The Role of Tonal Memory in Learning Novel Musical Tasks on Theremin

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Psychology

by

Benjamin S. Reynolds

Adviser: Linda Hayes, Ph.D.

August 2019
THE GRADUATE SCHOOL

We recommend that the dissertation prepared under our supervision by

BENJAMIN REYNOLDS

Entitled

THE ROLE OF TONAL MEMORY IN LEARNING NOVEL MUSICAL TASKS ON THEREMIN

be accepted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Dr. Linda J. Parrott Hayes, Adviser

Dr. Wilfred "Larry" Williams, Committee Member
Dr. Jean-Paul Perrotte, Committee Member
Dr. Ramona Houmanfar, Committee Member
Dr. Stephen C. Eubanks, Graduate School Representative

David W. Zeh, Ph.D., Dean, Graduate School

August-2019
Abstract

Tonal memory has been shown to significantly correlate with extent of musical achievement more than other auditory discriminative abilities (e.g. Bugg & Herpel, 1946; Henson & Wyke, 1982). Despite these findings, no study has yet determined whether tonal memory predicts future music learning potential. This collection of two studies sought to determine the correlates of music learning by implementing the Seashore Tonal Memory Test (1957) and Drake Musical Memory Test (1957) as assessments of tonal and musical memory followed by a novel direct instruction application using the theremin as a musical apparatus to teach absolute pitch production. Demographic information of participants was also correlated to tonal and musical memory scores as well as overall probe average performances and durations to completing the first learning task. Results showed that the Seashore test predicted prompted music learning typical of ensemble settings and the Drake test predicted both prompted and advanced unprompted music learning typical to absolute pitch. Additionally, gender and musical training of more than three years were shown to significantly affect Seashore and Drake composite score, though they were not predictive of overall probe average performance or duration to completion of the first learning task. Implications of these findings for music education and music learning research are discussed.

Keywords: tonal memory, music learning, absolute pitch, direct instruction, theremin
Dedication

To my wife, Christine, and my family, for everything you did to help me both large and small.

And a special note of appreciation to our cat, Franz “Bucket” Kafka, for long days and nights spent on the couch supporting this work. He hardly ever took breaks.
Acknowledgments

First acknowledgments are extended to the 61 UNR students who volunteered 2 hours each toward this project, with special mention of those few students who volunteered without course credit solely due to interest in the project. Distinguished acknowledgments are also reserved for the members of the dissertation committee who volunteered their time and invaluable feedback toward the improvement of both studies. The quality of this work would not be quite so exacting without the combined expertise of Drs. Hayes, Williams, Houmanfar, Eubanks, and Perrotte who brought to bear a considerable breadth of knowledge toward the refinement of my methods and analyses. I couldn’t have done it without any of you. I owe you all my sincerest thanks.
# Table of Contents

Abstract i

Dedication ii

Acknowledgments iii

Table of Contents iv

List of Tables v

List of Figures vi

Introduction 1

Assessment 10

Study 1 19

Study 1 Results 28

Study 1 Discussion 37

Study 2 41

Study 2 Results 49

Study 2 Discussion 68

Conclusion 74

References 77

Appendix A: Study 1 Procedural Integrity Data Sheet 81

Appendix B: Study 2 Procedural Integrity Data Sheet 84

Appendix C: Demographic Information Questionnaire 87

Appendix D: Study 2 Learning Task B Random Number Sheet 88
List of Tables

Table 1  Descriptive Statistics and Shapiro-Wilk Tests of Normality across Variables (Study 1)

Table 2  Descriptive Statistics and Shapiro-Wilk Tests of Normality across Variables (Study 2)
List of Figures

Figure 1. Overall distribution of Seashore Tonal Memory scores across 61 participants from Study 1 and Study 2

Figure 2. Distribution of Drake Musical Memory Form A and Form B combined scores across 61 participants from Study 1 and Study 2

Figure 3. Nonparametric correlation between Seashore and Drake scores across 61 participants from Study 1 and Study 2 with LOESS line

Figure 4. Training progression chart (Study 1)

Figure 5. Linear correlation between Drake Musical Memory A+B score and probe A average performance (Study 1)

Figure 6. Linear correlation between Drake Musical Memory A+B score and overall probe average performance (Study 1)

Figure 7. Top two study 1 participants’ (19 and 21) overall average probe scores (99th and 82nd percentiles) along with Seashore and Drake scores

Figure 8. Middle two study 1 participants’ (8 and 12) overall average probe scores (both 57th percentile) along with Seashore and Drake scores

Figure 9. Bottom two study 1 participants’ (20 and 14) overall average probe scores (11th and 9th percentiles) along with Seashore and Drake scores

Figure 10. Training progression chart (Study 2)

Figure 11. Non-linear correlation between Seashore Tonal Memory score and duration to final probe A with LOESS line (Study 2)

Figure 12. Non-linear correlation between Seashore Tonal Memory score and total trial blocks to final probe A with LOESS line (Study 2)
Figure 13. Non-linear correlation between Seashore Tonal Memory score and highest instructional level attained with line of best fit (Study 2)

Figure 14. Non-linear correlation between Drake Musical Memory A+B score and duration to final probe A with LOESS line (Study 2)

Figure 15. Non-linear correlation between Drake Musical Memory A+B score and total trial blocks to final probe A with LOESS line (Study 2)

Figure 16. Non-linear correlation between Drake Musical Memory A+B score and highest instructional level attained with line of best fit (Study 2)

Figure 17. Linear correlation between Drake Musical Memory A+B score and probe A average performance (Study 2)

Figure 18. Linear correlation between Drake Musical Memory A+B score and overall probe average performance (Study 2)

Figure 19. Top two study 2 participants’ (50 and 26) overall average probe scores (99th and 98th percentiles) along with Seashore and Drake scores

Figure 20. Middle two study 2 participants’ (30 and 43) overall average probe scores (both 50th percentile) along with Seashore and Drake scores

Figure 21. Bottom two study 2 participants’ (49 and 24) overall average probe scores (10th and 7th percentiles) along with Seashore and Drake scores
The Role of Tonal Memory in Learning Novel Musical Tasks on Theremin

Introduction

An extensive history of attempts to measure and predict musical aptitude has pervaded the science of Psychology from very early in the field’s development. An attempt to survey all tests of musical aptitude may prove unfeasible, considering that as of 1962, 792 distinct tests of musical aptitude had been compiled, with some presumably used before the advent of recording equipment capable of standardizing procedures (Wing, 1962). Given the vast array of tests available with only minor differences in method, the Psychology of Music has had to make do with a small sample of popular musical batteries developed contemporaneously with the widespread availability of phonograph, gramophone, and record players. These tests include the Bentley Musical Aptitude Test, Drake Music Memory Test, Gordon’s Musical Aptitude Profile, the Karma Auditory Structuring Ability Test, the Kwalwasser-Dykema music tests, the Seashore Measures of Musical Talents, the Wallenton Musical Ear Test, and the Wing Musical Aptitude Test (LeVan, 1937; Law & Zentner, 2012).

Another innovation of these technologically standardized assessments was the use of standardized test stimuli. In the case of the Seashore Measures of Musical Talents, beat frequency oscillators were used to play pure tones at set frequencies, and a prototype tape timing apparatus was built to allow for musical rhythm by opening and closing the frequency circuit through spooling a roll of paper with conductive metal strips attached (Saetveit, 1939). These innovations in standardizing test stimuli rendered obsolete test administrations dependent on the skill of instrumentalists to reliably play test items from standard notation sheet music. This work occurred shortly after historical development of psychometric
technologies of data analysis including the correlation coefficients of Pearson, Spearman, and Kendall (Prokhorov, 2011). Pearson’s correlation coefficient was a common test in the works of both Seashore and Drake.

Some test developers such as Raleigh M. Drake of the Drake Musical Memory Test, criticized Seashore and colleagues’ approach to establishing psychometrics of musical discrimination on the grounds that such measures were deemed inconclusive in predicting musical from nonmusical participants (Treichler, 2013). However, an argument may be made that the initial purpose of psychometrically profiling participants with respect to auditory discrimination may constitute an important subject matter irrespective of application, just as measures of visual acuity need not demonstrate predictive value in determining artists from nonartists to be both reliable and valid at describing participants’ visual capacities. According to Saetveit, Lewis, & Seashore (1940), the authors of the Seashore Measures, the battery was intended to provide “the measurement of aptitude as distinguished from achievement,” defining aptitude as one’s “capacity for discriminating pitch, loudness, duration, and timbre… rhythm and tonal memory,” and defining achievement as one’s “extent of acquired knowledge and skill” in music (Saetveit et al., 1940, pg. 8). Therefore, one may delineate the objectives of the tests on the basis of whether they intend to determine the psychometrics of auditory discrimination in diagnostic tools like the Seashore test, or predict potential for musical achievement in applied tools like the Drake test.

These musical test batteries were developed contemporaneously with Watsonian and Skinnerian behavior science, though the fields appear to have developed in parallel, with little opportunity for collaborative exchange. Nevertheless, these tests may be run and interpreted
without resorting to mentalistic descriptions, and later work did historically bridge music education with the methods and concepts of behavior science from both Skinnerian (e.g. Madsen, Greer, & Madsen, 1976) and Kantorian (e.g. Lundin, 1967) perspectives.

**Seashore Measures of Musical Talent**

The psychometric assessment of musical aptitude in an easily distributed and replicable format was first accomplished by an academic grandchild of Wilhelm Wundt, Carl Seashore, who developed a battery of musical measurements intended to assess basic musical skills in young students. Seashore’s Measurements of Musical Talents were originally intended to be investigated as experimental psychometrics of auditory discrimination rather than as skills tests, though they were certainly used for that purpose. This test differed from earlier applied musical assessments by its explicitly experimental purpose. These included measurements of discrimination for pitch, loudness, rhythm, time, timbre (tone), and recalling tonal memory. The identification of these component skills was predicated on Seashore’s multi-factor theory of musical talent (discussed and investigated by Bugg & Herpel, 1946). Originally, the measurements were tested as a single set (Seashore, 1919). After a variety of critical works evaluated and cited this original set, Seashore (1939) proposed 3 experimental sets intended for different purposes: Set A for standard scoring of musical skills in the general populace, Set B for evaluation of student preparedness for admission to a musical conservatory training environment, and Set C for individual testing of musical skills. The final iteration of the test discarded both Set B and Set C and retained Set A following extensive component analyses assessing the relative difficulties of individual test items to rank order a general test of musical auditory psychometrics (Seashore, Lewis, & Saetveit, 1957). Component analyses of test items and ranking by difficulty may be lacking in other tests of tonal memory such as the Bentley
musical aptitude test (Bentley, 1966), to the extent that the Seashore Measures were preferred by Buckton (1977) when assessing the musical aptitudes of schoolchildren. Buckton’s participants were unable to respond correctly on the Bentley Tonal Memory Test (1966) frequently enough to warrant its inclusion but were able to respond correctly often enough on the Seashore Measures (Seashore, Lewis, & Saetveit, 1957) to allow for determination of normative scores for the young population.

**Drake Musical Memory Test**

The Drake Musical Memory Test (Drake, 1957) appears to place special emphasis on musical memory due to Drake’s dissertation findings that Seashore’s (Seashore, 1939) and Kwalwasser-Dykema’s (Kwalwasser & Dykema, 1930) measures of tonal memory had the highest reliability and predictive validity scores relative to measures of pitch discrimination, rhythm, and other measures. He also showed his measures of musical memory to have superior predictive validity to Seashore’s and Kwalwasser-Dykema’s tests (Treichler, 2013). He referred to this capacity as musical ‘retentivity,’ though a monistic interpretation could simply rephrase it ‘discriminative responding after a latency.’

Drake endorsed his test as highly predictive of musical success, and when independently tested, it did differentiate elite musicians from musical students and nonmusical students, though the original percentile tables were not replicated (e.g. Griffin & Eisenman, 1972). Use of this test in applied musical instruction settings to differentiate student performance conducted many years after the release of the test underscores the potential applied value of the test across settings.
Gordon’s Music Aptitude Profile

The latest contribution to musical aptitude testing has been Edwin Gordon’s Music Aptitude Profile (MAP), which demonstrates good convergent validity in predicting participant performance on other musical tests, such as the Seashore (1957), Wing (1948), Bentley (1966), Karma (1973), and Wallentin (2010) assessments, though its convergent validity with the Drake test is unknown (Law & Zentner, 2012). Gordon (1969) echoed specific critiques of the Seashore Measures’ emphasis on component analysis of musical skills as “atomistic” according to Mursell (1937). He also appeared to initially agree that musical ability constitutes an ‘omnibus’ skill, as posited by the contemporary European approaches to psychology of music. The omnibus view portrays music as a general skill, similar in form to general intelligence and thus irreducible to component abilities (Lowery, 1940; Mainwaring, 1947), though Gordon later refers to this purported “general ability” as “multidimensional” which contradicts this view and nevertheless appears to indicate some potential for component analyses (Gordon, 1999). Gordon initially noted his view that Seashore’s atomistic approach is “unmusically-oriented” (Gordon, 1969).

The MAP purports to correct this alleged shortcoming by investigating “audiation” or “aurally perceiving and then giving meaning to” musical stimuli, which sets the foundation for his main point: “Audiation is the basis of musical aptitude” (Gordon, 1999). The initial MAP subtests included measures of Tonal Imagery, Rhythm Imagery, and Musical Sensitivity. Through a comparative analysis of his measures with the Seashore measures, Gordon (1969) concluded that “the closest relationships between the two batteries are to be found among the MAP Tonal Imagery subtests and the Seashore Pitch and Tonal Memory subtests” as verified through conducting both tests with 157 participants and comparing relative scores with participants’ musical backgrounds and academic performances as measured by ACT (American College
Testing) score (Gordon, 1969). Despite this acknowledgement, Gordon neglected to explicitly include the admittedly important metric of tonal memory in his own test. According to Cuttieta (1991), Gordon’s 2nd test, “has basically discounted the importance of tonal memory as measured and defined in other tests, such as those created by Seashore, Wing, and Drake in favor of the audiation process”.

The discounting of tonal memory may constitute a shortcoming of Gordon’s Musical Aptitude Profile. Considering the large emphasis placed on this concept by Seashore, Drake, and others, and the later empirical finding that tonal memory may be predictive of elite musical performance (Henson & Wyke, 1982), the abandonment of this concept may be premature, and further study on tonal memory is warranted. Another difficulty of the MAP is philosophical. Just as identifying ‘verbal understanding’ as a prerequisite for reading does not offer any clues for exactly how to teach a child to read when they are academically delayed, ‘audiation’ as a concept does not offer any more utility in actually teaching a student to play and understand music better than their baseline levels of ability. This shortcoming arises from the broadness of the proposed concept and its compatibility with nonoperational constructs as well as a tendency to reify ‘audiation’ as a potentially circular cause and effect of musical aptitude. As such, Gordon approaches music from an incompatible perspective with behavior science, wherein component analyses of the skills that comprise reading or playing music are crucial to teaching the composite skills. Consequently, component analysis of musical ability necessarily entails reducing ‘audiation’ into distinct skills that can be assessed and potentially improved through training. This difference, being a philosophical one, may only be exacerbated and not resolved through comparisons of mere data. Interestingly, Gordon’s near contemporary, Raleigh Drake, was unable to find a correlation between measures of musical aptitude (Musical Memory,
Rhythm, and Pitch Discrimination), forcing Drake to abandon the omnibus view of musical talent (Drake, 1954). Despite these concerns, Seashore’s and Drake’s atomistic views are now largely unrepresented in musical assessment and Gordon’s omnibus view has persisted as a popular assessment philosophy, possibly as a result of traditional or cultural factors of the musical academy.

For purposes of study, the Gordon MAP may actually be methodologically redundant compared with the Drake Test. Both of these tests assess a participant’s ability to judge changes in the rhythm, key, and melody line of tonal sequences. Further, both tests demonstrate similar predictive validity with respect to actual musicians (Gordon, 1969; and Griffin & Eisenman, 1972) and both successfully predict the composite skill of musical achievement. Drake (1957) names this process ‘music memory,’ which renders the discrimination only potentially mentalistic, but Gordon (1969) appears to duplicate elements of this method, but gives it the more definitively mentalistic names of ‘audiation’ or ‘tonal imagery.’ These errors in philosophy of concepts are unfortunate, since the operational terms and atomistic suppositions discarded by Gordon may still offer advantages to the music researcher above the omnibus perspective, which does much to obfuscate the contribution of tonal memory to musical learning, and may unnecessarily stymie a component analysis meant to confirm or disconfirm it.

**Study Purpose**

This study will utilize a non-reified, naturalistic perspective, which has certain advantages relative to traditionally mentalistic accounts of musical learning. Among them are 1) a predictive and comparative analysis of musical pitch discrimination skill relative to pitch production skill, which may provide either confirmation or refutation of earlier research.
isolating this skill as a variable correlating with future music learning potential, 2) an understanding of tonal memory as a psychometric measure in a population of students, regardless of applied significance, 3) a conceptual reversal of the omnibus perspective which attributes deficient musical performances of students to unmodifiable qualities of the student themselves rather than to failure of the instructional approach to correct the deficiency, and when deficiency in skill is evident, 4) the atomistic analysis may allow for a more thorough description of which component skills are lacking and suggest possible remedial trainings that may be obfuscated by the non-specific omnibus perspective. To Gordon’s credit, he also appeared to grasp the value of individual education of musical skills when he stated, “there is nothing so unequal as the equal treatment of students of unequal potential” (Gordon, 1999).

In addition, Gordon extensively studied these exact assessment tools and lent support to the hypothesis that tonal memory is potentially predictive of musical success, though he did not pursue this matter further. The objective of this study will not be to refute Gordon as much as to revitalize the psychometrics of basic musical abilities as assessment tools for music education from a different set of operating assumptions. The relative diminution of this area of research is a testament to Gordon’s success, but may overshadow an important and now nearly forgotten tradition of auditory psychometrics that are compatible with naturalistic postulates and incompatible with prevailing mentalistic assumptions that guide contemporary assessment practices. Identification of specific component skills in music learning has the potential to elevate music education to its rightful place in the broader canon of sciences through the discovery of component skills and the refinement of philosophical postulates underscoring current musical assessment and instruction practices.
Research Questions

Extensive correlations were conducted between the Seashore Tonal Memory Test, other Seashore Measures, Oregon Musical Discrimination Test, and the Kwalwasser-Dykema Test of Tonal Movement by Bugg and Herpel (1946) across cohorts of 182, 82 and 38 undergraduate students. Their analyses revealed that “marked superiority in pitch discrimination, timbre discrimination, rhythm discrimination, time discrimination, music appreciation, or the perception of tonal movement rarely occurs in the absence of superior tonal memory” (Bugg & Herpel, 1946, pg. 15). Tonal memory also appears to correlate with musical success, and professional musicians score at the 99th percentile on only this measure compared to the typical population as assessed by the Seashore Measures of Musical Talents (1957) (Henson & Wyke, 1982). However, these demonstrations have never examined whether tonal memory facilitates novel music learning, or whether superior tonal memory is merely an emergent byproduct of other musical skills and training. The following studies will seek to address this question:

Does tonal memory score, as assessed through standardized tests (Seashore Measures of Musical Talent; Drake Musical Memory Test), predict rate of musical learning and mastery of basic musical tasks on a novel musical instrument (the theremin) by undergraduate students recruited from UNR?

Given the specific scope of this research hypothesis, the concerns stated above, the comparatively broad and applied scope of the Gordon test, and its redundancies with the Drake test, which possesses comparable predictive validity (Griffin & Eisenman, 1972; and LeVan, 1937), the historical Seashore and Drake tests will be used for completion of tonal memory
assessment, and the theremin apparatus and learning tasks will be used to demonstrate novel
music learning.

Assessment

Materials

The Tonal Memory Test from the Seashore Measures of Musical Talent (1957) was
administered as well as the Musical Memory Test from the Drake Musical Aptitude Tests (1957).
An experimental packet was developed consisting of an institutional review board approved
Consent Document, Seashore blank sheet, Demographic Information Questionnaire, Drake
practice sheet, Drake blank sheet, and standardized test stimuli sheet for Learning Task B (See
Appendix D). Participants for Study 2 were administered the full packet, while Participants from
Study 1 were administered all items except for the Demographic Information Questionnaire and
standardized test stimuli sheet for Learning Task B, which were added in order to improve
standardization of Study 2. A tape cassette player set to a standard volume was used to play
cassette copies of the Seashore and Drake assessments. The test key for the Seashore Tonal
Memory test was recreated from the standard notation notes included in the appendix of
Saetveit, Lewis, & Seashore’s (1940) test revisions. The complete answer key for the other
unadministered Seashore measures was irretrievable through interlibrary loan over 6 months
and repeated requests. The Drake test key was included with the Drake Musical Aptitude Tests
LP recording. The Rhythm section of the Drake tests was not administered.

Procedures

Assessment sessions were scheduled as Part 1 of the study. Administration of tonal
memory tests (Part 1) was identical between Study 1 and Study 2 participants, though there
were differences in standardization of the demographic questionnaire and Part 2 stimuli between Study 1 and Study 2. The Part 1 session required 45-60 minutes total and a maximum of five participants at a time were allowed to undertake Part 1 due to space constraints. The majority of appointments were arranged through SONA for course credit in UNR Psychology undergraduate courses. Approximately 10 participants were recruited directly from the UNR campus community through flyers with the expressed disclaimer that no incentive in pay or course credit could be provided for their participation beyond a better understanding of their tonal memory.

- Participants were first administered the Consent Document and were given as much time as desired to consider participation and sign. Next, participants were asked to verify their scheduled Part 2 dates and times and write these on the top page of their data packet to aid in organizing follow up appointments.

- The Seashore Tonal Memory Test script was read from the manual and one non-included example item was played on a computer keyboard.

- The Seashore Tonal Memory Test was administered over 15 minutes.

- A two minute standardized rest period was provided, during which demographic information was solicited through in person interview for Study 1 participants, and on the Demographic Information Questionnaire for Study 2 participants, which was folded by each participant and unexamined until after all participants’ Part 2 data was entered to blind the experimenter to participant musical history. Only the tonal memory scores were blinded for Study 1.

- The Drake Musical Memory Test script was read from the manual and the practice sheet was completed over three minutes.
• The Drake Musical Memory Test, Form A was administered over 15 minutes.
• Another two minute standardized rest period was provided.
• The Drake Musical Memory Test, Form B was administered over 15 minutes.
• Part 1 participation was concluded.

Data Collection and Analysis

All data was initially collected on paper then entered into Microsoft Excel 2010. Analyses and figures were generated either with Microsoft Excel 2010 or ‘R,’ and ‘R Studio,’ which are free, open source statistics software (R Core Team, 2018).

Results

Figure 1 shows the distribution of Seashore Scores across the sample of 61 participants from Study 1 and Study 2. A Shapiro-Wilk test rejected the assumption of normality for the distribution ($W = .84, p < .001, M (61) = 24.9, SD = 5.43$).
Figure 1. Overall distribution of Seashore Tonal Memory scores across 61 participants from Study 1 and Study 2

Figure 2 shows the distribution of Drake Scores across the sample of 61 participants from Study 1 and Study 2. A Shapiro-Wilk test was not able to disconfirm the assumption of normality ($W = .98, p = .55, ns, M (61) = 65.41, SD = 14.6$), hence the Drake test distribution may be assumed normal.
Figure 2. Distribution of Drake Musical Memory Form A and Form B combined scores across 61 participants from Study 1 and Study 2

Figure 3 shows the nonparametric correlation between Seashore and Drake scores across 61 participants from Study 1 and Study 2. A Kendall Rank Correlation Test showed that scores on the Seashore test significantly predicted scores on the Drake test ($r_s = .56, p < .001$). A LOESS (local regression) line of best fit was calculated and added as a nonparametric indicator of trend (i.e. a line that doesn’t assume a linear relationship or normal distribution of data, given the non-normality of the Seashore distribution, see Prabhakaran, 2017).
**Figure 3.** Nonparametric correlation between Seashore and Drake scores across 61 participants from Study 1 and Study 2 with LOESS line

**Demographic Comparisons**

For gender, scores were compared using an independent samples t-test due to the normality of the residuals (i.e. the differences between observed scores and overall mean score, as described by Rochon, Gondan, & Kieser, 2012) \( W = .96, p = .29, ns \) and the equal variances of the samples verified through a Levene test \( F(1,59) = .02, p = .88, ns \). The t-test showed that composite scores (i.e. Seashore score + Drake score) differed significantly across gender,
such that female participants ($M (31) = 94.94, SD = 17.67$) scored significantly higher than male participants ($M (30) = 85.53, SD = 18.48$) in overall composite score ($t(58.64) = 2.03, p = .047, ns$). A follow up Mann-Whitney test evaluating only Seashore score across gender found no significant relationship ($U = 590, p = .07, ns$), and neither did a follow up independent samples t-test comparing Drake scores across gender ($t(58.58) = 1.96, p = .055, ns$).

For extent of musical instruction, scores were compared using an independent samples t-test due to the normality of residuals ($W = .96, p = .40, ns$) and the equal variances of the samples verified through a Levene test ($F(1,59) = 1.04, p = .31, ns$). The t-test showed that composite scores differed significantly with musical training, such that musically trained participants of three years or more ($M (32) = 99.12, SD = 15.01$) scored significantly higher than participants with less than 3 years of musical training ($M (29) = 80.59, SD = 17.33$) in overall composite score ($t(55.74) = -4.45, p < .001, s$).

For age, a Kendall rank correlation determined no significant relation between age and composite score ($r_t (59) = .04, p = .71, ns$).

**Discussion**

**Seashore Tonal Memory Test.** The Seashore distribution of scores demonstrates a skew toward high average scores in the population sampled by both Study 1 and Study 2 to the extent that it constituted a ceiling on test precision. This problem might be expected with any assessment designed to detect discrimination capacity of typically lower performing individuals.
such as young children. Since the original Seashore test was predominately administered to school aged children with only limited study of college aged participants, it may be surmised that this shortcoming limits the descriptive validity of the Seashore Tonal Memory Test for the purpose of predicting musical learning in a general population. Of course, this limitation was uncovered inductively over the course of test administration, though there is evidence Seashore and colleagues may have quietly surmised this limitation as well, since their own norm tables for college aged participants in the revised test manual similarly reveals a decided skew (Seashore, Lewis, & Saetveit, 1957).

By way of improvement, it may be suggested that if an additional trial block of test stimuli were appended to the original test, it might have allowed for the completion of the high range of the bell curve through the addition of more difficult test items. The Seashore Tonal Memory test (1957) only gets as difficult as five tone comparisons, so an additional 10 trial blocks of six tone comparisons, or even further extension to seven tone comparisons, would likely allow for the generation of a normally distributed sample and provide a test more representative of tonal memory performance by older participants while requiring only a slight extension of the test to 40 or 50 items. Future development of this assessment might extend the test items in this way as well as rerecord the test items in digital form appropriate for online test administration with automated data entry which could facilitate much larger sample sizes in future administrations as well as preserve this important method of tonal memory measurement for future research. However, it must be considered that the test items selected by the 1940 test revision were not selected arbitrarily but rather were subject to extensive component analyses of test item difficulty. This process would need to be repeated with hundreds of modern participants for any extension of this test, which would entail pilot testing
and revisions of the test materials to the standards of the original authors, which would be no small endeavor.

**Drake Musical Memory Test.** The Drake distribution of scores revealed a nearly perfect bell curve distribution of scores ranging from 36 to 97 out of 108 test items. Neither a floor nor ceiling effect on participant performance was evident. Thus, the Drake Musical Memory Test appears to be a superior psychometric test to the Seashore Tonal Memory Test for adult participants and may be more appropriate for a precise description of musical memory capacity in the general population. This finding is ironic, since Drake’s Musical Memory Test was not primarily intended to function as a psychometric test, but as an applied screening tool. These results reveal a distribution of scores similar to the norm scores for nonmusical participants in the original Drake test, which is a further replication of historical psychological findings.

**Correlation between Seashore and Drake tests.** The Kendall Rank Correlation revealed a highly significant correlation between the Seashore Tonal Memory Test and the Drake Musical Memory Test. It should be noted that the correlation is nonlinear and shows a distinct upward skew explainable by the ceiling effect on the Seashore test (i.e. since no participant can score higher than 30/30, high scores on the Drake test are bunched together at the Seashore top score). This finding indicates that both assessments likely measure similar capacities, such that score on one metric significantly predicts score on the other. This finding replicates Drake’s (1933) historical finding that his Musical Aptitude Test correlates with the Seashore Measures of Musical Talents. Given the acknowledged replication crisis in the field of psychology, this current replication is particularly timely as a demonstration of these psychometric tools’ reliability (Diener & Biswas-Diener, 2019).
Comparison across Demographic Categories. Analyses of the composite Seashore and Drake test scores compared across gender, musical experience, and age revealed significant effects of gender and musical experience on overall composite score and no significant relationship between age and overall composite score.

The observed advantage for female over male participants in their composite scores of Seashore and Drake tests supports the findings of Miles, Miranda, & Ullman (2016) who found that female participants scored higher at recognizing familiar melodies than males even when controlling for previous musical training. Miles and colleagues attributed this to a more general female advantage in declarative memory and deduced that musical recognition might therefore be a verbal process.

Finally, the higher observed scores of those trained musically for 3 or more years supported the findings of Seashore (1957) and Drake (1957) that those with musical training exhibit higher mean scores than the musically untrained.

Study 1

Introduction

Music Research in Behavior Science. Behavior science has a modest history of studies attempting to teach musical skills. Early studies often took the form of consultative work in music education or experimentation with basic music discrimination skills (e.g. Madsen et al., 1976). Studies which attempted to directly teach musical skills include the study of Williams and Engelmann (1989) who used a Direct Instruction methodology to teach absolute pitch to a cohort of 11 trained elementary school participants, who were shown to outperform a control group of 15 untrained high school aged participants. This study assessed both pitch
discrimination and pitch production. The authors were limited however, by an acknowledged lack of a standardized assessment of musical learning aptitude. Therefore, they could conclude Direct Instruction was an effective method of acquiring musical responses in 11 out of 18 students recruited, but could not account for individual differences across subjects in absolute pitch acquisition or predict which students were likely to succeed in the task. This oversight may have accounted for 7 participants out of 18 being initially removed from the study following failure to learn from Direct Instruction. More recent work includes a study by Snyder (2004) who demonstrated that absolute pitch accuracy within one whole tone of the original reference tone could be maintained across a delay of 24 hours with two adult female participants. Lastly, but not exhaustively, contemporary work by Domjan (2016) demonstrates how behavior analytic principles may be applied to traditional music instruction in his Tertis/Pavlov Project.

This literature demonstrates that Direct Instruction has shown some success in teaching absolute pitch and also that behavior analytic approaches to music education have a longstanding and continuing precedent. The teaching progression to be introduced (See Learning Tasks A and B) was derived from a modified multiple baseline with probe research design that constitutes a novel extension of instructional design derived from behavior science to the application of musical learning (Reynolds & Hayes, 2015). The method allows for a demonstration of which learning tasks result in probe performance improvement across participants, and allows for the researcher to identify not only how well skills are learned, but when those skills are initially acquired, and whether skilled performances are retained over successive trainings. This method combined with the extensive tonal and musical memory pre-assessments allows both the evaluation of the teaching method as well as a potential explanation for students who demonstrate difficulty in progressing through the learning tasks.
Potential Contributions to Music Education. Research has demonstrated that elite musical performance predicts high tonal memory (e.g. Henson & Wyke, 1982; et al.), but no study has yet attempted to discern whether tonal memory ability predicts novel musical learning. This first study will undertake demonstration of musical learning with the theremin as a novel apparatus that is unfamiliar to most participants, even those with a history of musical instruction. Playing the theremin is not commonly used as a novel musical skill in experimental settings, and the use of this apparatus has advantages in an experimental setting over singing or playing instruments such as customization or standardization of the playing field, an inconsistent root tone which renders proprioceptive feedback unreliable as the instrument warms up, simple gross motor behaviors used in playing tones, and a nonstrenuous playing technique that can be sustained with minimal fatigue over extended sessions. The use of this device improves on the use of singing from the early Master’s thesis iteration of this Direct Instruction method (Reynolds & Hayes, 2015). This study will demonstrate whether tonal memory grants an advantage for music learning, which is a novel contribution to music education research.

Participants

Sixteen participants (9 female, 7 male, aged 18-35, $M = 20.13, SD = 4.13$) were recruited through the Psychology department experimental sign-up system (SONA) at University of Nevada, Reno and completed the Seashore and Drake tonal memory assessments as well as a demographic questionnaire. Of this group, all participants completed learning tasks on the theremin. All participants received course credit for their time with the exception of 3-5 students who volunteered for no credit. All procedures and materials used with participants were granted exempt status by the UNR Institutional Review Board.
Materials and Apparatus

For Learning Tasks A and B, a Burns B3 deluxe theremin, guitar cable, Danelectro Honeytone N-10 mini amplifier, and 9 volt adapters were used for tone production. The VoceVista Video Pro Software was used for real time visual feedback of musical performance and audio recordings of sessions. Scrap paper was used to cover the screen during probe trials, randomize number sequences, and track conditions.

Procedures

Tonal Memory Assessment. Both the Seashore and Drake tonal memory assessments were administered as described previously (See Assessment: Procedures). All tonal memory assessment data were handled in a double blind administration. Neither the experimenter nor the participants were made aware of the tonal memory scores prior to the learning tasks in Part 2 of the experiment.

Learning Tasks: Preliminary Equipment Operating Procedures. The participants sat adjacent to the experimenter so they could see the visual feedback display. The root tone of the theremin was initially standardized to G#/Ab4 on a keyboard within the program, which played a pure sine tone when a note was selected and the left mouse key depressed. The digital keyboard also produced auditory feedback during learning task steps requiring it (see Study 1: Learning Task A and Learning Task B). The sine tone was identical in timbre to the theremin tone produced by the amplifier. The participant was asked to practice the following requests prior to beginning the learning tasks: mute, play, higher, and lower. When the participant demonstrated correct responding at least twice on each request, they were eligible to begin the
study and the learning tasks began. The end of this phase also marked the beginning of the 60 minute time limit of the study.

**Learning Task A: Playing C Major Scale (See Figure 1).** Probe A consisted of asking the participant to play the 8 notes of the C major scale by number (C, D, E, F, G, A, B, C) without auditory or corrective feedback (The experimenter said “play 1”, the participant attempted to play C for 3 seconds, the experimenter said “mute”, the experimenter said “play 2”, the participant attempted to play D for 3 seconds, the experimenter said “mute”, etc.). There was no accuracy criterion due to this being a probe conducted before and between learning trials to test for acquisition of the skill. No prompts were given during these trials and the screen was covered. This probe was presented before Learning Task A Level 1, between all levels, and after mastery of Level 5.

In all Learning Task A levels, if the participant matched a tone within 2 seconds of the prompt, the trial was correct. If the participant required more than 2 seconds matching a tone, the trial was incorrect. If 6 out of 8 trials were correct, the participant advanced to the next step after a probe trial. If less than 6 out of 8 trials were correct, the participant repeated the current step until they scored 6 out of 8 correct, or study time was exhausted.

In Learning Task A level 1, the participant was asked to play the C major scale with 1 whole tone allowed variability with an auditory tone present during playing (“play 1”, C auditory tone was played while the participant attempted to play C, the experimenter said “mute” after the participant matched within 1 whole step of C for 2 seconds, “play 2”, D auditory tone was played while the participant attempted to play D, the experimenter said “mute” after the participant matched within 1 whole step of D for 2 seconds, etc.).
In Learning Task A Levels 2 and 3, the participant was asked to play the C major scale with ½ tone allowed variability (Level 2) or ¼ tone allowed variability (Level 3) with an auditory tone present during playing. Instructions were identical in these steps to Level 1.

In Learning Task A level 4, the participant was asked to play the C major scale with ¼ tone allowed variability with an auditory tone delayed by 2 seconds after the participant began playing (“play 1”, the participant attempted to play C, the experimenter played a C tone 2 seconds after the participant began playing, the experimenter said “mute” after the participant matched within ¼ step of C for 2 seconds, “play 2”, the participant attempted to play D, the experimenter played a D tone 2 seconds after the participant began playing, the experimenter said “mute” after the participant matched within ¼ step of D for 2 seconds, etc.).

In Learning Task A level 5, the experimenter played a C reference tone (note 1) before the participant was asked to play the C major scale with ¼ tone allowed variability with an auditory tone delayed by 2 seconds. Instructions were identical to Level 4 with the exception that the C tone is played as a reference tone prior to prompting the participant to play.

Regardless of participant progress, they were asked to conclude participation after 60 minutes, or completion of final probe B, whichever came first.

**Learning Task B: Playing Random Notes in C Major** (See Figure 1). After achieving final probe A, participants were inducted into Learning Task B and began Probe B.

Probe B consisted of asking the participant to play 8 random nonrepeating notes in C major without auditory or corrective feedback (The experimenter said “play 4”, the participant attempted to play F for 3 seconds, the experimenter said “mute”, the experimenter said “play 2”, the participant attempted to play D for 3 seconds, etc.). There was no accuracy criterion due
to this being a probe conducted between learning trials to test for acquisition of the skill. The first trial block of Probe B was conducted immediately after final probe A before Learning Task B Level 1, between Levels 2 and 3, and after passing the Level 3 progression criterion.

In all Learning Task B levels, if the participant matched a tone within 2 seconds of the prompt, the trial was correct. If the participant required more than 2 seconds matching a tone, the trial was incorrect. If 6 out of 8 trials were correct, the participant advanced to the next step after a probe trial. If less than 6 out of 8 trials were correct, the participant repeated that step until they scored 6 out of 8 correct, or study time was exhausted.

In Learning Task B level 1, the experimenter played a C reference tone before the participant was asked to play 8 random nonrepeating notes in C major with ¼ tone allowed variability with an auditory unison tone delayed by 2 seconds. This step is essentially Learning Task A Level 5 with the note order randomized (The experimenter played a C reference tone, said “play 4”, the participant attempted to play F, the experimenter played an F tone 2 seconds after the participant began playing, the experimenter said “mute” after the participant matched within ¼ step of F for 2 seconds, the experimenter played a C reference tone, said “play 3”, the participant attempted to play E, the experimenter played an E tone 2 seconds after the participant began playing, the experimenter said “mute” after the participant matched within ¼ step of E for 2 seconds, etc.).

In Learning Task B level 2, the experimenter played a C reference tone before the participant was asked to play 8 random nonrepeating notes in C major with ¼ tone allowed variability. Level 2 was the same as Level 1 with the exception that no feedback auditory tone
was provided. The participant was provided only visual feedback to verify correct tones (i.e. the antecedent C reference tone was played, but no feedback tone).

In Learning Task B level 3, the participant was asked to play 8 random nonrepeating notes in C major with ¼ tone allowed variability. Level 3 discontinued all auditory feedback. Only the visual feedback display was provided.
Inter-observer agreement and procedural integrity. Out of the 16 Learning Task sessions included in Study 1, 37.5% (≥ 33.33%) of sessions (6 sessions) were reviewed by a research assistant, who independently reviewed video recordings and software audio/visual recordings to code learning task data and review procedural integrity adherence.
Inter-observer agreement (IOA) was calculated for each session according to the following formula:

\[
\frac{\text{# of agreements}}{\text{total # of agreements} + \text{disagreements}} \times 100 = \text{IOA}\%
\]

Procedural integrity (PI) data was coded using a procedural integrity data sheet in conjunction with session video recordings. Procedural integrity was calculated according to the following formula:

\[
\frac{\text{# of adhered items}}{\text{total # of data sheet items}} \times 100 = \text{PI}\%
\]

Calculated IOA averaged 92.44% across six sessions (ranging from 89.60% to 96.78%).

Calculated PI averaged 96.35% across six sessions (ranging from 92.68% to 100%).

**Study 1 Results**

**Correlations**

Comparisons between tonal memory scores and measures of performance during the learning tasks demonstrated the relation between these variables. First, data distributions were assessed with the Shapiro-Wilk test for normality to assess whether the parametric Pearson or nonparametric Kendall correlation coefficient was most appropriate for given comparisons. The Shapiro-Wilk test results also informed the choice between a linear visual model or LOESS (local regression) line for visualizing trends, where applicable (See Table 1 for Shapiro-Wilk and summary results).
Table 1

*Descriptive Statistics and Shapiro-Wilk Tests of Normality across Variables (Study 1)*

<table>
<thead>
<tr>
<th>Variables (N = 16, unless otherwise specified)</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>W</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seashore Score (out of 30)***</td>
<td>26.13</td>
<td>5.04</td>
<td>.76</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Drake A+B Score (out of 108)</td>
<td>69.19</td>
<td>14.41</td>
<td>.96</td>
<td>.71</td>
</tr>
<tr>
<td>Duration to Final Probe A (in minutes)</td>
<td>15.39</td>
<td>4.40</td>
<td>.95</td>
<td>.48</td>
</tr>
<tr>
<td>Total Trial Blocks to Final Probe A**</td>
<td>11.44</td>
<td>3.88</td>
<td>.84</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Probe A Average</td>
<td>2.21</td>
<td>1.10</td>
<td>.93</td>
<td>.27</td>
</tr>
<tr>
<td>Probe B Average</td>
<td>3.08</td>
<td>1.02</td>
<td>.96</td>
<td>.65</td>
</tr>
<tr>
<td>Overall Probe Average</td>
<td>2.55</td>
<td>0.94</td>
<td>.93</td>
<td>.28</td>
</tr>
<tr>
<td>Duration to Final Probe B (n = 12)</td>
<td>15.17</td>
<td>3.93</td>
<td>.93</td>
<td>.36</td>
</tr>
<tr>
<td>Additional Trial Blocks to Final Probe B (n = 12)</td>
<td>7.58</td>
<td>3.45</td>
<td>.89</td>
<td>.09</td>
</tr>
<tr>
<td>Top Score Probe A</td>
<td>4.69</td>
<td>1.89</td>
<td>.93</td>
<td>.24</td>
</tr>
<tr>
<td>Top Score Probe B</td>
<td>4.56</td>
<td>1.09</td>
<td>.90</td>
<td>.07</td>
</tr>
<tr>
<td>Highest Instructional Level Attained (out of 9)***</td>
<td>8.56</td>
<td>0.81</td>
<td>.56</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Level Before Highest Probe Score</td>
<td>4.44</td>
<td>2.31</td>
<td>.94</td>
<td>.31</td>
</tr>
</tbody>
</table>

* α is ≤ .05 (i.e. distribution is assumed non-normal)

** α is ≤ .01 (i.e. distribution is assumed non-normal)

*** α is ≤ .001 (i.e. distribution is assumed non-normal)
Seashore tonal memory score did not significantly predict duration to final probe A $(r_c(14) = -.16, p = .43, ns)$, nor total trial blocks to final probe A $(r_c(14) = .03, p = .89, ns)$, nor probe A average performance $(r_c(14) = .15, p = .43, ns)$, nor probe B average performance $(r_c(14) = .37, p = .06, ns)$, nor overall probe average $(r_c(14) = .27, p = .17, ns)$, nor duration to final probe B mastery $(r_c(10) = .06, p = .80, ns)$, nor additional trial blocks to final probe B $(r_c(10) = 0, p = 1, ns)$, nor top score for probe A $(r_c(14) = 0, p = 1, ns)$, nor top score for probe B $(r_c(14) = .25, p = .23, ns)$, nor highest instructional level attained $(r_c(14) = .38, p = .09, ns)$, nor level before highest probe score $(r_c(14) = .05, p = .82, ns)$. Kendall Rank Correlation was used for all comparisons due to the non-normal distribution of Seashore scores. No measures of interest from Learning Tasks A or B were significantly predicted by the Seashore tonal memory scores of Study 1 participants.

Drake musical memory test score did not significantly predict duration to final probe A $(r(14) = -.39, p = .13, ns)$, nor total trial blocks to final probe A $(r_c(14) = -.06, p = .75, ns)$, nor probe B average $(r(14) = .33, p = .21, ns)$, nor duration to final probe B $(r(10) = .23, p = .46, ns)$, nor additional trial blocks to final probe B $(r(10) = .23, p = .46, ns)$, nor top score for probe A $(r(14) = .31, p = .24, ns)$, nor top score for probe B $(r(14) = .23, p = .40, ns)$, nor highest instructional level attained $(r_c(14) = .19, p = .37, ns)$, nor level before highest probe score $(r(14) = -.30, p = .26, ns)$. All comparisons were assessed using either Pearson’s Product-Moment Correlation ($r$) for normal distributions or Kendall’s Rank Correlation ($r_c$) for non-normal distributions.
However, Drake test score did significantly predict probe A average 

\( r(14) = .57, p = .02, s\), see Figure 5) and overall probe average \( r(14) = .54, p = .03, s\), see Figure 6).

\[ \text{Figure 5. Linear correlation between Drake Musical Memory A+B score and probe A average performance (Study 1)} \]
Figure 6. Linear correlation between Drake Musical Memory A+B score and overall probe average performance (Study 1)

Sample Participant Performances

In order to illustrate typical learning trajectories of participants, six sample participants are shown, corresponding with the top two, middle two, and bottom two performers out of the entire Study 1 sample, as assessed through overall probe average scores. These learning trajectories are shown alongside both Seashore and Drake scores to demonstrate whether individual results validate the correlations shown above.
The top two performers in overall probe average (participants 19 and 21) demonstrated immediate performance improvements, followed by maintenance through Learning Task A, in the case of participant 19, and a decline of accuracy through Learning Task A in the case of participant 21. Transition to Learning Task B precipitated a loss of accuracy and accelerated learning in participant 19, while participant 21 exhibited a similar asymptotic increase followed by a decrease in Learning Task B. The results were consistent with the scores of each participant on the Seashore and Drake tests (See Figure 7).

The two performers with overall probe average scores closest to the sample average (participants 8 and 12) demonstrated a modest increase in accuracy early in Learning Task A, followed by a decrease in accuracy toward the end of the task. Both performers exhibited an acquisition of Learning Task B accuracy, with both exhibiting a higher average score during Learning Task B than Learning Task A. Seashore and Drake scores were within the expected range for participant 8, however participant 12 scored the lowest on the Drake test of all sample participants, which contradicts the general correlation exhibited by the study sample (See Figure 8).

The two performers with the lowest overall probe average scores (participants 20 and 14) demonstrated initial minor increases in accurate responding which declined through Learning Task A. Participant 20 demonstrated a steady increase in accuracy throughout Learning Task B. Participant 14 demonstrated low scores throughout session, and with the exception of the outlier score of participant 12, these two scored lower than the other sample participants on the Drake test (See Figure 9).
Seashore Score out of 30 | Drake Score out of 108
--- | ---
30 | 93

Seashore Score out of 30 | Drake Score out of 108
--- | ---
29 | 86

*Figure 7.* Top two study 1 participants’ (19 and 21) overall average probe scores (99\textsuperscript{th} and 82\textsuperscript{nd} percentiles) along with Seashore and Drake scores
Figure 8. Middle two study 1 participants’ (8 and 12) overall average probe scores (both 57th percentile) along with Seashore and Drake scores
Figure 9. Bottom two study 1 participants’ (20 and 14) overall average probe scores (11th and 9th percentiles) along with Seashore and Drake scores
**Study 1 Discussion**

**Descriptive Data**

The Shapiro-Wilk tests of normality revealed that all variables may be assumed normally distributed with the exception of the Seashore test results, total trial blocks to final probe A, and highest instructional level attained (out of 9). Those particular variables could not be assumed to vary continuously, and the nonlinear Kendall correlation was most appropriate for comparing them to other variables. All other variables of interest could be assumed to vary continuously such that the Pearson correlation was an appropriate test for their comparisons.

**Correlations**

The Seashore assessment was not shown to predict performance on any variables of interest during Learning Tasks A and B. This indicates that the test may insufficiently differentiate music learners of limited capacity from those of greater capacity. This may owe to the aforementioned ceiling effect on scores, such that the test is insufficiently difficult to profile the tonal memory of college aged participants. This may also be an artifact of limitations in Study 1 that are corrected in Study 2 (See Limitations). It should be noted that this finding only pertains to the tonal memory section of the Seashore test, which consists of other tests as well that were not evaluated in these studies.

The Drake assessment, by contrast, was significantly predictive of overall probe average score and probe A average score, which validates the use of the Drake assessment for its original purpose of predicting likelihood of musical learning potential. Inability of the assessment to predict time to complete Learning Tasks A and B or highest instructional level attained may owe to limitations in Study 1 design that are corrected in Study 2 (see Limitations), or be indicative of
greater predictive validity for accuracy than learning speed or overall progress. Top scores were also not predictable by Drake score, which may indicate that this assessment cannot predict which participants are capable of infrequent instances of high accuracy, a phenomenon which was sometimes observed, even among low scorers.

Discussion of Sample Participants’ Results

Five out of six of the sample participants demonstrated a reduction of accuracy through Learning Task A following an initial increase in accuracy. This may constitute behavioral habituation, “in which familiar stimuli lead to a reduction in responsiveness” (Huron, 2013, pg. 8). Interestingly, 13 out of 16 participants actually scored higher on average during Learning Task B, the higher difficulty task, than Learning Task A. This may indicate that the variability introduced through random presentation of notes 1-8 dishabituated probe performance, “defined as the recovery of responsiveness due to… intervening novel [stimuli]” (Huron, 2013, pg. 11). The unpredictable variability of the Learning Task B trial block stimuli evoked higher performance in Learning Task B. However, some reduction in responding accuracy nevertheless occurred within Learning Task B for 9 out of 16 participants, indicating that the dishabituating stimuli tended to lose their effectiveness across repeated trial blocks.

The sample of 6 participants demonstrated contradictions between the correlation results and individual performances. For example, participant 12 scored at the 57th percentile on overall probe average score, but scored lower than the bottom two participants on the Drake test. This indicates that although a significant correlation exists between performance on the Drake Musical Memory Test and overall probe average score, the test may not be individually predictive in every instance. It is recommended on that basis to utilize the Drake test or similar
psychometric tools in conjunction with direct performance based measures such as Learning Tasks A and B implemented in these studies to make fully informed placements of students in musical instruction.

Limitations

Study 1 included only 16 participants for comparative data which was cut from 22 total recruited participants. This was due to participants 1-3 being shown visual stimuli for probe trials, which was recognized as a procedural inconsistency and necessitated the removal of their data from analysis. Participant 7 was removed since they scored the lowest on both the Seashore and Drake assessments, yet scored among the highest on the direct Learning Tasks, constituting an extreme outlier, likely as a result of poor attention during the assessment batteries. Participant 22 was run to replace Participant 7’s contribution. Participants 10 and 17 were not included since they did not attend Part 2 of the experiment and therefore had no Learning Task data for analysis. The high proportion of lost data indicates that standardization of experimental procedures and protocols for encouraging repeat attendance are important for refinement of this project, in addition to recruitment of a greater number of participants to buffer against the effects of outliers.

Probe stimuli were also not standardized by latency and tone of voice, which might have subtly changed these administrations across participants and led to situations where tone of voice and latency between stimuli might have been modulated in such a way as to influence accuracy. Additionally, the randomized tones for Learning Task B varied across participants, which may have inadvertently created more or less difficult sequences of stimuli from one participant to the next. To address these problems, standardized probe prompts from
recordings, longer allowed latencies for Learning Task A Levels 4/5 and Learning Task B Levels 6/7, and standardized stimulus sequences for Learning Task B are necessary for further revision.

Though the tests were administrated in a double blind fashion, such that neither the Seashore nor Drake assessments were scored prior to completion of the Learning Tasks and their data entry, the demographic information of participants was not blinded to experimenters, which may have unwittingly left a potential source of bias. This indicates that blinding of demographic information and continued double blinding of test scores will be important for future revision.

The visual feedback display may have constituted an important confound in measuring variables of interest with fidelity. Because the waveform of produced tones was visually perceptible, some proportion of participants may have matched visual stimuli to progress further or faster than their tonal memory in isolation would allow. This would have the effect of progressing participants further on the basis of visual discrimination rather than tonal memory alone. Hence, further revision of this study will remove all visual stimuli from both Learning Tasks and probe conditions, which will also entail a removal of Learning Task B Step 3, which used exclusively visual feedback. Also, the pause before feedback provided in Task A, Levels 4-5 and Task B, Levels 1-3 will be lengthened to 4 seconds to allow enough time for participants to find the correct notes and prevent false scoring of incorrect responses. The revised method for Study 2 is portrayed below (Figure 10).

Selection bias and participant expectancy may have constituted other potential confounds, in that nearly every participant in Study 1 approached the experimenter to volunteer participation, citing personal interest in music or the theremin. This may have biased
participation toward those who had a greater affinity for music or interest in the theremin, and such participants might therefore be more likely to have acquired musical skill than randomly sampled participants. Further revision will feature recruitment practices designed to actively recruit participants who may or may not have any affinity for music or interest in the theremin.

**Study 2**

**Introduction**

Study 2 improved on the materials, recruitment, procedures, and methods of Study 1 to increase the number of participants enrolled, further standardize procedures, blind the experimenters to both assessment scores and demographic information, remove visual feedback from all learning tasks, and proactively recruit a disinterested sample of undergraduate participants for course credit administered through the UNR SONA system of credit. All assessment procedures were replicated from Study 1 except demographic information collection was double blinded, and both learning tasks were revised to remove visual feedback (See Figure 10).

Unlike Study 1, in which participants were passively enrolled after they demonstrated interest in the study, participants in Study 2 were recruited in greater numbers due to proactive advertisement of the study to participants in UNR Psychology classes as well as through direct approach of UNR students on campus to solicit participation with or without course credit. This increased the number of correlations from 16 to 37. Only IRB approved advertisement materials, methods, and consent documents were utilized.

Standardization of Study 2 was improved by playing probe trials from recordings with standardized latencies between prompts accomplished by timing instructions with an inaudible
60 bpm metronome. Additionally, the random number sequences presented to participants in Learning Task B were standardized to prevent varying difficulties of sequences across participants.

The double blind procedure of Study 1 was maintained and demographic information was also blinded by asking participants to complete their information on a folded sheet of paper which was not opened until after both Learning Tasks were completed and data coded and entered. This prevented the experimenters from having any information on participant musical history or tonal memory prior to administration of the Learning Tasks.

Visual feedback was removed from the Learning Tasks to control for the possibility of participants using visual feedback to progress faster through the Learning Tasks than their tonal memory alone might allow. Also, Task A, Levels 4-5 and Task B, Levels 1-2 increased the latency between participant attempt and feedback from 2 seconds to 4 seconds, to allow the participant more time to find notes and prevent the possibility of falsely marking incorrect responses.

Lastly, both selection bias and subject expectancy of participants who may have taken a personal interest in the experiment during Study 1 were controlled during Study 2 by advertising to participants without explicitly mentioning the theremin (the experiment was referred to simply as a ‘tonal memory study’ without mentioning the theremin) and actively seeking out participants from a disinterested population of students seeking credit for Psychology courses.

Participants

Thirty-seven participants (17 female, 20 male, aged 18-47, $M = 20.49, SD = 4.78$) were recruited through the Psychology department experimental sign-up system at University of Nevada, Reno. All completed the Seashore and Drake tonal memory assessments as well as a
demographic questionnaire. Of this group, all participants completed learning tasks on the theremin. All participants received course credit for their time. All procedures and materials used with participants were granted exempt status by the UNR Institutional Review Board. Two female participants were recruited who did not attend part 2 of the experiment and could not be contacted for follow up appointments. Their assessment data was included (See Assessment), though their lack of Learning Task data precluded them from inclusion in Study 2 data analyses.

Materials and Apparatus

For Learning Tasks A and B, a Burns B3 deluxe theremin, guitar cable, Danelectro Honeytone N-10 mini amplifier, and 9 volt adapters were used for tone production. The VoceVista Video Pro Software was used for real time visual analysis of musical performance and audio recordings of sessions. Visual feedback was unavailable to participants at all times.

Procedures

Tonal Memory Assessment. Both the Seashore and Drake tonal memory assessments were administered as described previously (See Assessment: Procedures). The one exception to the previous method was the use of a demographic information questionnaire (See Appendix C) which was folded and unexamined by experimenters until after final data for the learning tasks were coded to facilitate a blind administration. All tonal memory assessment data continued to be handled according to a double blind administration. Neither the experimenters nor the participants were made aware of tonal memory scores prior to the learning tasks in Part 2 of the experiment.
Learning Tasks: Preliminary Equipment Operating Procedures. The participants sat opposite to the experimenter to prevent viewing of the visual feedback display. The root tone of the theremin was standardized to G#/Ab4 on a keyboard within the program, which played a pure sinusoid tone when a note was selected and the left mouse key depressed. The digital keyboard also produced auditory feedback during learning task steps requiring it (see Study 2: Learning Task A and Learning Task B). The sinusoid tone was identical in timbre to the theremin tone produced by the amplifier. The participant was asked to practice the following requests prior to beginning the learning tasks: mute, play, higher, and lower. When the participant demonstrated correct responding at least twice on each request, they were eligible to begin the study and the Learning Tasks began. The end of this phase also marked the beginning of the 60 minute time limit of the study.

Learning Task A: Playing C Major Scale (See Figure 10). Probe A consisted of playing a standardized recording asking the participant to play the 8 notes of the C major scale by number (C, D, E, F, G, A, B, C) without auditory or corrective feedback (The experimenter said “play 1”, the participant attempted to play C for 3 seconds, the experimenter said “mute”, the experimenter said “play 2”, the participant attempted to play D for 3 seconds, the experimenter said “mute”, etc.). There was no accuracy criterion due to this being a probe conducted before and between learning trials to test for acquisition of the skill. No prompts were given during these trials and the screen was unobservable. This probe was presented before Learning Task A Level 1, between all levels, and after mastery of Level 5.

In all Learning Task A levels, if the participant matched a tone within 4 seconds of the prompt, the trial was correct. If the participant required more than 4 seconds matching a tone, the trial was incorrect. If 6 out of 8 trials were correct, the participant advanced to the next step
after a probe trial. If less than 6 out of 8 trials were correct, the participant repeated the current step until they scored 6 out of 8 correct, or study time was exhausted.

In Learning Task A level 1, the participant was asked to play the C major scale with 1 whole tone allowed variability with an auditory tone present during playing (“play 1”, C auditory tone was played while the participant attempted to play C, the experimenter said “mute” after the participant matched within 1 whole step of C for 2 seconds, “play 2”, D auditory tone was played while the participant attempted to play D, the experimenter said “mute” after the participant matched within 1 whole step of D for 2 seconds, etc.).

In Learning Task A Levels 2 and 3, the participant was asked to play the C major scale with ½ tone allowed variability (Level 2) or ¼ tone allowed variability (Level 3) with an auditory tone present during playing. Instructions were identical in these steps to Level 1.

In Learning Task A level 4, the participant was asked to play the C major scale with ¼ tone allowed variability with an auditory tone delayed by 4 seconds after the participant began playing (“play 1”, the participant attempted to play C, the experimenter played a C tone 4 seconds after the participant began playing, the experimenter said “mute” after the participant matched within ¼ step of C for 2 seconds, “play 2”, the participant attempted to play D, the experimenter played a D tone 4 seconds after the participant began playing, the experimenter said “mute” after the participant matched within ¼ step of D for 2 seconds, etc.).

In Learning Task A level 5, the experimenter played a C reference tone before the participant was asked to play the C major scale with ¼ tone allowed variability with an auditory tone delayed by 4 seconds. Instructions were identical to Level 4 with the exception that the C tone is played as a reference tone prior to prompting the participant to play.
Learning Task A was concluded with a final probe A trial block and probe A was conducted between every level as well.

Regardless of participant progress, they were asked to conclude participation after 60 minutes, or completion of final probe B, whichever came first.

**Learning Task B: Playing Random Notes in C Major (See Figure 10).** After achieving final probe A, participants were inducted into Learning Task B and began Probe B. All randomized prompt sequences were standardized across participants for Study 2.

Probe B consisted of playing a standardized recording asking the participant to play 8 standardized, random, nonrepeating notes in C major without auditory or corrective feedback (The experimenter said “play 4”, the participant attempted to play F for 3 seconds, the experimenter said “mute”, the experimenter said “play 2”, the participant attempted to play D for 4 seconds). There was no accuracy criterion due to this being a probe conducted between learning trials to test for acquisition of the skill. The first trial block of Probe B was conducted immediately after final probe A before Learning Task B Level 1, between Levels 2 and 3, and after passing the Level 3 progression criterion.

In all Learning Task B levels, if the participant matched a tone within 4 seconds of the prompt, the trial was correct. If the participant required more than 4 seconds matching a tone, the trial was incorrect. If 6 out of 8 trials were correct, the participant advanced to the next step after a probe trial. If less than 6 out of 8 trials were correct, the participant repeated that step until they scored 6 out of 8 correct, or study time was exhausted.

In Learning Task B level 1, the experimenter played a C reference tone before the participant was asked to play 8 random nonrepeating notes in C major with ¼ tone allowed
variability with an auditory unison tone delayed by 4 seconds. This step is essentially Learning Task A Level 5 with the note order randomized (The experimenter played a C reference tone, said “play 4”, the participant attempted to play F, the experimenter played an F tone 4 seconds after the participant began playing, the experimenter said “mute” after the participant matched within ¼ step of F for 2 seconds, the experimenter played a C reference tone, said “play 3”, the participant attempted to play E, the experimenter played an E tone 4 seconds after the participant began playing, the experimenter said “mute” after the participant matched within ¼ step of E for 2 seconds, etc.).

In Learning Task B level 2, the experimenter played a C reference tone before the participant was asked to play 8 random nonrepeating notes in C major with ¼ tone allowed variability. Level 2 was the same as Level 1 with the exception that no feedback auditory tone was provided. No visual feedback was provided either to verify correct tones (i.e. the antecedent C reference tone was played, but no feedback tone). In the event of an incorrect response, or 4 seconds without the target frequency, the experimenter would indicate “higher” or “lower” to guide the participant to the correct note and then say “hold” to allow the participant to hear and remember the correct tone for 2 seconds.

Learning Task B was concluded with a final probe B trial block and probe B was conducted between every level as well.
Figure 10. Training progression chart (Study 2)

Inter-observer agreement and procedural integrity. Out of the 37 Learning Task sessions included in Study 1, 35.14% (≥ 33.33%) of sessions (13 sessions) were reviewed by a research assistant, who independently reviewed video recordings and software audio/visual recordings to code learning task data and review procedural integrity adherence.

Inter-observer agreement (IOA) was calculated for each session according to the following formula:

$$\frac{\text{# of agreements}}{\text{total # of agreements + disagreements}} \times 100 = IOA\%$$
Procedural integrity (PI) data was coded using a procedural integrity data sheet in conjunction with session video recordings. Procedural integrity was calculated according to the following formula:

\[
\frac{\text{# of adhered items}}{\text{total # of data sheet items}} \times 100 = \text{PI}\% 
\]

Calculated IOA averaged 96.20% across 13 sessions (ranging from 91.89% to 100%).

Calculated PI averaged 98.50% across 13 sessions (ranging from 95.12% to 100%).

**Study 2 Results**

**Correlations**

Comparisons between tonal memory scores and measures of performance during the Learning Tasks demonstrated the relations between these variables. First, data distributions were assessed with the Shapiro-Wilk test for normality to assess whether the parametric Pearson or nonparametric Kendall correlation coefficient was most appropriate for given comparisons. The Shapiro-Wilk test results also informed the choice between a linear visual model or LOESS (local regression) line for visualizing trends, where applicable (See Table 2 for Shapiro-Wilk results).
Table 2

*Descriptive Statistics and Shapiro-Wilk Tests of Normality across Variables (Study 2)*

<table>
<thead>
<tr>
<th>Variables (N = 37, unless otherwise specified)</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>W</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seashore Score (out of 30)***</td>
<td>24.24</td>
<td>5.67</td>
<td>.86</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Drake A+B Score (out of 108)</td>
<td>63.51</td>
<td>13.22</td>
<td>.98</td>
<td>.63</td>
</tr>
<tr>
<td>Duration to Final Probe A (in minutes)**</td>
<td>21.03</td>
<td>9.82</td>
<td>.90</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Total Trial Blocks to Final Probe A**</td>
<td>13.97</td>
<td>6.75</td>
<td>.91</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Probe A Average**</td>
<td>2.55</td>
<td>1.41</td>
<td>.89</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Probe B Average (n = 34)</td>
<td>3.21</td>
<td>1.36</td>
<td>.94</td>
<td>.07</td>
</tr>
<tr>
<td>Overall Probe Average**</td>
<td>2.67</td>
<td>1.30</td>
<td>.90</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Duration to Final Probe B (n = 24)**</td>
<td>14.98</td>
<td>5.78</td>
<td>.87</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Additional Trial Blocks to Final Probe B (n = 24)**</td>
<td>7.25</td>
<td>3.9</td>
<td>.83</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Top Score Probe A*</td>
<td>5.32</td>
<td>1.87</td>
<td>.94</td>
<td>.05</td>
</tr>
<tr>
<td>Top Score Probe B (n = 34) *</td>
<td>4.5</td>
<td>1.5</td>
<td>.93</td>
<td>.04</td>
</tr>
<tr>
<td>Highest Instructional Level Attained (out of 8)***</td>
<td>7.35</td>
<td>1.06</td>
<td>.67</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Level Before Highest Probe Score*</td>
<td>3.43</td>
<td>1.83</td>
<td>.93</td>
<td>.02</td>
</tr>
</tbody>
</table>

*α is ≤ .05 (i.e. distribution is assumed non-normal)*

**α is ≤ .01 (i.e. distribution is assumed non-normal)*

***α is ≤ .001 (i.e. distribution is assumed non-normal)*
Seashore tonal memory score did not significantly predict probe A average performance ($r_\tau(35) = .19, p = .11, ns$), nor probe B average performance ($r_\tau(32) = .01, p = .94, ns$), nor overall probe average ($r_\tau(35) = .22, p = .07, ns$), nor duration to final probe B mastery ($r_\tau(22) = -.11, p = .45, ns$), nor additional trial blocks to final probe B ($r_\tau(22) = -.08, p = .63, ns$), nor top score for probe A ($r_\tau(35) = .15, p = .22, ns$), nor top score for probe B ($r_\tau(32) = .09, p = .48, ns$), nor level before highest probe score ($r_\tau(35) = -.05, p = .72, ns$). Kendall Rank Correlation was used for all comparisons due to the non-normal distribution of Seashore scores.

However, Seashore tonal memory score did significantly predict duration to final probe A ($r_\tau(35) = -.52, p < .001, s$, see Figure 11), total trial blocks to final probe A ($r_\tau(35) = -.49, p < .001, s$, see Figure 12), and highest instructional level attained ($r_\tau(35) = .27, p = .049, s$, see Figure 13).
Figure 11. Non-linear correlation between Seashore Tonal Memory score and duration to final probe A with LOESS line (Study 2)
Figure 12. Non-linear correlation between Seashore Tonal Memory score and total trial blocks to final probe A with LOESS line (Study 2)
Figure 13. Non-linear correlation between Seashore Tonal Memory score and highest instructional level attained with line of best fit (Study 2)

Drake musical memory test score did not significantly predict duration to final probe B ($r_t(22) = -.24, p = .10, ns$), nor additional trial blocks to final probe B ($r_t(22) = -.21, p = .18, ns$), nor top score for probe A ($r_t(35) = .12, p = .32, ns$), nor top score for probe B ($r_t(32) = .12, p = .37, ns$), nor probe A average ($r_t(35) = .08, p = .48, ns$), nor probe B average ($r(32) = .32, p = .06, ns$), nor overall probe average.
\( r_\tau (35) = .18, p = .13, ns \), nor level before highest probe score \( r_\tau (35) = -.09, p = .47, ns \).

All comparisons were assessed using either Pearson’s Product-Moment Correlation \( (r) \) for parametric distributions or Kendall’s Rank Correlation \( (r_\tau) \) for nonparametric distributions.

However, Drake test score did significantly predict duration to final probe A \( r_\tau (35) = -.44, p < .001, s \), see Figure 14, total trial blocks to final probe A \( r_\tau (35) = -.40, p < .001, s \), see Figure 15, and highest instructional level attained \( r_\tau (35) = .30, p = .02, s \), see Figure 16. Interestingly, Drake test score also significantly predicted probe A average performance \( r(35) = .37, p = .03, s \), see Figure 17, and overall probe average performance \( r(35) = .44, p < .01, s \), see Figure 18, but only if the distributions for those data were assumed to conform to a linear correlation and analyzed using a Pearson correlation. No other variables of interest demonstrated the unusual correspondence of highly significant findings with the Pearson correlation and highly insignificant findings with the Kendall correlation. These significant findings may be generally considered robust in instances where the value of \( p \) approaches 0, as it does in these comparisons, though caution would be warranted in making a claim of nonsignificance (Kowalski, 1972).
Figure 14. Non-linear correlation between Drake Musical Memory A+B score and duration to final probe A with LOESS line (Study 2)
Figure 15. Non-linear correlation between Drake Musical Memory A+B score and total trial blocks to final probe A with LOESS line (Study 2)
Figure 16. Non-linear correlation between Drake Musical Memory A+B score and highest instructional level attained with line of best fit (Study 2)
Figure 17. Linear correlation between Drake Musical Memory A+B score and probe A average performance (Study 2)
Figure 18. Linear correlation between Drake Musical Memory A+B score and overall probe average performance (Study 2)

Sample Participant Performances

In order to illustrate typical learning trajectories of participants, six sample participants are shown, corresponding with the top two, middle two, and bottom two performers out of the entire Study 2 sample, as assessed through overall probe average scores. These learning trajectories are shown alongside both Seashore and Drake scores to demonstrate whether individual results validate the correlations shown above.
The top two performers in overall probe average (participants 50 and 26) demonstrated immediate performance improvements. Participant 50 exhibited acquisition to 8 out of 8 notes correct after level 1 of Learning Task A and was the only participant to acquire maximal performance so early in instruction. Participant 50 probe performance remained nearly perfect through Learning Tasks A and B and both Seashore and Drake scores were the highest observed of all participants.

Participant 26, the second highest overall probe average performer, showed similarly high performance, but only for Learning Task A. Learning Task B performance reduced to between 3-5 out of 8 and duration to completion of the learning tasks increased by 2-4 total minutes of instruction. The Seashore score was the same as participant 50, but Drake performance was reduced by six points (See Figure 19).

The two performers with overall probe average scores closest to the sample average (participants 30 and 43) demonstrated initial increases in Learning Task A performance followed by decrements in performance across repeated probes. Both participants showed longer duration to completion times compared to the top 2 performers. Following advancement to Learning Task B, participant 30 demonstrated a further increase in performance while participant 43 showed maintenance of accuracy with high variability until the final probe of Learning Task B. Seashore scores were 26 and 23 and Drake scores were 60 and 68, respectively (see Figure 20).

The two performers with the lowest overall probe average scores (participants 49 and 24) demonstrated low and variable accuracy throughout session. Participant 49 demonstrated a slow acquisition of accuracy throughout Learning Tasks A and B, beginning with 0-1 out of 8
correct through Learning Task A, followed by a sudden increase in accuracy with the beginning of Learning Task B and a maximum of 3 out of 8 notes accuracy in probe performance. Despite the low accuracy, the participant was able to meet progression criteria on all Learning Tasks in a moderately short duration. The Seashore score was 28 and Drake score was 70, which are unexpectedly high for the probe performances observed, but consistent with the short observed durations to progression criterion.

Participant 24 demonstrated low average performance and did not progress past Learning Task A in the 60 minute session (See Figure 21). Performance was improved to a maximum of 2 out of 8 correct on Learning Task A. The Seashore score was 21 and the Drake Score was 49, which would predict both the long duration of instruction in Learning Task A, as well as low average probe scores (see Figure 21).
Figure 19. Top two study 2 participants’ (50 and 26) overall average probe scores (99th and 98th percentiles) along with Seashore and Drake scores.
Figure 20. Middle two study 2 participants’ (30 and 43) overall average probe scores (both 50th percentile) along with Seashore and Drake scores
Figure 21. Bottom two study 2 participants’ (49 and 24) overall average probe scores (10th and 7th percentiles) along with Seashore and Drake scores.
Demographic Comparisons

For gender, overall probe average scores were compared using a Mann-Whitney rank sum test due to the non-normality of the residuals \((W = .84, p < .01, s; \text{female:})\)

\[
\text{Mdn}(17) = 2.56, \text{IQR} = .67; \text{male: Mdn}(20) = 2.44, \text{IQR} = 1.86)
\]

and the equal variances of the samples verified through a Levene test of medians \((F(1,35) = .98, p = .33, ns)\). The Mann-Whitney comparison showed that the effect of gender on overall probe average was not significant \((U = 166.5, p = .93, ns)\).

For extent of musical instruction, overall probe average scores were compared using a pooled variances t-test due to the normality of residuals \((W = .90, p = .07, ns; \text{untrained:})\)

\[
M(18) = 2.68, SD = .83; \text{trained: } M(19) = 2.67, SD = 1.65\]

and the equal variances of the samples verified through a Levene test \((F(1,35) = 2.14, p = .15, ns)\). Participants were categorized as trained if they reported at least 3 years of musical instruction. Participants were categorized as untrained if they reported less than 3 years of musical instruction. The t-test showed that the effect of self-reported prior musical instruction of 3 or more years on overall probe average score was not significant \((t(35) = .03, p = .98, ns)\).

For gender, durations to final probe A were compared using a pooled variances t-test due to the normality of the residuals \((W = .96, p = .70, ns; \text{female:})\)

\[
M(17) = 20.53, SD = 10.82; \text{male: } M(20) = 21.45, SD = 9.16\]

and the equal variances of the sample verified through Levene test \((F(1,35) = .85, p = .36, ns)\). The t-test showed that the effect of gender on duration to probe A performance was not significant.
\[ t(35) = -0.28, \ p = .78,ns. \]

For extent of musical instruction, durations to final probe A were compared using a Mann-Whitney rank sum test due to the non-normality of residuals \( W = .81, \ p < .01, s; \)
untrained: \( Mdn(18) = 21.75, IQR = 10.75; \) trained: \( Mdn(19) = 14.5, IQR = 9 \) and the equal variances of the samples verified through a Levene test of medians \( F(1,35) < .001, \ p = .99,ns. \) The Mann-Whitney test showed that the effect of musical instruction of 3 or more years on duration to final probe A was not significant \( (U = 233, \ p = .06,ns). \)

For age, Kendall rank correlations determined no significant relations between age and overall probe average score \( (r_c(35) = .10, \ p = .41,ns) \) or duration to final probe A \( (r_c(35) = -.01, \ p = .94,ns). \)

Comparing scores across gender, Seashore and Drake composite scores (Seashore +
Drake scores) of Study 2 participants exhibited normal distributions \( W = .97, \ p = .82,ns; \)
female: \( M (17) = 91.88, SD = 17.15; \) male: \( M (20) = 84.25, SD = 17.33 \) and equal variances \( (F(1,35) = .31, \ p = .58,ns). \) Unlike the overall assessment analysis, a pooled variances t-test showed no significant effect of gender on Study 2 composite scores \( (t(35) = 1.34, \ p = .19,ns), \) which may account for the lack of significant relations seen in Study 2 overall average probe scores across gender as well.

Across musical training experience, Seashore and Drake composite scores of Study 2 participants exhibited normal distributions \( W = .98, \ p = .92,ns \) and equal variances
In keeping with the overall assessment analysis, an independent samples t-test showed a significant difference in Study 2 composite scores between musically untrained ($M = 80.89, SD = 16.57$) and trained ($M = 94.26, SD = 16.03$) participants, though this difference between the musically trained and untrained was not seen in overall probe average scores or durations to final probe A.

**Study 2 Discussion**

**Descriptive Data**

Study 2, unlike Study 1, revealed a high incidence of non-normal data distributions. The Shapiro-Wilk tests of normality revealed that most variables were non-normal with the exception of the Drake A+B test scores and probe B average performance. The results indicate that most variables could not be assumed to vary continuously, and the nonparametric Kendall correlation was most appropriate for comparing them to other variables. Only the Drake scores and probe B averages could be assumed to vary continuously, though linear correlations were observed in the absence of the normality assumption for Probe A averages and overall probe averages compared to Drake scores.

The irregularity of data distributions in Study 2 may have constituted an artifact of unbiased recruitment procedures wherein active solicitation of random participants with minimal advertisement was used to encourage uninterested participation from a typical undergraduate student population. Despite these measures, a small proportion of high performing participants greatly skewed the data distribution, which can be expected to recur across repeated administrations. In order to obtain a normal distribution of performance scores, screened recruitment of trained musicians separate from typical students would likely be
necessary, which would have defeated the purpose specific to this study: assessing musical potential in a general population of students.

Correlations

The Seashore and Drake assessments were both shown to significantly predict duration to completion of Learning Task A, total trial blocks or degree of instructional support to completion of Learning Task A, and highest instructional level attained. The removal of visual feedback for Study 2 likely contributed to these significant findings, when no such findings were detectable in Study 1. Without visual feedback, participants could only progress using tonal memory and auditory discrimination abilities in Study 2. The variables predicted by both Seashore Tonal Memory and Drake Musical Memory scores might be assumed to correspond with quickly learning a musical task with prompts as well as completing a course of musical instruction, which validates one of the original intents of both Seashore and Drake: to describe the auditory discrimination skills of students in order to determine their capacity to benefit from musical instruction.

The significant findings for Seashore test support its use for the purpose of predicting the likelihood of a student succeeding in a general course of musical study, though the lack of significant correlations with probe average scores indicates that the test might not be able to predict advanced musicianship in the absence of instructional prompting, as in composition and improvisation. The brief nature of the tonal memory assessment (app. 15 minutes duration) may lend well to screening a large population of students for the purposes of general music ensemble placement without needing to run the complete battery of Seashore measures.
The Drake test additionally predicted probe A average score and overall probe average score, but only with the parametric Pearson correlation. This replicated the findings of Study 1, although the score distributions in Study 2 were significantly non-normal, likely owing to the recruitment of both musically trained and untrained participants, which resulted in a skewed distribution toward lower scores with a couple of very high scores as outliers. These findings indicate that the Drake test may provide all the information of the Seashore test, as well as predicting the likelihood a student is capable of unprompted accurate musical pitch production as a byproduct of training, also known as absolute pitch. It may be surmised that the Drake test better predicts individual capacity in skills that involve pitch memory, such as improvisation, composition, and other creative musical pursuits in addition to predicting simpler prompted echoic and textual musicianship such as that observed in ensembles and among musicians who rely on standard notation to play accurately. Complicated musical capacities exemplified by composition and improvisation may be conceptualized as a form of autoclitic tacting, defined as elements of a musical piece evoking elaborate verbal descriptions of those elements (i.e. key, timing, chords, song structure, etc.) (Reynolds & Hayes, 2017). In cognitive psychology and musicology, these concepts have been referred to more conventionally as ‘listening or aural skills,’ ‘melodic, harmonic, and polyphonic dictation,’ and ‘musical reading and performing skills’ (Karpinski, 2000). These more complex verbo-musical skills deserve further research and the behavioral perspective in particular stands to benefit their conceptualization in non-mentalistic terms.

Neither the Seashore nor Drake tests were able to predict any variables of interest specific to Learning Task B, which indicates that absolute pitch of random notes may require additional abilities beyond tonal memory alone. Given that the production of notes occurs
relative to others, this may involve implicit verbal responding. Assessment of the implicit verbal responding component of creative musicality may require the development of novel assessment tools specific to that purpose.

**Sample Participants’ Results**

The habituation and dishabituation of accurate responding observed in Study 1 was not observed in the sample participants of Study 2. Only one out of six participants exhibited a clear reduction in accuracy toward the end of Learning Task A. As such, this effect in Study 1 may have perhaps been a byproduct of the visual prompting strategy. The learning trajectories observed in Study 2 sample participants were idiosyncratic and not clearly indicative of a general trend. This finding should prompt research into the extent to which habituation and dishabituation vary across sensory modalities in human learning.

A contradictory outlier was observed in Study 2, similar to the one observed in the Study 1 sample participants. Participant 49 exhibited much lower overall average probe scores than assessment results would indicate. This replicated outlier supports the conclusion that assessment of musicians should include both tonal memory tests and direct performance measures due to the high likelihood that as many as 1/6 of performances might contradict predictive assessments. Given this variability in correlation, music educators are advised to be cautious in using music discrimination tests on an individual basis in the absence of performance measures as a means of predicting student potential and degree of instructional support needed. Utilization of both tonal memory and direct performance assessments is strongly recommended.
Demographic Comparisons

In contrast to the significant effects of gender and musical training on Seashore and Drake composite scores, no significant effects of gender, musical training, or age on overall probe average scores or durations to final probe A were evident. However, this may owe to the lack of evident relationship between gender and composite score in the Study 2 sample. The relationship between musical training and overall probe average was contradictory though, with no relationship between musical training and average performance or duration to final probe A, but a significant relationship between musical training and composite assessment score. This finding underscores how tonal memory may provide an indirect advantage to musical learning. Even though tonal memory has been shown to be the most predictive variable for music learning capacity, the act of producing pitch accurately or acquiring musical responses likely owes to the coordinated integration of many behavioral capacities in tandem, which complicates simple correlations between discriminative capacities and performance abilities. These findings also indicate that the extent of traditional music instruction alone may not improve absolute pitch capacity in typical undergraduate students, indicating that a specialized procedure such as the direct instruction protocol used in this study is likely necessary to demonstrably improve absolute pitch ability.

Limitations

Many of the limitations from Study 1 including low recruitment numbers, non-standardization of prompts, non-blinded demographic information, visual display being visible to participants, and passive recruitment of interested participants (See Study 2: Introduction) were satisfactorily corrected in the revised Study 2 method.
Despite these improvements, certain limitations inherent to this sort of study remained in Study 2. One observation of interest is the highly skewed distribution of probe averages toward low scores. This non-normal distribution should be reliably expected in a general sample of university students, who will generally score quite low on absolute pitch probe performances with the exception of a couple rare high scorers. Because high performance in absolute pitch is an exceptionally rare capacity in the general population, unbiased recruitment of more participants from a general population will tend to further confirm a low skewed, non-normal distribution of overall probe average scores with a few high performers.

The lack of significant predictive validity of the Seashore or Drake assessments on Learning Task B duration and probe averages is indicative of the fact that tonal memory does not significantly predict the production of random notes to the same degree that it predicts playing a stepwise scale. This should prompt further research into the role of verbal behavior in advanced musical abilities, specifically absolute pitch of random notes.

**Future Directions**

Further refinement of these methods should replicate and extend measurements of tonal memory as a concept, which has not received popular recognition as an important predictor of musical capacity, despite findings verifying exceptional tonal memory among elite musicians. Both the Seashore and Drake tests have demonstrated utility nearly 65 years after their releases, yet have not been revised or extended in novel forms since that time. Further replication of work in this area as well as technological extensions of the Seashore and Drake assessments would help revitalize the investigation of tonal and musical memory. With regard to assessment of auditory verbal relations, it may also be possible to generate a more predictive
battery of tests that remain consistent with naturalistic postulates and the scientific method without resorting to mentalistic explanations.

The technology used in follow up studies of this kind should aspire to automation of the study processes. The assessments could be digitally rendered and data could be automatically entered. This would simplify assessment and allow for a potentially greater recruitment of participants with less paper filing. Automation of the learning tasks could also allow for feedback deliverable through software, which would also remove the need for inter-observer agreement and procedural integrity data in addition to removing the extensive time requirements of manual data coding. Automation of this kind may well prove obligatory for replications given the admittedly rigorous demands of the manualized research method.

Lastly, expansions of this study specifically interested in comparisons across gender, musical instruction, and age of participants should specifically recruit and group participants based on these categories to generate more representative comparisons of these groups.

**Conclusion**

The research thesis of both studies was to examine whether tonal memory grants an advantage in music learning. Both studies demonstrated that tonal memory allows for quicker musical learning while the capacity to describe remembered sequences with autolitic tacts predicts corresponding improvements in stepwise absolute pitch. Further, analysis of demographic factors revealed an unexpected advantage for females in composite scores for the Seashore and Drake tests and the expected finding that musical training significantly improves Seashore Tonal Memory and Drake Musical Memory scores. These effects were not evident in the Learning Tasks, nor were the demographic differences reflected in the Study 2 assessment
scores. Also, a distinction was shown between duration to learning a musical task to criterion and probe average performance.

Duration to learning a musical task to criterion corresponds with prompted echoic instruction, such as that used in elementary music teaching. Probe average performance corresponds with the absolute pitch and unprompted independent tone production most commonly exhibited in the context of composition, improvisation, and other creative musical behaviors. It was shown that the Seashore Tonal Memory test predicts elementary music learning, but not unprompted absolute pitch acquisition. The Drake Musical Memory test, however, was shown to significantly predict both types of music learning. On this basis, the Drake test, or related tests based on memory and autoclitic tacting (i.e. musical appreciation) of musical pieces, is recommended when testing time allows, as this best predicts student potential. The findings of these studies also demonstrate the importance of tonal memory and music appreciation as fundamental contributors to musical learning, which may also constitute important learning goals for introductory musical instruction. Further study should also examine whether improvements in tonal and musical memory result in improved music performances.

On the basis of the findings, music educators are well advised to implement tests of tonal or musical memory to predict students likely to succeed in musical instruction as well as identify students who require remedial training to keep pace in larger ensemble instruction settings. The wide range of tonal memory and musical memory scores demonstrated by the participants across both studies also indicate that individualized instruction may be of particular importance in music education.
The stark increase in significant correlations between assessment scores and variables of interest in Study 2 following the removal of the visual prompts used in Study 1 provides evidence that allowing visual or proprioceptive feedback during instruction of absolute pitch easily confounds demonstrations of the skill. It is recommended on that basis to avoid visual or proprioceptive prompting strategies in assessment of absolute pitch ability.

The studies also constituted novel demonstrations of the theremin as a useful tool for assessing the musical skills of participants, especially in conjunction with software capable of visualizing performance in real time. The theremin is recommended for absolute pitch assessment and instruction since it easily controls for musical experience and technical skill compared to other musical instruments by rendering root tones and proprioceptive feedback ambiguous.

Finally, the direct instruction protocol for absolute pitch was successful at increasing participant performance above baseline levels for all 55 participants who participated in the learning tasks without the use of punitive feedback and with a method allowing exact determination of exactly which levels evoked improved performances for each participant. These results demonstrate that instruction methods derived from behavior science may be of great utility in music education. Furthermore, the particular method of direct instruction used in these studies was shown to reliably and efficiently improve absolute pitch ability with participants spanning all levels of musical experience and discriminative capacity in one hour of instruction. It should be noted that these results were attained not only with simple tasks, but with absolute pitch, which is typically considered an advanced musical skill.
References


Domjan, M. (2016). The Tertis/Pavlov Project. The University of Texas at Austin. https://sites.utexas.edu/tertispavlovproject/

Drake, R. M. (1930). Tests of musical talent. Doctoral thesis, University College, University of London. A copy is also held by Special Collections and Archives, Kent State University/Drake papers.


Education, 10(1), 39-46.
## Appendix A: Study 1 Procedural Integrity Data Sheet

### Assessment Session Procedural Integrity Checklist

<table>
<thead>
<tr>
<th>Question</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Are consent forms and answer sheets present in the room at the start of session?</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>Are the tape player and tapes present and ready to play from the correct starts (Side B of Seashore before Tonal Memory Test, Side A of Drake) with volume set to 4(1) for Seashore and Drake?</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>Are pens available for all participants present?</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>Did the experimenter collect Name, Participant Number, Age, instrumental history, vocal history, theremin history, and handedness?</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>Did the experimenter verify participants understood the Seashore and Drake tests prior to beginning?</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>Did the experimenter pause 2 minutes between Seashore and Drake A, and between Drake A and Drake B?</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>Did the experimenter rewind the tapes to the correct starts after each test was complete?</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>Were test sheets completed by both participants and verified complete by the experimenter?</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>Were sheets stored ungraded and locked in the filing cabinet with the tape player and tapes?</td>
<td>YES</td>
<td>NO</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Score</th>
<th>9</th>
<th>Percentage</th>
</tr>
</thead>
</table>

### Theremin Session Procedural Integrity Checklist

<table>
<thead>
<tr>
<th>Question</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is the theremin, speaker, and laptop present in the room, disinfected, plugged in and ready to play prior to the participant arriving?</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>Does the experimenter have a scratch sheet to keep track of conditions?</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>Does the experimenter calibrate the tone to G#/Ab before beginning any conditions?</td>
<td>YES</td>
<td>NO</td>
</tr>
</tbody>
</table>
Does the experimenter verify that the participant can comply with mute/play and higher/lower prompts?  
YES  NO

Does the experimenter begin Learning Task A with a probe trialblock?  
YES  NO

Does the experimenter proceed to Task A Level 1?  
YES  NO

Does the experimenter verify at least 6 out of 8 correct before proceeding to the next probe?  
YES  NO

Is the next probe conducted?  
YES  NO

Does the experimenter proceed to Task A Level 2?  
YES  NO

Does the experimenter verify at least 6 out of 8 correct before proceeding to the next probe?  
YES  NO

Is the next probe conducted?  
YES  NO

Does the experimenter proceed to Task A Level 3?  
YES  NO

Does the experimenter verify at least 6 out of 8 correct before proceeding to the next probe?  
YES  NO

Is the next probe conducted?  
YES  NO

Does the experimenter proceed to Task A Level 4?  
YES  NO

Does the experimenter verify at least 6 out of 8 correct before proceeding to the next probe?  
YES  NO

Is the next probe conducted?  
YES  NO

Does the experimenter proceed to Task A Level 5?  
YES  NO

Does the experimenter verify at least 6 out of 8 correct before proceeding to the next probe?  
YES  NO

Is the next probe conducted?  
YES  NO

Does the experiment then conduct Probe B immediately after Probe A is done?  
YES  NO

Does the experimenter proceed to Task B Level 1?  
YES  NO
Does the experimenter verify random numbers on every trial and at least 6 out of 8 correct before proceeding to the next probe?  
YES  NO

Is the next probe conducted?  
YES  NO

Does the experimenter proceed to Task B Level 2?  
YES  NO

Does the experimenter verify random numbers on every trial and at least 6 out of 8 correct before proceeding to the next probe?  
YES  NO

Is the next probe conducted?  
YES  NO

Does the experimenter proceed to Task B Level 3?  
YES  NO

Does the experimenter verify random numbers on every trial and at least 6 out of 8 correct before proceeding to the final probe?  
YES  NO

Is the final probe conducted before session is concluded?  
YES  NO

Does the experimenter take the assessment sheets home for grading?  
YES  NO

Is all data added within 1 day to the excel file?  
YES  NO

__________/32  ________%
Appendix B: Study 2 Procedural Integrity Data Sheet

Assessment Session Procedural Integrity Checklist

Are consent forms and answer sheets present in the room at the start of session?  YES  NO

Are the tape player and tapes present and ready to play from the correct starts (Side B of Seashore before Tonal Memory Test, Side A of Drake) with volume set to 4(1) for Seashore and Drake?  YES  NO

Are pens available for all participants present?  YES  NO

Did the experimenter collect Name, Participant Number, Age, instrumental history, vocal history, theremin history, and handedness on a folded paper?  YES  NO

Did the experimenter read the scripts of the Seashore and Drake tests prior to beginning?  YES  NO

Did the experimenter pause 2 minutes between Seashore and Drake A, and between Drake A and Drake B?  YES  NO

Did the experimenter rewind the tapes to the correct starts after each test was complete?  YES  NO

Were test sheets completed by participants and verified complete by the experimenter?  YES  NO

Were sheets stored ungraded and locked in the filing cabinet with the tape player and tapes?  YES  NO

Theremin Session Procedural Integrity Checklist

Is the theremin, speaker, and laptop present in the room, disinfected, plugged in and ready to play prior to the participant arriving with left or right handedness accounted for?  YES  NO

Does the experimenter have the Standardized Random Number Sheet?  YES  NO

Does the experimenter calibrate the tone to G#/Ab before beginning any conditions?  YES  NO
Does the experimenter verify that the participant can comply with mute/play and higher/lower prompts? YES NO

Does the experimenter begin Learning Task A with a probe trialblock? YES NO

Does the experimenter proceed to Task A Level 1? YES NO

Does the experimenter verify at least 6 out of 8 correct before proceeding to the next probe? YES NO

Is the next probe conducted? YES NO

Does the experimenter proceed to Task A Level 2? YES NO

Does the experimenter verify at least 6 out of 8 correct before proceeding to the next probe? YES NO

Is the next probe conducted? YES NO

Does the experimenter proceed to Task A Level 3? YES NO

Does the experimenter verify at least 6 out of 8 correct before proceeding to the next probe? YES NO

Is the next probe conducted? YES NO

Does the experimenter proceed to Task A Level 4? YES NO

Does the experimenter verify at least 6 out of 8 correct before proceeding to the next probe? YES NO

Is the next probe conducted? YES NO

Does the experimenter proceed to Task A Level 5? YES NO

Does the experimenter verify at least 6 out of 8 correct before proceeding to the next probe? YES NO

Is the next probe conducted? YES NO

Does the experiment then conduct Probe B immediately after Probe A is done? YES NO

Does the experimenter proceed to Task B Level 1? YES NO
<table>
<thead>
<tr>
<th>Question</th>
<th>YES</th>
<th>NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Does the experimenter verify random numbers on every trial and at least 6 out of 8 correct before proceeding to the next probe?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Is the next probe conducted?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Does the experimenter proceed to Task B Level 2?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Does the experimenter verify random numbers on every trial and at least 6 out of 8 correct before proceeding to the final probe?</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>Is the final probe conducted before session is concluded?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Does the experimenter take the assessment sheets home for grading?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Is all data added within 1 day to the excel file?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

________/29     _______%
Appendix C: Demographic Information Questionnaire

Participant #:

1. Do you have any experience playing the theremin?

2. Have you heard of the theremin before, or ever seen one, including online or in entertainment media?

3. Have you received formal instruction in playing a musical instrument?

4. If so, what type(s) of musical instrument(s) and how many years each?

5. Have you received formal vocal instruction, such as choir or singing lessons?

6. If so, how many years?

7. How do you gender identify?

8. What is your age in years?

9. Are you right or left handed?

Please fold this sheet in half when you are done. Thank you!
### Appendix D: Study 2 Learning Task B Random Number Sheet

<table>
<thead>
<tr>
<th>Number 1</th>
<th>Number 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>57413682</td>
<td>25176384</td>
</tr>
<tr>
<td>71648253</td>
<td>37251486</td>
</tr>
<tr>
<td>24716835</td>
<td>63827514</td>
</tr>
<tr>
<td>63852471</td>
<td>51638247</td>
</tr>
<tr>
<td></td>
<td>48315726</td>
</tr>
<tr>
<td></td>
<td>74261358</td>
</tr>
<tr>
<td></td>
<td>37286154</td>
</tr>
<tr>
<td></td>
<td>28643175</td>
</tr>
<tr>
<td></td>
<td>48261573</td>
</tr>
<tr>
<td></td>
<td>62583741</td>
</tr>
<tr>
<td></td>
<td>57413682</td>
</tr>
<tr>
<td></td>
<td>71635824</td>
</tr>
<tr>
<td></td>
<td>25374168</td>
</tr>
<tr>
<td></td>
<td>48613527</td>
</tr>
<tr>
<td></td>
<td>37416285</td>
</tr>
</tbody>
</table>