ISSN: 2083-389X
eISSN: 2084-3127

J Hear Sci, 2021; 11(4): 55–71
DOI: 10.17430/JHS.2021.11.4.7

PRELIMINARY PSYCHOACOUSTIC INVESTIGATION OF THE SIAM PROCEDURE TO MEASURE FREQUENCY DIFFERENCE LIMENS

Contributions:
A Study design/planning
B Data collection/entry
C Data analysis/statistics
D Data interpretation
E Preparation of manuscript
F Literature analysis/search
G Funds collection

Aurora J. Weaver[1,A-F], Jeffrey J. DiGiovanni[2,A,D-F], Ryleigh Edwards[1,B,D-F], Erin E. Lynch[3,D-E], Dennis T. Ries[3,A-D], Hayes Vinson[1,E-F], Anne Rankin Cannon[1,B-F]

[1] Speech, Language, and Hearing Sciences, Auburn University, United States
[2] Communication Sciences and Disorders, University of Cincinnati, United States
[3] Communication Sciences and Disorders, Ohio University, United States

Corresponding author: Aurora J. Weaver; Speech, Language, and Hearing Sciences, Auburn University, 1199 Haley Center, 36849, Auburn, United States; email: ajw0055@auburn.edu

Abstract

Background: The study investigated the cause of varying estimates of frequency difference limens (DLs) in delayed comparison tasks involving pitch retention in auditory working memory (AWM). Using procedures adapted from the method of constant stimuli (MCS) and the single-interval adjustment matrix (SIAM), we sought to determine via 3 experiments whether the disparity in frequency DLs obtained using each procedure was due to the method of measurement (Experiment 1), the response format (Experiment 2), or performance feedback (Experiment 3).

Material and methods: Five adults (ages 21 to 38 years) with hearing within normal limits participated in Experiments 1 and 2, and seven adults (ages 20 to 30 years) with hearing within normal limits participated in Experiment 3. Delayed comparison tasks were used to evaluate frequency DLs under SIAM and MCS.

Results: Our preliminary results suggest that DL values for pitch discrimination are more influenced by response format than by the measurement procedure or performance feedback. Regardless of the method used, DL values were greater in the condition containing intervening stimuli compared to the condition lacking intervening stimuli.

Conclusions: Preliminary findings suggest there is consistency in the listener's adopted criterion (i.e., judgment rationale) across the psychoacoustic methods investigated. Performance measures suggest that SIAM is as accurate as MCS, but it is noteworthy that two SIAM measurement runs using the "same/different" response format is more efficient than four runs with MCS. Future application of the SIAM procedure for measuring DL values might, with larger sample sizes, identify additional factors that contribute to performance and the listener's adopted criterion, since data collection time is appreciably shorter with SIAM.

Key words: frequency difference limen • just noticeable difference • auditory working memory • single-interval adjustment matrix • method of constant stimuli

WSTĘPNE BADANIE PSYCHOAKUSTYCZNE ZASTOSOWANIA PROCEDURY SIAM DO POMIARU PROGÓW RÓŻNICOWANIA CZĘSTOTLIWOŚCI

Streszczenie

Wprowadzenie: Zbadano przyczynę występowania różnic oceny progów różnicowania częstotliwości (DL) w zadaniach opóźnionego porównania związanych z zapamiętywaniem częstotliwości w słuchowej pamięci roboczej (AWM). Stosując procedury zaadaptowane z metody stałych bodźców (MCS) i macierz dostosowań po pojedynczych interwałach (SIAM), przeprowadziliśmy 3 eksperymenty w celu określenia, czy rozbieżność DL częstotliwości uzyskanych z zastosowaniem każdej z procedur, była związana z metodą pomiaru (Eksperyment 1), formatem odpowiedzi (Eksperyment 2) czy informacją zwrotną na temat wyników (Eksperyment 3).

Materiał i metody: Pięć osób dorosłych (w wieku 21–38 lat) z normalnym słuchem wzięło udział w Eksperymentach 1 i 2 oraz siedem osób dorosłych (w wieku 20–30 lat) z normalnym słuchem wzięło udział w Eksperymencie 3. Oceny DL częstotliwości z użyciem SIAM i MCS dokonano na podstawie zadań opóźnionego porównania.

Wyniki: Wstępne wyniki sugerują, że wartości DL dla rozróżniania tonacji są w większym stopniu zależne od formatu odpowiedzi niż od metody pomiaru czy informacji zwrotniej o wynikach. Niezależnie od użytej metody wartości DL były wyższe w warunkach zawierających wtrącone bodźce niż w warunkach bez wtrąconych bodźców.

Wnioski: Wstępne wyniki sugerują, że przyjęte przez słuchacza kryterium (tj. podstawa oceny) jest spójne dla wszystkich zbadanych metod psychoakustycznych. Wykonane pomiary sugerują, że SIAM jest równie dokładna jak MCS, jednak należy zwrócić uwagę, że dwa pomiary SIAM z zastosowaniem formatu odpowiedzi „taki sam/różny" są bardziej efektywne niż cztery pomiary z MCS. Przyszłe zastosowanie procedury SIAM do pomiaru wartości DL mogłoby, przy większych próbach, zidentyfikować inne czynniki, które przyczyniają się do osiąganych wyników i kryterium przyjmowanego przez słuchacza, ponieważ czas zbierania danych przy zastosowaniu SIAM jest znacznie krótszy.

Słowa kluczowe: częstotliwościowy próg różnicy • najmniejsza obserwowana różnica • słuchowa pamięć robocza • macierz dostosowań po pojedynczych interwałach • metoda stałych bodźców

Background

Pitch discrimination is a critical aspect of both speech recognition and melody perception in music [1]. Psychoacoustic research aims to enhance knowledge of what factors contribute to differences in listeners' abilities, including their limits of pitch discrimination [1,2]. The frequency difference limen (DL) measurement is the smallest perceived change in frequency during behavioral listening tasks. This measurement can help evaluate pitch perception abilities in auditory working memory (AWM) during a delayed comparison task where an individual must retain information about a standard pitch for a period of several seconds prior to making a DL judgment. AWM is the cognitive processing of auditory information necessary to carry out a mental task, which is subject to decay over time and can be interfered with when additional auditory information is present [3–5]. Frequency DL values are typically measured using the Method of Constant Stimuli (MCS) procedure, which is lengthy and requires a listener's extended attention across 100 trials. Within an AWM experimental design each set of 100 trials (i.e., 1 run) may take approximately 35–45 minutes to complete. In the MCS, multiple runs are collected to plot the psychometric value for each experimental condition, which makes data collection time-consuming (e.g., time of each run × number of runs × conditions). For this reason, research designs tend to make use of repeated measures and small sample sizes when using the MCS [6]. Further, since with MCS attention declines with measurement time, studies have investigated the use of other methods such as the single-interval adjustment matrix (SIAM) procedure to obtain DL measurements in a more efficient manner with fewer trials [7]. However, procedural differences between these two measurement methods have yet to be examined. Therefore, the overarching aim of this study was to provide an initial investigation of the differences between these two methods of obtaining frequency DL measures.

Deutsch's [3] model of pitch memory includes three dimensions: the strength of the standard pitch value, how long the perception needs to be maintained (i.e., duration or retention interval), and the acoustic proximity (i.e., similarity) of opposing pitch perceptions. This model results in a Gaussian (bell-shaped) representation of pitch which broadens with elapsed time. The longer the time-span to retain pitch information, the less precise pitch representation becomes. Subsequent psychoacoustic and neural modeling studies have confirmed that newer stimuli can interact with older stored representations, which also alters the stored representations of pitch in AWM of target/standard stimuli [8–12]. While the Deutsch model for pitch representation is informed by measuring the frequency DL in a delayed comparison task, psychophysical methods still produce significantly different absolute DL values for the same listeners [11,12]. It has been proposed that pitch retention and discrimination abilities in AWM might be less affected by training because performance feedback is not crucial [3–5,8–12].

Ries and DiGiovanni [11] used the MCS, which measures the discrimination between a standard and comparison tone, to obtain frequency DLs [13]. They found that absolute values were consistent with the prior literature [3,11,14–21]; however, implementing this method was inefficient due to the number of points above and below the performance point of interest (e.g., 75% correct, which approximates *d*-prime (d') = 1). The performance point of interest is obtained to achieve a fit to the psychometric function, which is used to describe a listener's performance and is dependent upon the physical stimulus [22–24]. The psychometric function is a relationship between the intensity of the stimulus and the participant's tendency to say "yes" to hearing a presented tone [13]. As just stated, the SIAM procedure by Kaernbach [7] is a faster method and, in prior work, employing SIAM has helped improve the efficiency with which frequency DL values can be obtained [12]. The SIAM procedure reported for use in measuring frequency DL prompts a participant to decide if the comparison tone is either the "same" or "different" compared to the standard tone presented; the color of the button that corresponds to the correct answer then changes to provide performance feedback [12]. This application of the SIAM for determining frequency DL uses feedback regardless of what the participant selected, which deviates from the originally described procedure which only provides feedback when the participant makes an error [12]. This procedure cuts data collection time approximately in half by using an adaptive tracking paradigm in which the tones presented are absolute threshold measures comparable to the ones found when using a two-interval forced-choice procedure that controls for response bias [12]. While the use of different psychometric methods has been known to affect measurements, both procedures employed by Ries & DiGiovanni [11,12] were designed to target the same psychometric point of performance of 75% correct [7,25,26].

Each completion of the SIAM procedure, referred to as a run or track, estimates participant performance for a targeted performance point (e.g., 75% correct) along the psychometric function by using rules that average reversal points. Ries and DiGiovanni [12] found that absolute DL values obtained using the SIAM procedure were higher than those obtained with the MCS procedure, even though the pattern of results was the same across similar conditions [11]. The authors postulated that the difference was likely due to different response formats among the studies [12]. Therefore, further evaluation of the frequency DL measurement and calculation method is merited when using the SIAM procedure [12].

Signal detection theory

The signal detection theory framework relates choice behaviors to psychological decisions made during either the signal or noise condition (i.e., targets and foils) trials during perception tasks. In pitch discrimination tasks, the possible answer types include: Hit (H), Miss (M), Correct Rejection (CR), or False Alarm (FA). The accurate detection of a pitch difference is measured as H, whereas the failure to identify a pitch difference is measured as M. When a participant correctly recognizes that two pitches are the same, this is referred to as CR, while inaccurate identification of a pitch difference where there is none is referred to as FA. Thereafter, the performance measures, H rate and FA rate, estimate both detection sensitivity d' and β (see Brophy [27] and Macmillan & Creelman [13] for calculation details). Both detection sensitivity and bias are measured in this approach and may explain the differences found in the absolute frequency DL values via both the MCS and SIAM procedures [12,28–30].

Sensitivity refers to a listener's ability to discriminate perceptual alternatives of a task (e.g., the frequency DL in a pitch discrimination task), which can be indexed by calculating the *d*-prime (*d*′) discriminability index. MacMillan & Creelman [13] note that researchers using a 2-alternative forced-choice (2-AFC) task often look for accuracy between 60 and 90%. An accuracy of 60% corresponds approximately to $d' = 0.5$ (e.g., if a participant is unbiased ($c = 0$), H rate = .60, and FA rate = .40, then $d' = 0.51$). An accuracy ~90%, on the other hand, would derive from an H rate = .90 and FA rate = .10, producing a $d' = 2.6$. As mentioned previously, the MCS has been used to approximate $d' = 1$ using 75% performance criteria in the design. (Note, $d' = 0$ would reflect a change in performance to H rate =.5 and FA rate = .50). Response bias refers to the inclination of a participant to select signals (e.g., pitches are different) versus noise (i.e., pitches that are the same) during their perceptual task. A participant's d' and decision criterion (β) are dependent upon the particular detection task(s) with either signal or noise trials, and responses are coded into the four categories mentioned previously.

In this framework, the optimum β ($\beta = 1$) corresponds to the internal criterion that would provide the optimal blend of missed detections (M) and false alarms (FA) when responding in a discrimination task [30]. $\beta = 1$ is a neutral point, whereas values between 0 and < 1 are considered liberal criterions and values >1 are considered conservative criterions (on a base-10 logarithmic scale). This β value is determined by the probability on a given trial of the standard and comparison tones being the same or different frequency, combined with the perceived value of H and CR, and compared to the cost of M and FA. Since the 1960s, more recent research has indicated that the measure *c* is preferable [28,31,32]. The value of *c* refers to the distance between the participant's adopted criterion and the *neutral point.* The participant's adopted criterion can be either *liberal* or *conservative*. Participants who are likely to respond that a *signal is present* are categorized as having a *liberal* criterion, while those likely to respond that there is *no signal* are categorized as having a *conservative* criterion. When $c = 0$ neither response (i.e., conservative or liberal) is preferred; negative values of *c* indicate a liberal criterion (i.e., $c < 0$); and positive values of *c* indicate a conservative criterion (i.e., $c > 0$) [13].

The current study

There are several ways that an experimental task may affect a participant's optimum β. Two possibilities relevant to the measurement of frequency DL are: 1) the general procedure, and 2) the response format. The way in which the participant's response is elicited during the SIAM procedure alters the values of CR and H, as well as the costs of FA and M. In comparison, the MCS procedure uses a randomized presentation of predetermined pitch differences. Additionally, the SIAM procedure gives performance feedback for each trial when incorrect responses are provided, which may draw more attention to performance than the MCS procedure (which provides no feedback).

A second characteristic of the task that may affect the optimum β is the response format. Two response formats – same or different (S/D) and higher or lower (H/L) – have been adopted for pitch discrimination tasks [11,12,16]. Wickelgren [33] suggested different decision mechanisms may be employed by participants when they are asked to make S/D judgments as opposed to H/L judgments, which cannot yield correct rejections. If an H/L response format is implemented (MCS procedure) [11], the participant is then primed to detect fine differences between the standard and comparison tones since the participant knows that the stimuli on each trial will always differ in pitch. When primed in this way, the participant is more likely to adopt a liberal optimum β (i.e., criterion), resulting in a seemingly finer degree of acuity in detecting pitch changes. In contrast, many participants are more likely to adopt a stricter criterion when the S/D response format is used. Participants are then more likely to wait until the frequency perception is clearly different before they are willing to select a S/D response (SIAM procedure) [12]. This specific response format (S/D) tends to make the participant overlook smaller frequency differences between the standard tones and comparison tones due to the categorical nature of the response format.

Research in categorical speech perception using S/D response formats have demonstrated that participants only answer D if they are very sure of their decision [34]. Therefore, the S/D response tends to make the participant think more conservatively, while the H/L response has the effect of making the participant take a more liberal approach. The SIAM procedure was designed to minimize response bias through differential adjustment of step size, whereby adjustments in steps depend on the listener's responses within the context of signal detection theory [7]. Signal detection theory attempts to explain the difference in frequency DL values between the MCS and SIAM procedures [29]. It suggests that the likelihood that participants will adopt an extremely liberal or conservative criterion is constrained with the SIAM procedure. However, with the MCS procedure, there is no inherent penalty or few active means by which to limit such actions; thus, listeners can adopt a more extreme response bias. Within the MCS procedure, the only method of examining the effect of such bias is through analyses of participant responses on catch trials following data collection. The examiner inspects the catch trials, which are presented as if they were 'noise trials' (i.e, with 0 Hz difference between the standard and the comparison), but forcing the participant to select H/L. The examiner aims to see whether the participant will respond when there is no 'signal' present (recall that a 'signal trial' means the pitches are different whereas a 'noise trial' means the pitches are the same). In this way, the amount of bias can be gauged.

The aim of the present research was to further explore, through evaluating both general procedures and response format for the same participants, why the MCS and SIAM procedures each produce different estimates of the frequency DL. A secondary aim was to determine whether a version of the SIAM procedure could effectively replace the lengthy MCS procedure for determining frequency DLs. To do so, three experiments were conducted in which the 75% correct recognition point of the psychometric function was targeted. For the SIAM measurement, this was achieved by aiming directly for this point, while with the MCS it was done by fitting a psychometric function to the data points.

The three experiments included in this study and their hypotheses are as follows:

1) Experiment 1 compared five potential reversal calculation rules and the number of data collection runs to determine the optimum frequency DL calculation using the SIAM procedure. The null hypothesis was that the reversal rules did not significantly affect the frequency DL estimates. Additionally, the null hypothesis that fewer than four SIAM runs would significantly affect the frequency DL estimates was also tested.
2) Experiment 2 compared data collected with the MCS procedure, SIAM procedure, and a hybrid task to evaluate the influences of the general procedure and the response format. It was hypothesized that response format would significantly contribute to performance differences, with the use of the H/L format in MCS resulting in smaller DLs.
3) Experiment 3 explored the influence of performance feedback and use of increment and decrement (I/D) judgements in the SIAM procedure. Specifically, use of I/D was hypothesized to produce smaller DLs as the participant could switch their internal criterion (judgment rationale) similar to MCS-H/L. Additionally, performance feedback was hypothesized to produce smaller frequency DLs.

Each experiment measured frequency DLs in the following two conditions: 1) in the presence of a silent intercomparison interval (ICI); and 2) in the presence of tones within the ICI. These two conditions were used to compare, and extend, the study findings in the previous AWM literature – specifically, the effects of time on the accuracy of stored pitch representations [3] and potential interactions between new stimuli and old (stored) representations in AWM [8–12].

General method

Participants

Following receipt of Institutional Review Board (IRB) approval at Ohio University, five adult participants (four females, one male), ages 21 to 38 years old, from Ohio University's student and staff population were recruited. After consenting, they participated in and completed Experiments 1 and 2. Four of the five participants had previous experience with psychoacoustic experiments, which were broadly consistent with classical psychoacoustic studies. However, this aspect may make them unrepresentative of the general untrained population (see *Design and data analysis* for power analysis details). For Experiment 3, seven females ages 20 to 30 years, were recruited and consented for participation. Two of the seven participants had previous experience with psychoacoustic experiments, one of whom also participated in Experiments 1 and 2. Participants across all three experiments had pure-tone air-conduction thresholds within normal limits (20 dB HL or better) at octave frequencies from 0.25 to 8 kHz [35]. Additionally, all participants were asked about absolute pitch abilities, which would have excluded them from participation; no participant reported absolute pitch [12]. One additional participant was recruited for the study, and enrolled in Experiment 1, but discontinued their voluntary participation after the first session. Their data was not included in the formal analysis due to missing repeated measures.

Stimuli

All stimuli were created using System III hardware (Tucker Davis Technologies, Alachua, FL) controlled by a Pentium 4 processor (Dell, Round Rock, TX) running Matlab 2007 and RPVD (Tucker Davis Technologies) software. A dynamic signal analyzer (Stanford Research Systems, Sunnyvale, CA) and oscilloscope (Tektronix, Richardson, TX) were used to verify the signals electronically. A 2250-S sound level meter and 2 cc coupler (Brüel & Kjær, Nærum, Denmark) were used to evaluate the acoustic signals. ER-2 insert earphones (Etymotic Research, Elk Grove Village, IL) were employed to present stimuli to the right ear of the participants, who were seated within a double-walled, sound attenuating booth (Industrial Acoustics Corporation, Bronx, NY).

All stimuli (i.e., standard, intervening, and comparison sounds) were 200 ms in duration including 20 ms cosine-squared onset and offset ramps, generated at a sampling rate of 24.414 kHz, and low-pass filtered at 12 kHz. All stimuli were presented at a level of approximately 80 phons. The levels producing equal loudness were based on loudness-matching data collected from two highly trained listeners who had previously participated in a wide array of psychoacoustic experiments including multiple studies on loudness. The level of the 1000-Hz standard in the matching paradigm was 80 dB SPL. None of the subjects reported any notable change in the loudness of the stimuli throughout the study. In addition, prior related work indicated that an intensity difference of 12 dB between the level of the intervening tones and the standard/comparison tones resulted in no significant change in the measured frequency DL [10,11]. The frequency of the standard tone (the first tone presented in a trial) was roved randomly on each trial within a range spanning from 395 to 475 Hz [36]. The comparison tone (i.e., the last sinusoidal tone presented in a trial) was presented 4800 ms after the offset of the standard tone. The timing skeleton for the experiment,

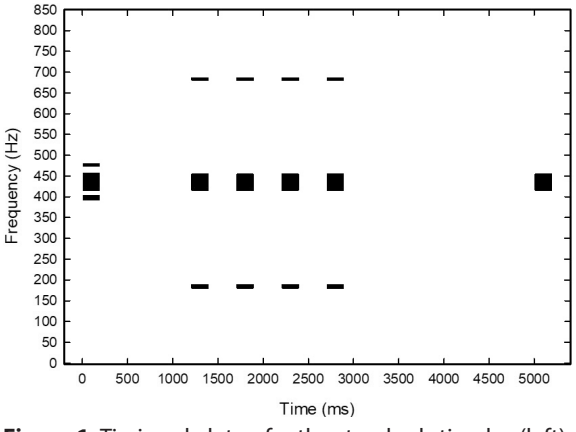

Figure 1. Timing skeleton for the standard stimulus (left), comparison stimulus (right), and four intervening stimuli (1, 2, 3, 4). Intervening stimuli were not presented in the intercomparison interval for the NoINT condition. Reprinted with permissions from [10]

adapted from Deutsch [3,14,15], is shown in Figure 1. The frequency of the comparison tone was adjusted adaptively to target the 75% point on the psychometric function. Four intervening stimuli (tones), with frequencies ranging from 183 to 691 Hz (randomly selected), were presented in the ICI between the standard tones and the comparison tones in one of the conditions.

Conditions

All three experiments consisted of two ICI conditions. In the first condition, hereafter referred to as NoINT, there were silent ICIs with no intervening stimuli present. Participants were presented with a standard tone followed by the onset of the comparison tone 4800 ms later during NoINT (see Figure 1). Following the presentation of the comparison tone, and dependent upon the procedure, participants provided responses (i.e., S/D, H/L, Y/N described later) to the pitch of the comparison tone by pressing a square on the touch-screen monitor. The second ICI condition, hereafter referred to as ToneINT for tonal intervening stimuli, had the same general configuration as NoINT except for the presence of four intervening stimuli (tones). The frequencies ranged from 183 to 691 Hz and were presented randomly in the ICI between the standard and comparison tones in this condition. The onset of the first intercomparison tone occurred 1000 ms after the offset of the standard. Each of the four intervening tones had a duration of 200 ms and each subsequent tone was separated from the prior one by 300 ms. For all three experiments, participants were trained under both ICI conditions (for approximately 15 min prior to formal data collection for the SIAM and for 35–45 min for the MCS). Participants were trained to ensure they understood how to complete each task, consistent with prior work in this area [12].

Measurement of frequency DL

Across the three experiments, six variations of the SIAM procedure were created that involved alterations to response format, frequency comparison, and performance feedback. Two variations of the MCS, involving alterations to the response format, were employed to investigate the research questions outlined above. Table 1 provides a summary of the variations of the frequency DL data collection procedures. For Experiments 2 and 3 in which different data collection procedures were compared, the order of runs of trials was randomized across the two conditions within a given method prior to moving on to the next method. For Experiments 2 and 3, the order of the frequency DL data collection methods was counterbalanced across participants.

Table 1. Summary of the three experimental designs and their purpose

Experiment	General procedure	Response format	FB	Frequency comparison	Frequency DL estimation	Runs averaged	ICI
1	SIAM (target d' = 1) [a]	S/D	yes	I	Rule 1	2 Runs (max. 200 trials)	NoINT
	SIAM (target d' = 1) [a]	S/D	yes	I	Rule 2	3 Runs (max. 300 trials)	ToneINT
	SIAM (target d' = 1) [a]	S/D	yes	I	Rule 3	4 Runs (max. 400 trials)	
	SIAM (target d' = 1) [a]	S/D	yes	I	Rule 4		
	SIAM (target d' = 1) [a]	S/D	yes	I	Rule 5		
Purpose	Investigate reversal rules and number of runs for SIAM – constructed to mimic frequency DLs measured in prior work (Ries & DiGiovanni, 2009)						
2	SIAM (target d' = 1) [a]	S/D	Yes	I	Rule 2	2 Runs (max. 200 trials)	NoINT
	MCS[b]	H/L	No	Fixed I/D	Sigmoid function	4 Runs (min. 400 trials)	ToneINT
	MCS[b]	S/D	No	Fixed I/D	Sigmoid function	4 Runs (min. 400 trials)	
Purpose	Comparisons of response formats and general procedures investigated in a small sample of participants trained in each task						
3	SIAM (target d' = 1) [a]	S/D	Yes	I	Rule 2	2 Runs (max. 200 trials)	NoINT
	SIAM (target d' = 1) [a]	S/D	No	I	Rule 2	2 Runs (max. 200 trials)	ToneINT
	SIAM (target d' = 1) [a]	Y/N	Yes	I	Rule 2	2 Runs (max. 200 trials)	
	SIAM (target d' = 1) [a]	Y/N	No	I	Rule 2	2 Runs (max. 200 trials)	
	SIAM (target d' = 1) [a]	Y/N	Yes	I/D	Rule 2	2 Runs (max. 200 trials)	
	SIAM (target d' = 1) [a]	Y/N	No	I/D	Rule 2	2 Runs (max. 200 trials)	
Purpose	Compare 6 SIAM adaptations (response format, performance feedback, and method of frequency comparison) in a small sample of participants trained in each task – make preliminary recommendations for future applications						

Note: FB = feedback, ICI= intercomparison-interval, SIAM = single-interval adjustment-matrix; d' = d-prime, MCS = method of constant stimuli, S/D = Same or Different frequency, H/L = Higher or Lower frequency, NoINT= silent ICI, ToneINT = Tonal stimuli in ICI, I = increment (a frequency difference always results in a higher comparison frequency; I/D= Increment/Decrement frequency adjustment (a frequency difference can result in a higher or lower comparison frequency)

[a] SIAM matrices: the value of the difference between the standard and comparison on any given trial is adjusted adaptively (i.e., Hit: –1 * step size; Miss: 1 * step size; False Alarm: 2 * step size; Correct Rejection: 0 * step size) based upon the matrix described by Kaernbach for a target performance of 0.5 which estimates 75% correct along the psychometric function

[b] Sigmoidal functions fitted to the MCS data were used to determine the 75% correct point on the psychometric function

Single-interval adjustment matrix (SIAM) procedure

For the SIAM procedure used in Experiment 1, the participant was prompted to provide an S/D response comparing the comparison tone to the standard tone using a touch screen. The SIAM procedure gave the participant correct performance feedback following each response by changing the color of the button corresponding to the correct answer from blue to yellow for 300 ms. In this way, the participants were therefore provided with feedback about correct and incorrect answers. This deviated from Kaernbach's original work, which provided only feedback on incorrect performance [6]; however, the modification is consistent with later work in this area [11]. The presentation of a Same or Different trial was determined randomly across trials. The likelihood of a Different trial was 75%, until the first reversal was obtained, and then it reduced to 50% thereafter. Each run began at a frequency that was roughly 20 Hz above a participant's likely DL value calculated from an earlier training run. Within a given experimental run, the step size was set to 4.0 Hz for the first four reversals, then to 1.0 Hz for the subsequent reversals. The value of the frequency difference between the standard and comparison on any given trial was adjusted adaptively (i.e., Hit: −1 * step size; Miss: 1 * step size; False Alarm: 2 * step size; Correct Rejection: 0 * step size) based upon the matrix described by Kaernbach to track the 75% line [7]. This matrix was created for a target performance of 0.5, which estimates 75% correct along the psychometric function (see Kaernbach (1990) for different adjustments for different target performances). These adjustments applied to a 'signal' trial (i.e., the pitches were different) and were incremental frequency comparisons (I). The frequency difference (e.g., 10 Hz) was set for each signal trail by the SIAM, however the comparison tone was higher in frequency (e.g., standard tone = 395 Hz vs comparison tone = 405 Hz). Additionally, for Experiment 3, an adaptation of the SIAM, referred to as SIAM-I/D Y/N, allowed the comparison tone frequency to be either higher (increment) or lower (decrement) in pitch than the standard (e.g., ±10 Hz (with random likelihood), standard tone = 395 Hz vs comparison tone = 405 or 385 Hz). This is similar to the MCS fixed I/D frequency adjustments, described later, but still controlled the frequency difference between the standard and comparison using the SIAM procedure.

A run consisted of 100 trials in order to match traditionally reported protocols for the MCS [3]. The SIAM rules adopted used the average H rate and CR rate to pinpoint the 75% correct point on the psychometric function. The frequency difference values discard the first four reversals obtained using the staircase procedure [10]. One complete condition measurement consisted of four individual runs. If one of the four individual runs differed from the average of the remaining three by more than two SDs, an additional run was collected. This occurred for 2/40 runs; and the additional run replaced the original run only if it was closer to the mean of the remaining three original run values. This occurred once, representing ~2.5% of the data collected in Experiment 1. Table 2 provides a description of each rule used in Experiment 1.

Number of runs included in SIAM DL

An additional consideration when evaluating the data collected using the SIAM procedure was determining the number of runs, or tracks, that should be included in the frequency DL estimate for a condition. In the MCS procedure, four runs are collected to develop one psychometric function for the condition (see next section). Prior work mentioned has collected and averaged four runs in the SIAM, following practice, to estimate the frequency DL for formal data analysis [12]. It has not been explored whether four runs with the SIAM procedure are necessary. In Experiment 1, the SIAM procedure was collected four times for each ICI condition, the aim being to determine whether all four runs were necessary to calculate performance when using the SIAM, as reported previously [12]. In more detail, for determining DL estimates, three options allowed averaging across either two, three, or four runs. These options allowed us to determine if fewer than four runs could be implemented to increase data collection efficiency when using the SIAM procedure. The choice was based on being able to compare a participant's performance on a minimum of two runs for stability. Therefore, the first option averaged the first and second runs (2 runs), the second option averaged the first three runs (3 runs), and the last option averaged all four runs (4 runs), as reported in prior work [12].

Method of Constant Stimuli (MCS) procedure

The MCS was only employed for Experiment 2. The two versions of the MCS used in Experiment 2 differed primarily in response format, although different frequency ranges were employed in some conditions. The frequency differences used in the MCS procedures are listed in Table 4. In the

Table 2. Single-Interval Adjustment Matrix (SIAM) reversal rules (Experiment 1)

Rule	SIAM reversals included in DL mean of 18 reversals	Rule basis
Rule 1	~~1, 2, 3, 4,~~ **5, 6, 7, 8,** ~~9, 10, 11, 12, 13, 14, 15, 16, 17, 18~~	Matrix count
Rule 2	~~1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14,~~ **15, 16, 17, 18**	Matrix count
Rule 3	~~1, 2, 3, 4,~~ **5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18**	Kaernbach's (1990) Human subjects
Rule 4	~~1, 2, 3, 4,~~ **5, 6, 7, 8, 9, 10, 11, 12, 13, 14,** ~~15, 16, 17, 18~~	Kaernbach's (1990) Monte Carlo sim.
Rule 5	~~1, 2, 3, 4, 5, 6, 7, 8, 9, 10,~~ **11, 12, 13, 14,** ~~15, 16, 17, 18~~	Ries & DiGiovanni (2009) Human subjects

Note: In calculating DL across the rules evaluated (column 1), the reversals dropped are shown in grey and strikeout, while those used are shown in bold. The basis for the rule is indicated in the final column; 'matrix count' indicates that the rule is derived from data collected in this study

first version (MCS H/L), participants indicated, by pressing a button, whether the comparison tone was higher (H) or lower (L) than the standard tone's frequency. In the second version (MCS S/D), the participants indicated whether the comparison tone was the same or different in pitch as the standard by pressing either a button labeled S (for same) or D (for different). A run in either MCS procedure consisted of 10 trials at each of the fixed frequency differences, and 50 trials at 0 Hz difference to obtain an estimate of the participant's response bias. Each run using either MCS procedure consisted of 100 trials (10 presentations × 5 fixed frequency differences = 50 'signal' trials + 50 'catch' trials at 0 Hz). This approach to the frequency of the comparison tone uses fixed increments and decrements, and is referred to as 'Fixed I/D' in this study (see Table 1). Each run took approximately 35–45 min to complete, including intermittent breaks offered after every 20 trials. For Experiment 2, three additional runs were completed, two of which replaced the original runs in the formal data analysis (2/80 = 2.5% of the data collected).

Sigmoidal functions fitted to the MCS data were used to determine the 75% correct point on the psychometric function. For the H/L response format, the values associated with the zero frequency difference were anchored at 50% correct chance performance for this response format (as an unbiased listener is equally likely to report the comparison tone was 'higher' or 'lower' when provided with a 0 Hz frequency difference). For the hybrid S/D response format, the fitted function was anchored at 0% correct as the chance of a participant responding 'different' increases as the frequency difference increases from zero to maximum performance. The frequency DL was considered to be the point on the sigmoidal function that produced correct responses 75% of the time. The zero frequency difference scores obtained using the MCS procedure were used to monitor response bias such that unbiased responses would result in 50% accuracy for these trials.

For MCS H/L, d' and c cannot be calculated due to the lack of CR for the catch trials based on the response format; however, bias towards a participant reporting the catch trial (0 Hz different trials) is likely to be reflected in the mean. The value reported for comparison to c (calculated for the SIAM and MCD-S/D), is referred to as the *c_estimate*. A *c_estimate* = 0 indicates the participant was equally likely to respond higher or lower on the 50 trials with 0 Hz difference, shown with an asterisk. Positive values reflect a bias toward responding that a comparison tone was higher in pitch that the standard. Data from the SIAM procedure was manipulated following rule 2 discussed in Experiment 1 to calculate DL and performance measures (d' and c).

Design and data preparation

Prior work using classical psychoacoustic methods has meant that, due to economic considerations, recruitment involved only a few participants who completed several hours of data collection within each experiment [6]. Instead, in this work, several hundreds of trials were collected per participant and measures of sensitivity (d') were computed from signal detection theory [6,29]. To ensure that the calculation method employed similar participant criteria, d' and c values were calculated when applicable. Both d' and c are measures of sensitivity or bias [13]. Following the recommendation by Brown & White, calculation of d' and c used a correction method adapted to avoid infinite discriminability (such as H = 100 and M = 0; CR = 100, and FA = 0). If a participant produced zero for any given category, a minimum value of 0.25 was added to all response counts (H = 100.25, M = 0.25, CR = 100.25, and FA = 0.25) in the calculation [37].

Additionally, retaining participants across each condition and experiment was prioritized over sample sizes. To keep experimentation time at a manageable level, small numbers of subjects were tested (N = 2 to 10) [38–40]. Therefore within-subject designs with Geisser–Greenhouse adjustment to control for Type I errors due to small sample sizes were employed [6]. An a priori power analysis using Gpower 3.1.9.4 indicated a minimum of 3 participants would be required to set the power at 0.80 for an F-test, with repeated measures ANOVA used within factors to identify a large effect size (i.e., ≥.70). In this way, 5 participants are sufficient to identify an effect size $f \geq 0.4$ (Critical f = 1.84). The use of planned comparisons (Experiment 2) and RM MANOVA (Experiment 3) reduced the likelihood of Type II errors, and so both approaches were employed. A minimum of 5 participants in each experiment was required to reduce Type II and Type I errors (Beta >.80 (Gpower v3.1.9.4) for planned within-subject factors.

Experiment 1: SIAM calculation for frequency DL

The purpose of Experiment 1 was to establish the optimal calculation protocol for the adaptive measurement runs. Using the SIAM procedure, there are two factors that need to be considered when calculating DLs: 1) the specific reversal points to average within a given run; and 2) the number of runs to collect which must be averaged to give the DL value. The most prudent rules for calculating the frequency DLs for the SIAM procedure in a delayed comparison task are not explicitly defined, but Table 2 incorporates three reversal rules derived from prior work [7,12]. In addition to these rules, matrices were planned for each run collected using the SIAM. There were a total of 80 matrices (2 metrics × 5 participants × 4 runs × 2 ICI conditions). The metrics, frequency DL, and standard deviation (*SD*) for each independent run were based on all possible combinations of reversals points, which were inspected to determine if additional reversal rules might emerge as optimal for each frequency DL calculation method. Rule 1 is the most efficient protocol because the participant undertakes the least amount of trials and reversals (see Table 2 for rule 1); we postulate that the most efficient calculation rules may sacrifice consistency of DL values.

Data preparation procedure

According to Kaernbach's [7] computer simulations, runs with greater than 16 reversals produce approximately the same total error regardless of whether the experimenter drops the first 2 or 4 reversals. However, for experiments involving humans, Kaernbach [7] opted to drop the first 2 reversals out of a total of 18 reversals. Therefore, the two rules adapted from Kaernbach's work were evaluated in comparison to the approach used by

Ries & DiGiovanni [12], with the expectation that one would provide the optimal balance of efficiency and consistency within frequency DL measurements. At the same time as these three reversal calculation rules were being investigated, the data was further inspected to determine if any additional reversal rules resulted in more efficient (fewer trials) or smaller DL values, such as those produced in the MCS in prior work.

Reversal rules were evaluated for the calculation of frequency DL by generating matrices for each run collected in the SIAM. Each cell in a matrix represented a different calculation scheme (i.e., rules to obtain a participant's DL). When developing the matrices, only the first 18 reversal points were included, as all 5 participants completed at least 18 reversals across the 100 trials in each of their 8 runs in the SIAM. Additionally, all calculations dropped the first 4 reversals collected, which used a larger SIAM frequency adjustment step size. All possible remaining reversal rules then averaged at least 4 reversal points thereafter (Note: total # reversals – # dropped reversals = # of reversal points used to calculate mean and variability of reversal points). The cells within the matrices containing the lowest DL mean and *SD* were identified for each of 40 matrices for each metric (i.e., 2 conditions × 4 runs × 5 participants). This process was conducted for both frequency DL and SD across reversal points for each of the 4 independent runs. Rule 1 produced the lowest DL value 10% of the time (4/40) and the smallest *SD* 12% of the time (5/40). Rule 2 produced the lowest DL value 53% of the time (21/40) and the smallest *SD* 38% of the time (15/40). This finding was consistent with those reported by Ries & DiGiovanni [12]. Therefore, five rules were analyzed further to determine which reversal rule provided the most stable and consistent result (see Table 2 and Figure 2). Figure 2 illustrates each of the five rules reported in Table 2 for participant 1. The performance measures, H rate and FA rate, were used to estimate both d' and c for all reversal rules tested (see [27] and [13] for calculation details).

Results

The average DLs were calculated across the four runs using each of the five rules. A repeated measure analysis of variance (ANOVA), with the rules serving as the independent variable and the DL values as the dependent variable, was conducted. A Geisser–Greenhouse adjustment (G-GA) was applied to correct for a violation of sphericity within the data and small sample size. The results of the analyses suggest there was a significant difference for ICI condition ($F(1,4) = 17.35$, $p = 0.014$; observed power = 0.89), indicating that, regardless of reversal rules employed across trials, the NoINT condition will always produce lower DL values (M = 12.48 Hz; SD = 3.10 Hz; SE = 0.62 Hz) than the ToneINT condition (M = 34.79 Hz; SD = 9.87 Hz ; SE =1.97 Hz).

Results indicated a significant difference across reversal rules [$F(4,16) = 6.62$, $p = 0.028$; observed power = 0.96]. Fisher's protected-t Least Significant Difference (LSD) Multiple-Comparisons Tests with an alpha level of 0.05 were used for post hoc pairwise comparisons [41]. Results of the post hoc analysis showed that the frequency DL values calculated using rules 1 (M = 25.32 Hz; SE = 4.78 Hz), 3 (M = 23.56 Hz; SE = 4.12 Hz), and 4 (M = 24.17 Hz; SE = 4.54 Hz) were significantly larger than those produced by rule 2 (M = 21.96 Hz; SE = 4.22 Hz). Rule 5 (M = 23.21 Hz; SE = 4.03 Hz) was not significantly different from rule 2. It appears that both rule 2 and rule 5 allow the listener to hone their measurement; however, the use of rule 2 resulted in smaller DL values without a concomitant increase in variability (SD). Therefore, the remainder of the data for the SIAM procedure tasks were calculated using rule 2: 18 reversals were collected, and only the last 4 reversals were averaged in to the participant's DL value (*Mreversals* 15–18).

Runs analysis

The DL values obtained from the four run options were analyzed using a repeated measures ANOVA with the numbers of runs (run option) serving as the independent variable and the DL values as the dependent variable. A Geisser–Greenhouse adjustment (G-GA) was applied to correct for a violation of sphericity within the data and the small sample size. The results of the analyses showed no significant difference for run option ($F(2,8) = 1.05$, $p = .37$; observed power = 0.18). Therefore, there was no significant difference in DLs across conditions and participants with the SIAM procedure using either 2 runs (M = 20.44 Hz; SE = 3.97 Hz), 3 runs (M = 21.81 Hz; SE = 4.48 Hz), or all 4 runs (M = 21.55 Hz; SE = 4.29 Hz).

To ensure the calculation method was employing similar participant criteria, d' and c values were calculated for each of the possible run options for rule 2. As rule 2 reduced the number of trials contributing to the calculation of DL values, a correction method was adopted (described in *Design and data preparation*). Table 3 reports the three different run averaging options (i.e., 2 runs, 3 runs, 4 runs) for the two conditions employed to compute overall DL values, standard error, mean d', and mean c across participants. Within each condition there was a small range of d' values (NoINT = 1.6 to 1.71, ToneINT = 1.42 to 1.65), indicating that the perceptual distance between the standard and comparison stimuli were adjusted only slightly as additional runs were added into the calculation. Additionally, the range of c values were small (NoINT = –0.01 to –0.10, ToneINT = <.01 to 0.09), reflecting small deviations from the neutral point (i.e., $c = 0$) criterion across participants for both conditions.

Discussion

The null hypotheses for Experiment 1 suggested that the reversal rules would not significantly influence the frequency DL estimates. The results support rejection of the null hypothesis. The null hypothesis, that fewer than four SIAM runs would significantly affect the frequency DL estimates, is also unsupported by the results. The findings from Experiment 1 indicate that various factors can influence the absolute DL values obtained by using the SIAM procedure. The two factors explored in this experiment were: 1) the implemented reversal calculation rule; and 2) the number of runs averaged to estimate a participant's DL value. The SIAM procedure is designed so that the outcome converges on a designated point within the psychometric function, depending on the adaptive adjustment values. These adjustment values were determined

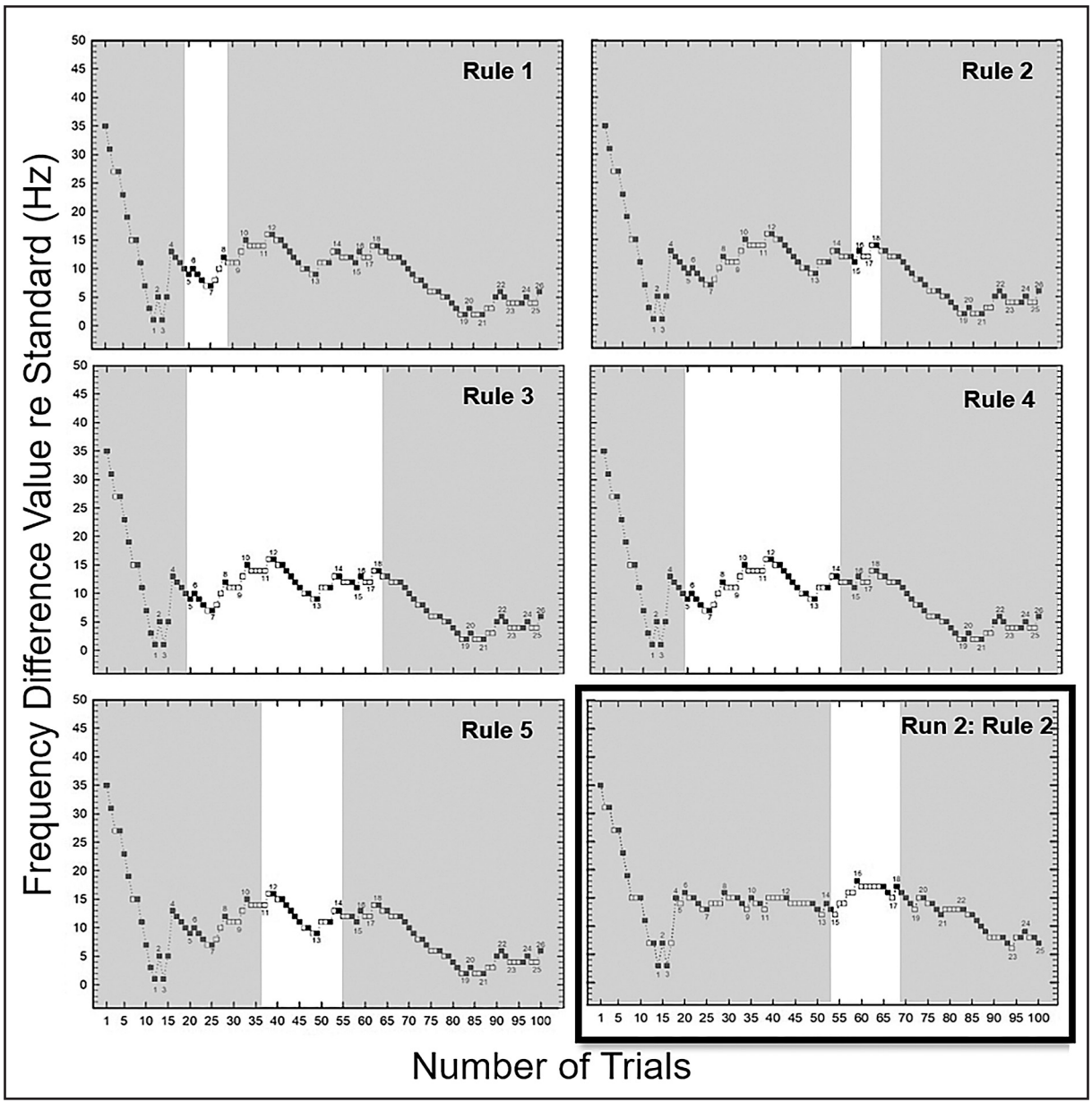

Figure 2. Reversal data for participant 1. The first five plots (labeled Rule 1 – Rule 5) show the reversals included in the calculation of frequency DL for each rule investigated for Run 1. Grey shading indicates reversals not used for the respective reversal rules (see Table 2). The bottom right plot (outlined with a black box) shows results for Run 2 Rule 2 of the same participant (DL = 15.5 Hz). When averaged with the DL obtained from Run 1 Rule 2 (12.5 Hz; top right plot), the estimated DL for participant 1 is 14.00 Hz (SD = 2.12 Hz), which is the recommended frequency DL estimated from the results of Experiment 1

Table 3. Descriptive statistics for the SIAM DL for condition and number of runs (Experiment 1)

		2 Runs	3 Runs	4 Runs
NoINT	Mean	11.55 Hz	11.54 Hz	11.88 Hz
	SE	1.05 Hz	1.44 Hz	1.35 Hz
	SD	2.11 Hz	5.30 Hz	3.61 Hz
	d'	1.60	1.71	1.60
	c	–0.10	–0.01	–0.09
ToneINT	Mean	29.38 Hz	32.08 Hz	31.22 Hz
	SE	5.45 Hz	5.96 Hz	5.85 Hz
	SD	10.44 Hz	11.40 Hz	10.53 Hz
	d'	1.65	1.42	1.54
	c	0.09	< 0.01	0.07

Note: Mean, standard error (SE), and standard deviation (SD) are reported in hertz (Hz). D-prime (*d'*) and criterion measure (*c*) were calculated for each participant and reported as an average across participants. NoINT = No Interpolated Tones; ToneINT = Interpolated Tones. There were no significant differences across the runs included in the DL values within each condition. Calculations involve adding 0.25 correction to all cells (instead of just adding a correction of 1 to the missing cell, see Brown & White [37])

by Kaernbach [7] using the SIAM procedure for measuring tonal thresholds when noise was present. Kaernbach determined that discarding the stimulus values associated with the first 4 reversals of a given run and averaging the next 8 to 10 reversals collected (i.e., rule 4 in this study) resulted in efficient estimation of a listener's performance with a low error rate.

The current study demonstrated that reversal rules are influential in estimating the DL produced during a delayed comparison task, rejecting the null hypothesis. Instead, it was found that discarding the frequency values associated with the first 14 of 18 reversals and averaging only the last 4 reversals (i.e., rule 2 in this study) produced significantly smaller estimates for frequency DL values, and are considered to be more consistent across reversal points. Since auditory discrimination is a more complex task than detection, discarding more reversal points in order to accurately quantify participant performance is logical for a task requiring more difficult auditory processing [24]. It should be noted the calculation rule applied by DiGiovanni & Ries [12] (i.e., rule 5), was not statistically different than rule 2; however, it becomes more efficient as the number of trials are reduced. However, additional factors related to the SIAM procedure would need to be investigated in order to obtain a comprehensive understanding of all the influential procedural choices (e.g., step size adjustments).

Additionally, the number of runs averaged within a DL calculation was investigated to determine if it was an influential factor during the SIAM procedure. When only the first two runs were included, efficiency increased, as fewer total trials were necessary. Including more than two runs provided additional data yet did not significantly alter the DL values produced, and resulted in similar perceptual distance for the task calculated by d'. The value of d' presumes that there are equal variance distributions for both the mean of a signal distribution and of a noise distribution. When attempting to streamline data collection and calculation of d', the number of FA and M are often reduced. Therefore, a correction method needs to be adopted to avoid infinite discriminability. While caution should be taken when using correction values to calculate d', the goal of this experiment was to determine if additional runs significantly altered d' when internal c was used as a reference rather than a dependent variable. Determining that additional runs after the first two did not significantly alter the DL values (or marginally change the perceptual distance for the task calculated by d') may serve to indicate that the SIAM procedure following these recommendations is quicker and more efficient when it uses a rule that terminates data collection following 18 reversals and reduces the number of total runs needed. Similar to observations with the MCS procedure, the DL values were smaller in NoINT than in the ToneINT condition when using the SIAM procedure. This experiment yielded an effective SIAM procedure that could then be used to compare DL values obtained with the MCS procedure, albeit not directly tested and compared within Experiment 1.

These results may indicate that, with the SIAM procedure, learning or adapting to the task is not necessary; thus, bias has less impact on results [7,42]. This indicates that this psychometric function does not require redundancy to stabilize participant performance. As mentioned before, larger d' values can either represent a smaller *SD* of the distributions or a greater perceptual distance between the participant's representation of the standard and comparison distributions [13]. If learning, or adaptation to the tasks, occurred across the four runs, we would expect a systematic change in the d' or c values. It is possible that there is a tendency for a participant's criterion to become more liberal as additional runs are collected (since the values became slightly negative in the NoINT condition), but the change was negligible.

Overall, the results of Experiment 1 indicate that the SIAM procedure approaches the set target point (i.e., the 75% point on the psychometric function) with fewer trials and does not require redundancy of four runs to stabilize participant performance. However, inspection of d' values suggests that in our current sample this is slightly above the 75% point (i.e., d' values produced across all reversal rules were on average greater than 1.0). Streamlining this task is optimal for many reasons, including removal of the potential for inattention or fatigue during data collection [43]. Based on these findings, the SIAM frequency DLs for the remainder of this study were calculated using rule 2 (*Mreversals* 15–18), which were calculated from the first two runs collected following the training procedures indicated above.

Experiment 2: General procedure and response format

The purpose of Experiment 2 was to determine the influence, dictated by the methods (SIAM vs. MCS), that the general procedure and response format have on the measurement of an individual's DL. The measurements of frequency DLs in the two conditions were evaluated by both the MCS and SIAM procedures, as implemented in prior studies [11,12], along with a hybrid procedure that combined the general procedure of the MCS with the SIAM response format used in the aforementioned investigations (i.e., MCS-SD). We expected that, if the general procedure had a major effect on performance, that 1) the use of the MCS procedure would result in similar DL values regardless of response format (i.e., H/L vs. S/D); and 2) DL values would differ significantly between the SIAM and MCS procedures with either response format. Conversely, should the response format dominate performance, it was expected that: 1) DL values would be similar between the SIAM and MCS procedures using the S/D response format; and 2) DL values from the MCS procedure implementing the H/L format would significantly differ from both the SIAM and MCS procedure paired with the S/D responses. This was investigated with two a priori planned comparisons.

Experimental variables

Experiment 2 used the same stimuli explained in the general method, as well as the same timing skeleton (see Figure 1). However, depending on the experimental condition used, the frequency of the comparison tone was either presented at one of several different fixed intervals and was selected randomly per trial (MCS procedure) or adjusted adaptively to target the 75% point of the psychometric

Table 4. Method of constant stimuli (MCS) – comparison of stimulus values

Response Format	No Interpolated Tones (NoINT)	Interpolated Tone (ToneINT)
H/L	10 trials ± 4, 8, 12, 16, 20 Hz 50 trials = 0 Hz	10 trials ± 10, 20, 30, 40, 50 Hz 50 trials = 0 Hz
S/D	10 trials ± 4, 8, 12, 16, 20 Hz 50 trials = 0 Hz	10 trials ± 15, 30, 45, 60, 75 Hz 50 trials = 0 Hz

Note: S/D = Same/Different; H/L = Higher/Lower

function (SIAM procedure). The fixed interval values used in the MCS procedure were determined from pilot data to ensure the responses for frequency differences were obtained at several points across the psychometric function. Table 4 lists the values of the comparison tone used in the MCS procedure. (Note: the difference in hertz of the fixed frequencies of the comparison tones across the MCS procedure's ICI conditions is required to find the frequency DL corresponding to the 75% correct point on the psychometric function.) Participants' frequency DL performance in each condition was measured using three methods: two versions of the MCS that differed in regard to response format (H/L or S/D), and one from data obtained using SIAM. Each participant completed four runs in each of the two conditions (NoINT, ToneINT) for each of the three methods (MCS-H/L, MCS-S/D, SIAM).

Results

The DL values obtained using all three methods were analyzed using a repeated measures ANOVA with ICI condition (NoINT, ToneINT) and method (MCS-H/L, MCS-S/D, SIAM) as the within-participant independent variables and DL as the dependent variable. The G-GA was applied to correct for a violation of sphericity and small sample size. From the full model, significant differences were identified for condition ($F(1,8) = 21.48$, $p = 0.01$; observed power = 0.93) and method ($F(2,8) = 4.64$, $p = 0.045$; observed power = 0.61). There were no significant interactions identified. Table 5 shows descriptive results. A priori planned comparisons, based on the hypotheses used for the three procedures, indicated the MCS-H/L procedure resulted in significantly

Table 5. Descriptive statistics for Experiment 2

ICI condition	MCS-H/L	Hybrid MCS-S/D	SIAM
NoINT			
Mean	8.30 Hz	10.60 Hz	11.55 Hz
SE	1.55 Hz	1.52 Hz	1.05 Hz
SD	3.50 Hz	3.40 Hz	2.11 Hz
d'		1.56	1.60
c	+6.8*	+0.12	–0.10
ToneINT			
Mean	19.6 Hz	26.8 Hz	29.38 Hz
SE	4.35 Hz	2.79 Hz	5.45 Hz
SD	9.70 Hz	6.20 Hz	10.44 Hz
d'		1.33	1.65
c	–3.0*	–0.11	+0.09

Note: ICI = intercomparison interval, SIAM = single-interval adjustment-matrix; *d'* = d-prime, MCS = method of constant stimuli, S/D = Same or Different frequency, H/L = Higher or Lower frequency, NoINT = silent ICI, ToneINT = tonal stimuli in ICI. Mean, standard error (SE), and standard deviation (SD) are reported in hertz (Hz). D-prime (*d'*, the sensitivity index) and the criterion measure in SDT (*c*, the distance from an unbiased neutral criterion) were calculated for each participant and reported as an average across participants
**c*_estimate: For MCS H/L, *d'* and *c* cannot be calculated for the MCS-H/L method due to the response format in which there is lack of correct rejections in catch trials. A value of 0 therefore indicates the participant was equally likely to respond higher or lower on 50 trials with 0 Hz difference, and is shown with an asterisk. Positive values reflect a bias toward responding that the comparison tone was higher in pitch than the standard

Table 6. Planned comparison for response format and procedure (Experiment 2)

Comparisons	MSD	n	SE	t	df
MCS(H/L) vs. MCS(S/D) and SIAM	12.28	5	4.34	2.83**	8
MCS(H/L) and MCS(S/D) vs. SIAM	10.42	5	4.34	2.40*	8

Note: MCS = method of constant stimuli; SIAM = single-interval adjustment matrix, H/L = higher or lower response format, S/D = same or different response format; MSD = mean square difference; SE = standard error. For planned comparisons, the error term in mean square (error) = 3.45 ($N = 5$)
*Significant at an alpha level of 0.05 (one-tailed)
**Significant at an alpha level of 0.025 (one-tailed)

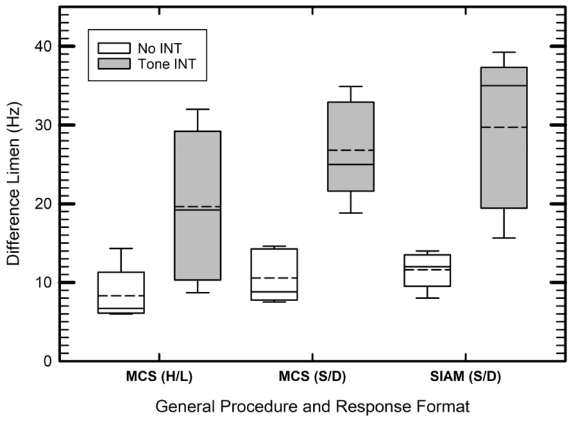

Figure 3. Box plots of limen values in hertz (*y*-axis) across participants for each method and response format (*x*-axis). The horizontal line within each box indicates the median and dashed lines indicate the mean. Upper and lower boundaries indicate the 25th and 75th percentiles, and whiskers indicate the highest and lowest values. White boxes represent performance with condition 1 (NoINT) and grey boxes represent performance with condition 2 (ToneINT)

smaller DL values (M = 13.97 Hz; SE = 2.88 Hz) compared to those obtained across the MCS-S/D (M = 18.68 Hz; SE =3.09 Hz) and SIAM (M = 21.54 Hz; SE = 4.16 Hz) procedures (p =.045). The results obtained using the MCS-S/D and SIAM procedures did not differ significantly from one another. Table 6 provides results of the comparison and Figure 3 illustrates differences in individual participant performance for each condition and method.

Discussion

The extent of interference produced in the two standard tasks and the hybrid task (i.e., MCS-S/D) were quantified through measurement of frequency DLs. As expected, the ToneINT condition, regardless of the procedure, produced significantly larger DL values relative to the other conditions, as illustrated in Figure 3. The results for the NoINT and ToneINT conditions are similar to the patterns of performance reported in previous research using the traditional MCS [11] and SIAM procedures [12].

The outcomes of this study also indicate that response format has a greater influence on DL values than the general method of measurement. Had the general method of measurement (i.e., general procedure) been the dominant factor, DL values obtained between similar methods (i.e., MCS-H/L and MCS-S/D) would have been comparable. Instead, actual DL values obtained using the MCS-H/L and MSC-S/D differed significantly. In contrast, response formats had an effect on absolute DL values obtained using the S/D response format, where the mean DLs were higher than the H/L response format. This difference was likely due to changes in the internal criterion adopted for each response type [13]. Additionally, listener performance did not differ significantly when the two methods (i.e., hybrid MCS-S/D and SIAM) having a common response format were compared. This is likely to occur because the S/D response format tends to make participants adopt a more conservative, stricter criterion for change detection [13]. This stricter criterion produces significantly larger DL values for the SIAM and MCS-S/D compared to the MCS-H/L.

Past research employing S/D response formats has shown that participants are likely to respond "D" only when they are certain they are answering correctly [34]. In contrast, listeners will adopt a more liberal criterion when they are prompted to make a judgment about a stimulus characteristic already assumed to differ along some dimension, as in the MCS-H/L procedure. Adopting the liberal decision criterion resulted in lower DL values, supporting Wickelgren's [33] rationale that different mechanisms may be responsible for S/D judgments. While the present data showed changes in internal criterion values based on response format, we note that internal c cannot be measured in the same manner. A liberal criterion corresponds to a positive c for the MCS-H/L, but to a negative c for the MCS-S/D; therefore, they were not introduced as a dependent factor for this study. Despite the inability to conduct a direct comparison between measured internal c's for MCS-H/L and MSC-S/D, these c values are still worth referencing for discussion based on findings from previous literature that suggest there is a more extreme response bias when employing MCS procedures [44]. These values for Experiment 2 are included in Table 5.

While absolute DL values can differ depending on response format, the general pattern of performance is maintained. Prior work has addressed issues that may arise when streamlining data collection (e.g., reducing the number of trials) and appropriately calculating d' without overestimating the participants' discriminability (i.e., ability to discern two pitches). Experiment 2 clarifies choices made when opting to perform pitch discrimination in AWM and highlights factors that may contribute to frequency DL values (i.e., internal c). Results from Experiment 3, to be discussed later, clarify the effects of response feedback, and increments and decrements (I/D), on the calculation of a listener's frequency DL, d', and c.

We note that our conclusion is speculative, since bias can only be monitored using the MCS-H/L, rather than with the MCS-S/D. When using the H/L response format, listeners already know that the stimulus will differ in some way, and so it makes sense to conclude that a DL value gathered from the MCS-H/L is likely to lead to a more *liberal* criterion and produce a more enhanced (artificial) DL value. While these values provide insight into the limits of discrimination for an individual, the SIAM may be a more reliable DL assessment, resulting in more meaningful (real world) DL values. An advantage of the SIAM procedure is its ability to calculate d' values from fewer trials, which reduces the possibility of listener fatigue. Support for this theory is based on the outcomes and stability gathered in Experiment 1, where DL values did not deteriorate following the addition of more runs.

Experiment 3: SIAM response format, feedback, and frequency adjustment

While Experiment 2 investigated differences in performance across two response formats (H/L and S/D) and

three procedures (MCS-H/L, hybrid MCS-S/D, and the SIAM), the SIAM provides listeners with feedback and the MCS does not. The latter point may have compromised comparison ability, which has been suggested by some researchers who argue that training/learning effects occur when a participant receives feedback, thereby allowing them to correct their internal criterion [13]. To account for this phenomenon, Experiment 3 was developed to examine the effects of feedback on performance measures (i.e., DLs) when using the SIAM. Further, the SIAM reported in Experiments 1 and 2, and in prior work [12], used increments in frequency of the comparison tone rather than both increments and decrements (I/D) used in the MCS (see Table 1). Therefore, six SIAM procedures were designed for Experiment 3 to compare performance by manipulating the following factors: 1) response format; 2) pitch change increments (I) vs. (I/D); and 3) feedback on performance. We expected that:

1) Performance feedback would produce smaller DLs, as the participants could refine their criterion based on the potential for learning (i.e., a training effect),
2) The Yes/No response format SIAM Y/N would not produce significantly different DLs to those of the SIAM S/D, despite a verbal and visual prime, but internal criterion values (*c*) may be altered, based on gaining more knowledge about the frequency comparisons, and
3) The SIAM-I/D condition would not result in significantly lower DLs than SIAM, similar to the hybrid (MCS-S/D) investigated in Experiment 2.

Response format and instructions

Three response formats were used for data collection in both ICI conditions. The written and verbal instructions for each were:

1) SIAM S/D (used in Experiments 1 and 2): "Was the comparison tone different than the standard? Select S (Same) or D (Different)",
2) SIAM Y/N: "Was the comparison higher than the standard? Select Y (Yes) or N (No)",
3) SIAM-I/D Y/N: "was the comparison higher or lower than the standard? Select Y (Yes) or N (No)".

For the last procedure (SIAM-I/D Y/N), the response format was again Y/N. This response format and general procedure differs from the SIAM S/D reported in Experiments 1 and 2 in that the comparison tone could be higher or lower in pitch, whereas with other versions of the SIAM the pitch difference was always higher (see the general methods section).

Procedures and data preparation

For each response format, a version of the SIAM procedure was created to either provide feedback or not provide feedback (hereafter referred to as FB in the procedure titles (see Table 1): SIAM S/D No FB, SIAM Y/N No FB, SIAM-ID Y/N No FB, SIAM S/D, SIAM Y/N, and SIAM-ID Y/N. For the last three versions, feedback was provided visually by turning the correct answer yellow after the participant responded to the stimuli. Participants completed four runs in each of two ICI conditions (NoINT, ToneINT) for all six methods ($2 \times 2 \times 6 = 24$ runs per participant). The rule for calculating the DL and performance measures (i.e., *d'* and *c*) from the SIAM procedure (rule 2) was adopted from the findings of Experiment 1.

Results

Data were analyzed using a multivariate analysis of variance with repeated measures (MANOVA), with the participants' response format, ICI conditions (NoINT, ToneINT), and feedback entered as the independent variables, and DL, *d'*, and *c* entered as the dependent variables. A significant difference between ICI (NoINT, ToneINT) conditions was found on the combined dependent variables ($F(3,70) = 13.13$; $p < 0.001$; Wilks' $\lambda = .64$; $\eta^2 = .36$). Univariate *F*-tests showed that the source of the significant multivariate effect was greater DLs produced in the ToneINT condition ($M = 30.25$ Hz, $SE = 1.43$ Hz) than in the NoINT condition ($M = 16.16$ Hz, $SE = 1.48$ Hz); $F(1,6) = 45.31$, $p < 0.001$; observed power .99. A Geisser–Greenhouse adjustment was applied for violation of sphericity and small sample size. A significant interaction between response format and feedback was found for the combined variables ($F(6,140) = 2.34$; $p = .035$. Wilks' $\lambda = .83$, $\eta^2 = .09$). Univariate *F*-tests indicated significant interaction for *c* ($F(2,12) = 3.69$, $p = .03$; observed power .80). Post hoc Fisher protected-*t* LSD indicated that participants

Table 7. Descriptive statistics for the six SIAM procedures (Experiment 3)

Condition	S/D NoFB	S/D FB	Y/N NoFB	Y/N FB	Y/N ID* NoFB	Y/N ID* FB
NoINT						
DL mean (Hz)	19.75	16.16	14.32	11.11	22.66	12.88
SD (Hz)	9.90	8.46	5.10	3.51	14.83	4.05
SE (Hz)	3.74	3.20	1.93	1.33	5.61	1.53
d'	1.57	1.33	1.25	1.69	1.63	1.21
c	–0.10	–0.01	0.23	0.13	0.03	–0.11
ToneINT						
DL Mean (Hz)	32.04	30.29	25.68	29.75	35.98	27.79
SD (Hz)	18.94	8.89	10.07	11.28	13.39	11.97
SE (Hz)	7.18	3.36	3.81	4.26	5.06	4.52
d'	1.41	1.64	1.60	1.99	1.71	1.96
c	–0.01	0.53	0.33	0.11	–0.03	–0.35

Note: All measures were found to be normally distributed (Shapiro–Wilk *W* test values = 0.82–0.99). Mean, standard error (SE), and standard deviation (SD) are reported in hertz (Hz). D-prime (*d'*) and criterion measure (*c*) were calculated for each participant and reported as an average across participants. *d'* and *c* values were calculated using the same approach described in Experiment 1 if either hit rate or false-alarm rate was found to be infinite

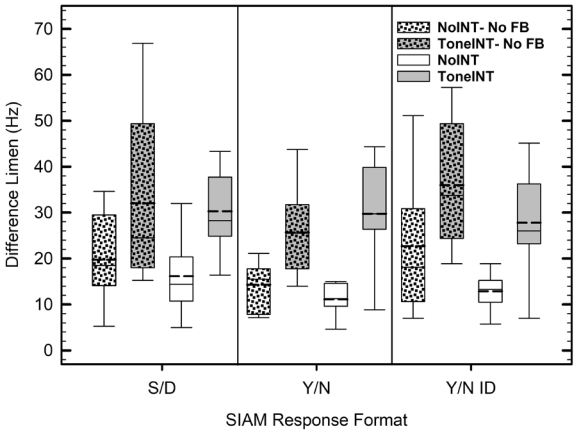

Figure 4. Box plots of difference limen values in hertz (*y*-axis) across participants for each response format used with the SIAM procedure. Key as per Figure 3 with the addition that patterned boxes indicate that no feedback was provided

adopted a more conservative criterion for the SIAM Y/N ($M = 0.20$; $SE = 0.06$) condition and adopted a more liberal criterion during the SIAM-I/D Y/N condition ($M = -.11$; SE = .06; $MSE = 0.12$; $DF = 12$; *Critical value* = 3.77). Performance on the S/D response format was not significantly different for either Y/N or Y/N-I/D ($M = 0.10$; $SE = 0.06$). No other significant differences were identified. Figure 4 provides box plots for the frequency DL measures across the six SIAM procedures investigated, and Table 7 gives descriptive results including d' and c for each SIAM procedure. The omnibus results of Experiment 3 indicate that neither the response format nor feedback affected DL values across the six SIAM procedures.

Discussion

Neither the response format nor feedback produced an overall difference in DL values across the six SIAM procedures/adaptations in Experiment 3. As seen in Experiments 1 and 2, the condition ToneINT produced significantly larger DL values compared to the NoINT condition regardless of response format or performance feedback. Performance feedback did not produce significant differences during the SIAM procedure, contrary to our expectations that feedback would produce differences among the SIAM adaptations. The second expected pattern of results was that response format would not produce significant differences and this was supported by the results. Finally, the data did not support the third expected outcome, that I/D would produce significantly smaller DLs; instead, our findings were similar to the MCS-S/D in Experiment 2.

The lack of impact from response format across the six versions of the SIAM procedure could be due to the possibility that participants adopted the same response criterion for all response formats, regardless of instructions. One could argue that this resulted from a learning effect, since participants received more exposure as the number of trials they completed throughout the study increased. However, this argument is challenged by the lack of significant performance differences in the presence or absence of feedback across all six methods employed for Experiment 3. Because performance feedback did not have a significant effect, this suggests any training effects within the protocol were minimal or non-existent. Additionally, when the identification of a specific change in pitch was requested, the ToneINT condition always increased the DLs. Overall, these findings suggest that pitch retention and discrimination in AWM is less likely to be impacted by training with a given experimental task than previously suspected [4,5,8-12]. This also suggests that performance feedback is not a critical aspect of the current adaptations to the SIAM procedure.

Lastly, there was little variation in the d' produced across the six SIAM protocols. This small range indicates that the participants' discrimination abilities were similar, regardless of the response format and feedback provided. The post hoc results for response bias indicated that participants were biased to adopt a *conservative* criterion in the Y/N response format, meaning they were less likely to say the comparison tone was higher than the standard. On the other hand, for the Y/N-I/D, participants were more *liberal* in reporting whether the comparison tone was higher or lower. Performance on the S/D response format was not significantly different from either Y/N or Y/N-I/D, as c was relatively neutral to both.

General discussion

The results across all three experiments support the overarching thesis that different psychoacoustic methods can influence participant performance during a delayed comparison task, including the measurement of frequency DL. The two common procedures investigated for obtaining frequency DLs included 1) the SIAM and 2) MCS. Researchers have speculated that many factors underlie measurement differences between these two methods. The current study explored the contributions of the following factors: 1) optimal calculation rule for number of sufficient runs (Experiment 1); 2) response format and general procedure (Experiment 2); and 3) feedback and frequency comparisons (I vs. I/D), arising from concerns raised by prior applications of the SIAM procedure (Experiment 3) [12]. This study has also discussed potential roles of AWM and the inclusion of both silent and interpolated tone ICI during a delayed pitch comparison task.

Within the background and rationale for Experiment 1, a range of reversal calculation rules in staircase procedures have been developed and modified by psychoacoustic researchers to increase the efficiency of data collection [7,11,12,45]. Reversal calculation rules specifically apply to the SIAM procedure, as it is an adaptive procedure which searches for the 75% correct point on the psychometric function. Based on data collected in Experiment 1, we have described in the general methods section novel rules which can be applied to the SIAM procedure. These matrices were investigated so as to uncover the optimal reversal points for calculating a listener's frequency DL, after which the optimal rule, rule 2, was used to identify the most efficient number of SIAM runs that can produce a DL value approximating $d' = 1$. Compared to the classical MCS procedure, which called for four runs of 100 trials (the SIAM 4 Runs condition) in Experiment 1 did not produce significant

differences from the DL obtained with only two or three runs. Implementing reversal rule 2 or rule 5 [12] with fewer runs (i.e., two) was more time efficient and minimized listener fatigue.

A consideration related to signal detection theory

The present findings support the idea that, when using the SIAM, there is a reduction in the number of runs necessary to calculate stable frequency DL values. That is, Experiment 1 included more runs but did not significantly alter DL values. Importantly, the number of runs and trials for Experiment 1 were derived within the framework of signal detection theory [29]. A participant's β (decision criterion) is dependent upon the task(s) presented. The optimum β ($\beta = 1$) corresponds to the criterion that an unbiased participant would adopt when responding in a discrimination task. The β value is determined by the probability of the standard and comparison tones having the same or different frequency. This is combined with the perceived value of H and CR, and compared to the cost of an M and FA. Incorporating these factors results in a criterion (on a continuum between conservative and liberal) that the participant has adopted. If the costs associated with incorrect responses, M and FA, outweigh the value of the participant's correct responses, H and CR, the participant's β will fall on the stricter, more conservative end of the continuum, therefore resulting in larger DLs. Probabilities and perceived values for H, CR, M, and FA can be affected by the number of possible trials in a run.

As the SIAM calculation tends to reduce the number of trials, we recommend, if a value of 0 is obtained for any response, adding a constant to each response count (i.e., H, CR, M, and FA) within the signal detection matrix, regardless of the number of trials conducted [37,46,47]. Adding a small value (e.g., 0.25–1) to each response count is a conservative correction to resolve this issue; however 0.25–0.5 is optimal (e.g., M = 0, H = 10, CR = 5, FA = 1, corrected to M = 0.25, H = 10.25; CR = 5.25; FA = 1.25) [37]. Although this correction likely underestimates the true *d'*, it still provides the "best guess" of discriminability [37,46,47]. The results of this study indicated little variation in the d' produced across SIAM runs in both Experiments 2 and 3. This supports the presence of similar discrimination abilities regardless of the number of runs undertaken; however, absolute frequency DL differences persisted across experimental response formats (S/D, H/L, and Y/N).

Experiment 2 is useful in presenting two forms of the MCS procedure (standard: MCS-H/L and hybrid: MCS-S/D) as well as the SIAM S/D procedure. Response format was introduced as a variable across all three experiments and was used to inspect potential bias and its effect on the internal criterion of the listener [13,44]. To reiterate, the internal criterion underlies selection and judgements of perceived pitches. In this study, participants were instructed to make S/D, H/L, or Y/N comparisons to standard tones across experiments and trials. The MCS traditionally employs a H/L judgement, while prior SIAM procedures have been developed using S/D judgements. H/L judgements tend to result in a *liberal* criterion, as they prime listeners that there will most likely be a difference between a comparison and standard tone. In contrast, the SIAM S/D response format induces a *conservative* criterion as listeners are *less likely* to judge a comparison tone as "different" than a standard [34]. These trends were reflected in the overall findings of this study. Results showed significant differences in frequency DL values when comparing both the MCS-S/D and the SIAM to the MCS-H/L procedure. No notable differences were observed between the MCS-S/D and the SIAM procedures, indicating that the general method has less effects on DL measurements than the response format. This finding was supported by the results from Experiment 3 with the SIAM-I/D Y/N, where a liberal criterion was adopted in the presence of tones varying in pitch both above and below the standard tone. These findings support existing reports of bias [11,12] and suggest that the general method (i.e., MCS or SIAM) did not affect frequency DL measures when the response format was the same (i.e., S/D).

Considerations related to performance feedback

The addition of Experiment 3 allowed us to draw novel connections between both the MCS and SIAM procedures, as the SIAM traditionally provides feedback and the MCS does not. Effects of feedback or *lack* of feedback has been presented in discussions regarding delayed pitch discrimination performance [11,12], where it is argued that the presence of feedback influences learning and listeners adopted criterion for constructing judgements. Should feedback induce a form of learning, frequency DLs would have been larger without compared to with feedback. However, results from Experiment 3 showed no effects of feedback across six conditions of the SIAM procedure varying in response format. These findings contrast existing concerns about the effects of feedback on frequency DL values [11,12]. Following analyses of this study's three main factors we conclude that: 1) a fewer number of runs are sufficient to elicit reliable frequency DL values (following rule 2); 2) response format impacts performance more than does the procedure used; and 3) the effects of feedback are minimal at most. As well as these three factors, this study also explored contributions from AWM and the contents of ICI (i.e., NoINT vs. ToneINT).

Considerations related to auditory working memory

To reiterate, AWM is responsible for maintaining representative traces of sounds heard, so that assessments of ICI can later be made. Should representations in AWM not be accurate, correct comparisons cannot be made. The integrity of representations of AWM may also relate to task duration or central effort. For example, a comparison tone varying by 1 Hz to a standard tone may be salient at short time intervals, in the absence of secondary stimuli or during a short task; however, AWM can no longer maintain precision as the time interval between presentations of stimuli increases, as secondary stimuli are introduced, or tasks become protracted. Due to the importance of AWM for successful pitch discrimination, especially in the presence of any delay, psychoacoustic researchers must consider the duration of the task they are presenting. Results from Experiment 1 show that reductions in the number of runs do not limit frequency DL values. Given the connection between AWM and discrimination ability, fewer runs may allow AWM to function optimally for better frequency DL values.

Across all three experiments, a similar pattern emerged, suggesting the involvement of AWM for success in frequency discrimination tasks, specifically concerning contents of the ICI (NoINT vs. ToneINT). Deutsch's [3] model of pitch memory further highlights the effects of increased time between presentation of a standard and comparison tone. Later studies have investigated the dissociative effects of filling an ICI with silence versus with a competing tone. Studies introducing tones show that new stimuli interfere with old stimuli, which spoils the integrity of pitch memory. Across all three experiments, frequency DLs were increased most in the presence of the ToneINT condition compared to the NoINT condition, regardless of general procedure, response format, or number of runs. In sum, frequency DL measurements are: 1) equally valid with a reduced number of runs; 2) affected more by response format than general procedure or method; 3) unaffected by the presence of feedback; and 4) most affected by interpolated tones compared to silent intervals due to their interference with stored representations in AWM.

Limitations

When using the MCS, internal criteria cannot be measured in the same manner as in the SIAM procedure. If the zero difference trials (catch trails) in the MCS are good indicators, it appears that the same individual is more likely to say that two tones are "Different" (M = 13.7 Hz) than decide that the first tone is "Higher" (M = 6.8 Hz). That is, they are likely to adopt a more liberal criterion, which will affect the frequency DL estimate. Alternatively, the SIAM matrix reported by Kaernbach [7] was developed for threshold detection and used in this study to target the 75% point on the psychometric function (i.e., $d' = 1$) for discrimination (frequency DL). The d' values calculated from this sample were found to be greater than 1.0, regardless of the manipulation of the SIAM used to calculate frequency DL (e.g., Experiment 1 reversal rules and number of runs; and Experiment 3: response format, pitch change increments (I) vs. (I/D), and feedback on performance). A limitation of the current work is that d' of the MCS-H/L cannot be calculated as CR are not produced. We note that the SIAM-I/D Y/N tended to produce the closest d' value to the target (i.e., $d' = 1.21$); however this was not statistically significant. Based on the current study's findings, it would be worthwhile to investigate larger samples with more efficient protocols and narrower focus. If such work is pursued, then different SIAM adjustments for target performances should also be explored (see [7] for details).

The general procedure of the MCS requires a decision to be made about the frequency differences selected (fixed I/D). Pilot data precluded use of the same fixed I/D frequency differences selected for the ToneINT condition in Experiment 2 (as the psychometric function derived from the MCS-S/D failed to incorporate the 75% point). Therefore, the interval differences had to be expanded to allow for the appropriate performance points to be reached. This pilot finding was an additional indication of the influence that the response format had on the participant's performance when using the same general procedure. For future studies, this preliminary finding should also be noted as a potential limitation for comparing the general procedure of the MCS. Note that one deviation from the original paper on the SIAM, providing feedback for wrong answers, mentioned its potential use for auditory detection [7]. Such a procedure would not change any button colors on H and CRs, which may reduce relevance to trial-by-trial performance. The current study replicated the feedback method reported by Ries & DiGiovanni [12]. In future, the method of introducing feedback should be studied to determine any systematic effect of using feedback in the SIAM.

Another potential limitation was the sample size for this study. However, with nearly all psychoacoustic study, the amount of time required of participants is extensive and is a major reason why samples sizes tend to be small. The current study took approximately 10–12 hours of data collection per participant for Experiments 1 and 2, and about 5–6 h for Experiment 3. While an a priori power analysis supports the use of our chosen sample sizes, this work needs to be interpreted with care, and future studies using modern psychoacoustics approaches should strive for samples that support generalizable performance measures. We believe the current findings, while preliminary, will support and inform work with larger samples. Future research is necessary to extend the knowledge base in regards to the effect of response format on the criterion adopted by a participant when using the SIAM procedure, as well as the general applicability of the frequency DL calculation methodology applied here to other measures of differential sensitivity in studies of AWM. Monte Carlo simulations of various measurement runs and rules would be a promising future step in refining measurement recommendations for streamlining frequency DL data collection.

Conclusions

Based on this preliminary study, adaptations to the reversal rules and number of runs required for the SIAM allow it to be an alternative data collection procedure to the MCS for determining frequency DLs, notably in fewer trials. Additionally, the SIAM has the advantage of producing CR values required to calculate d' and c. The SIAM procedure encompasses the efficiency of fewer runs while the response format and feedback are still sufficient to obtain reliable DL values similar to those of the MCS procedure. We recommend using 2 runs for the SIAM and S/D response format to maximize the utility of the test. The response format had a notable effect on the absolute DL values obtained with a Same or Different frequency format, which resulted in higher mean DL values than with a Higher or Lower frequency format. Whereas feedback did not make a notable difference, it did provide values closer to scores obtained with the MCS.

It is also concluded that regardless of the procedure used, the DL values were smaller in the NoINT condition than in the ToneINT condition. Although we did not require or specifically ask the participants to give comments on the study, we think it noteworthy that participants did report that when feedback was provided it made the tasks feel more "game-like" and "enjoyable" for them. Applications of the SIAM can be programmed so that d' is calculated with optimal correction values computed for each run; this would provide feedback on participant bias for each run with little calculation effort. Traditionally, the assessment of

perceptual differences on frequency DL uses the MCS; however, with a more efficient method of collecting DL values, participant factors would be easier to assess and control for in future studies. The current study found that the SIAM procedures seem to produce comparable DLs to the ones measured with the MCS procedure, but with fewer trials.

References

1. Barzelay O, Furst M, Barak O. A new approach to model pitch perception using sparse coding. PLoS Comp Biol, 2017;13(1): e1005338.
2. Plack CJ, Oxenham AJ, Fay RR. Pitch: Neural coding and perception. Springer, 2006.
3. Deutsch D. Mapping of interactions in the pitch memory store. Science, 1972; 175(4025): 1020–22.
4. Cowan N. Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. Psych Bull, 1988; 104(2): 163.
5. Cowan N. Attention and Memory: An integrated framework. Oxford University Press, 1998.
6. Oberfeld D, Franke T. Evaluating the robustness of repeated measures analyses: the case of small sample sizes and nonnormal data. Behav Res Meth, 2013; 45(3): 792–812.
7. Kaernbach C. A single-interval adjustment-matrix (SIAM) procedure for unbiased adaptive testing. J Acoust Soc Am, 1990; 88(6): 2645–55.
8. Deutsch D, Feroe J. Disinhibition in pitch memory. Percep Psychophys, 1975; 17(3): 320–24.
9. Schellenberg EG, Trehub SE. Frequency ratios and the discrimination of pure tone sequences. Percep Psychophys, 1994; 56(4): 472–78.
10. Schellenberg EG, Trehub SE. Children's discrimination of melodic intervals. Devel Psych, 1996; 32(6): 1039.
11. Ries DT, DiGiovanni JJ. Release from interference in auditory working memory for pitch. Hear Res, 2007; 230(1-2): 64–72.
12. Ries DT, DiGiovanni JJ. Effects of recurrent tonal information on auditory working memory for pitch. Hear Res, 2009; 255(1-2): 14-21.
13. Macmillan NA, Creelman CD. Detection Theory: A user's guide. Psychology Press, 2004.
14. Deutsch D. Dislocation of tones in a musical sequence: a memory illusion. Nature, 1970; 226(5242): 286–86.
15. Deutsch D. Tones and numbers: specificity of interference in immediate memory. Science, 1970; 168(3939): 1604–05.
16. Massaro DW. Retroactive interference in short-term recognition memory for pitch. J Exp Psychol, 1970; 83(1p1): 32.
17. Harris JD. Pitch discrimination. J Acoust Soc Am, 1952; 24(6): 750–55.
18. König E. Effect of time on pitch discrimination thresholds under several psychophysical procedures; comparison with intensity discrimination thresholds. J Acoust Soc Am, 1957; 29(5): 606–12.
19. Elliott LL. Pitch memory for short tones. Percep Psychophys, 1970; 8(5): 379–84.
20. Ross D, Olson IR, Gore JC. Absolute pitch does not depend on early musical training. Ann NY Acad Sci, 2003; 999(1): 522–26.
21. Ross D, Olson IR, Marks LE, Gore JC. A nonmusical paradigm for identifying absolute pitch possessors. J Acoust Soc Am, 2004; 116(3): 1793–99.
22. Wichmann FA, Hill NJ. The psychometric function: I. Fitting, sampling, and goodness of fit. Percep Psychophys, 2001; 63(8): 1293–313.
23. Levitt H. Transformed up–down methods in psychoacoustics. J Acoust Soc Am, 1971; 49(2B): 467–77.
24. Gelfand SA. Hearing: An introduction to psychological and physiological acoustics: CRC Press, 2017.
25. Wier CC, Jesteadt W, Green DM. Frequency discrimination as a function of frequency and sensation level. J Acoust Soc Am, 1977; 61(1): 178–84.
26. McNicol D. A Primer of Signal Detection Theory. Psychology Press, 2005.
27. Brophy AL. Alternatives to a table of criterion values in signal detection theory. Behav Res Meth Instr Comput, 1986; 18(3): 285–86.
28. MacMillian N, Creelman C. Characteristics of detection theory, threshold theory, and "nonparametric" indexes. Psychol Bull 1990;107: 401–13.
29. Green DM, Swets JA. Signal Detection Theory and Psychophysics. Wiley, 1966.
30. Lynn SK, Barrett LF. "Utilizing" signal detection theory. Psychol Sci, 2014; 25(9): 1663–73.
31. Stanislaw H, Todorov N. Calculation of signal detection theory measures. Behav Res Meth Instr Comput, 1999; 31(1): 137–49.
32. Furnham A. Response bias, social desirability and dissimulation. Pers Individ Diff, 1986; 7(3): 385–400.
33. Wickelgren WA. Associative strength theory of recognition memory for pitch. J Math Psychol, 1969; 6(1): 13–61.
34. Gerrits E, Schouten M. Categorical perception depends on the discrimination task. Percep Psychophys, 2004; 66(3): 363–76.
35. ASHA. Guidelines for the audiologic assessment of children from birth to 5 years of age, 2004.
36. Bull AR, Cuddy LL. Recognition memory for pitch of fixed and roving stimulus tones. Percep Psychophys, 1972; 11(1): 105–09.
37. Brown GS, White KG. The optimal correction for estimating extreme discriminability. Behav Res Meth, 2005; 37(3): 436–49.
38. Duncan J, Humphreys GW. Visual search and stimulus similarity. Psychol Rev, 1989; 96(3): 433.
39. Eriksen BA, Eriksen CW. Effects of noise letters upon the identification of a target letter in a nonsearch task. Percep Psychophys, 1974; 16(1): 143–49.
40. Ernst MO, Banks MS. Humans integrate visual and haptic information in a statistically optimal fashion. Nature, 2002; 415(6870): 429–33.
41. Ryan TA. Multiple comparison in psychological research. Psychol Bull, 1959; 56(1): 26.
42. Shepherd D, Hautus MJ, Stocks MA, Quek SY. The single interval adjustment matrix (SIAM) yes–no task: an empirical assessment using auditory and gustatory stimuli. Attent Percep Psychophys, 2011; 73(6): 1934.
43. Leek MR. Adaptive procedures in psychophysical research. Percep Psychophys, 2001; 63(8): 1279–92.
44. Taylor M, Forbes S, Creelman CD. PEST reduces bias in forced choice psychophysics. J Acoust Soc Am, 1983; 74(5): 1367–74.
45. Klein SA. Measuring, estimating, and understanding the psychometric function: A commentary. Percep Psychoacoust, 2001; 63(8): 1421–55.
46. Watson JE, Blampied NM. Quantification of the effects of chlorpromazine on performance under delayed matching to sample in pigeons. J Exp Anal Behav, 1989; 51(3): 317–28.
47. Jones BM, White KG. Sample–stimulus discriminability and sensitivity to reinforcement in delayed matching to sample. J Exp Anal Behav, 1992; 58(1): 159–72.