genomic analyses associate UGT8 variants with musical ability in a Mongolian population. *Journal of Medical Genetics*, *49*(12), 747–752. https://doi.org/10.1136/jmedgenet-2012-101209
17. Pfordresher, P. Q. (2022). Singing accuracy across the lifespan. *Annals of the New York Academy of Sciences*, *1515*(1), 120–128. https://doi.org/10.1111/nyas.14815
18. Pfordresher, P. Q., & Demorest, S. M. (2021). The prevalence and correlates of accurate singing. *Journal of Research in Music Education*, *69*(1), 5–23. https://doi.org/10.1177/0022429420951630
19. Dalla Bella, S., & Berkowska, M. (2009). Singing proficiency in the majority: Normality and "phenotypes" of poor singing. *Annals of the New York Academy of Sciences*, *1169*(1), 99–107. https://doi.org/10.1111/j.1749-6632.2009.04558.x
20. Pfordresher, P. Q., & Brown, S. (2007). Poor-pitch singing in the absence of "tone deafness". *Music Perception*, *25*(2), 95–115. https://doi.org/10.1525/mp.2007.25.2.95
21. Dalla Bella, S., Giguère, J.-F., & Peretz, I. (2009). Singing in congenital amusia. *Journal of the Acoustical Society of America*, *126*(1), 414–424. https://doi.org/10.1121/1.3132504
22. Berkowska, M., & Dalla Bella, S. (2013). Uncovering phenotypes of poor-pitch singing: The Sung Performance Battery (SPB). *Frontiers in Psychology*, *4*, 714. https://doi.org/10.3389/fpsyg.2013.00714
23. Wise, K. J., & Sloboda, J. A. (2008). Establishing an empirical profile of self-defined "tone-deafness": Perception, singing performance and self-assessment. *Musicae Scientiae*, *12*(1), 3–26. https://doi.org/10.1177/102986490801200102
24. Larrouy-Maestri, P., Wang, X., Vairo Nunes, R., & Poeppel, D. (2023). Are you your own best judge? On the self-evaluation of singing. *Journal of Voice*, *37*(4), 633.e7–633.e14. https://doi.org/10.1016/j.jvoice.2021.03.028
25. Larrouy-Maestri, P., Leveque, Y., Schon, D., Giovanni, A., & Morsomme, D. (2013). The evaluation of singing voice accuracy: A comparison between subjective and objective methods. *Journal of Voice*, *27*(2), 259 e251–259 e255. https://doi.org/10.1016/j.jvoice.2012.11.003
26. Demorest, S. M., Kelley, J., & Pfordresher, P. Q. (2017). Singing ability, musical self-concept, and future music participation. *Journal of Research in Music Education*, *64*(4), 405–420. https://doi.org/10.1177/0022429416680096
27. Krause, A. E., Forbes, M., & Lowe-Brown, X. (2022). Does reality television-style singing influence singing self-concept? *Journal of Voice*. Advance online publication. https://doi.org/10.1016/j.jvoice.2022.06.024
28. Barrett, M. S., Zhukov, K., Brown, J. E., & Welch, G. F. (2020). Evaluating the impact of a generalist teacher-led music program on early childhood school children's singing skills and attitudes to music. *Psychology of Music*, *48*(1), 120–136. https://doi.org/10.1177/0305735618790355
29. Tan, Y. T., Peretz, I., McPherson, G. E., & Wilson, S. J. (2021). Establishing the reliability and validity of web-based singing research. *Music Perception*, *38*(4), 386–405. https://doi.org/10.1525/mp.2021.38.4.386
30. Nichols, B., Hua, A., & Wang, Z. (2022). Defining "happy" in happy birthday: Singing accuracy a construct based on range and intervals. *Psychology of Music*, *51*(4), 1414–1423. https://doi.org/10.1177/03057356221137932
31. Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. *PLoS ONE*, *9*(2), e89642. https://doi.org/10.1371/journal.pone.0089642
32. R Core Team. (2022). *R: A language environment for statistical computing*. R Foundation for Statistical Computing.
33. Revelle, W. (2019). *psych: Procedures for psychological, psychometric, and personality research*. CRAN.
34. Watkins, M. W. (2018). Exploratory factor analysis: A guide to best practice. *Journal of Black Psychology*, *44*(3), 219–246. https://doi.org/10.1177/0095798418771807
35. Thurstone, L. L. (1935). *The vectors of mind*. University of Chicago Press.
36. Davis, D. A., Mazmanian, P. E., Fordis, M., Van Harrison, R., Thorpe, K. E., & Perrier, L. (2006). Accuracy of physician self-assessment compared with observed measures of competence: A systematic review. *JAMA*, *296*(9), 1094–1102. https://doi.org/10.1001/jama.296.9.1094
37. Westbrook, B. W., Buck, R. W., Wynne, D. C., & Sanford, E. (1994). Career maturity in adolescence: Reliability and validity of self-ratings of abilities by gender and ethnicity. *Journal of Career Assessment*, *2*(2), 125–161. https://doi.org/10.1177/106907279400200203
38. Jamshidi, J., Williams, L. M., Schofield, P. R., Park, H. R. P., Montalto, A., Chilver, M. R., Bryant, R. A., Toma, C., Fullerton, J. M., & Gatt, J. M. (2020). Diverse phenotypic measurements of wellbeing: Heritability, temporal stability and the variance explained by polygenic scores. *Genes, Brain and Behavior*, *19*(8), e12694. https://doi.org/10.1111/gbb.12694
39. Murphy, K., Lam, J., Cutler, T., Tyler, J., Calais-Ferreira, L., Li, S., Little, C., Ferreira, P., Craig, J. M., Scurrah, K. J., & Hopper, J. L. (2019). Twins Research Australia: A new paradigm for driving twin research. *Twin Research and Human Genetics*, *22*(6), 438–445. https://doi.org/10.1017/thg.2019.101

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Yeom, D., Stead, K. S., Tan, Y. T., McPherson, G. E., & Wilson, S. J. (2023). How accurate are self-evaluations of singing ability? *Ann NY Acad Sci.*, *1530*, 87–95. https://doi.org/10.1111/nyas.15081