A perceptual evaluation of the efficacy of Sound Exposure Level in the rating of annoyance to helicopter noise

Matthew A. Boucher
Research Engineer
NASA Langley Research Center
Structural Acoustics Branch
Hampton, VA, USA

Andrew W. Christian
Research Engineer
NASA Langley Research Center
Structural Acoustics Branch
Hampton, VA, USA

Siddhartha Krishnamurthy
Research Engineer
NASA Langley Research Center
Structural Acoustics Branch
Hampton, VA, USA

Stephen A. Rizzi
Senior Researcher for Aeroacoustics
NASA Langley Research Center
Aeroacoustics Branch
Hampton, VA, USA

ABSTRACT

A psychoacoustic test was performed that aimed to test how well Sound Exposure Level (L_{AE}) is at indicating changes in annoyance to helicopter noise for simple changes in design as well as for realistic changes found from flight tests. In particular, L_{AE} was evaluated for auralizations of optimized designs of rotor geometries when compared to a baseline design and for recordings that compare different helicopters and maneuvers. Paired comparisons consisted of the 10 dBA-down portion of flyovers, which is the same portion used to calculate L_{AE}. When played at the same L_{AE}, annoyance responses showed in which cases L_{AE} is a good indicator as well as when other aspects not included in the calculation of L_{AE} may be important. Annoyance responses for relative differences in L_{AE} allowed the calculation of an Equal Annoyance Point, giving further insight to the performance of L_{AE}. Confidence intervals calculated with Monte Carlo simulations showed when responses are statistically different from the equal L_{AE} comparison and also gave an indication of the necessary reduction for designers to be confident that a low noise design is impactful or that a difference in rotorcraft or maneuver is perceptually favorable.

INTRODUCTION

All helicopters must meet certain noise certification requirements. For light helicopters, the metric used is the Sound Exposure Level (L_{AE}), which is an integration of the sound energy contained in a noise event. Despite successful noise certification, complaints related to helicopter noise still persist (Ref. 1). Highly variable operations, flying low over communities and qualitative aspects of the sound may all be contributing factors to the number of complaints. The implication is that A-weighted, integrated noise metrics do not capture the complete human response to helicopter noise, and laboratory tests have confirmed that annoyance responses correlate with more than just time-integrated measures of loudness such as L_{AE}. The psychoacoustic test presented in this work seeks to answer the question of whether helicopter noise mitigation strategies (such as optimized rotor blade geometry and specified maneuvers) based on L_{AE} lead to undesired changes in sound quality. This paper is the first reporting of results from the 2nd Rotorcraft Sound Quality Metric (RoQM-II) psychoacoustic test, which was completed in December 2019 at the NASA Langley Research Center (LaRC). RoQM-II focuses on potential changes in sound quality (1) when rotor blades are optimized for low noise and (2) due to unsteady sounds common to helicopter maneuvers.

In general, Sound Exposure Level is the integrated, A-weighted sound energy of a noise event over a given time interval. For use as a noise certification metric for light helicopters, L_{AE} is specified in Annex 16 to the Convention on International Civil Aviation (Ref. 2). For flyover events that are sampled at regular time intervals, Δt, L_{AE} can be approximated by

$$L_{AE} = 10 \log_{10} \frac{1}{T_0} \sum_{k_1}^{k_2} 10^{0.1 L_A(k)} \Delta t \quad , \qquad (1)$$

in which $L_A(k)$ is the k-th sample of the A-weighted sound pressure level, L_A. With a reference duration of $T_0 = 1\,\text{s}$, L_{AE} represents the total energy, not the average. The summation is done for the portion of L_A that is within 10 dBA of its maximum, $L_{A,max}$, such that k_1 and k_2 are the first and last samples, respectively, that satisfy $L_A(k) \geq L_{A,max} - 10\,\text{dBA}$.

Presented at the VFS International 76th Annual Forum & Technology Display, Virginia Beach, Virginia, USA, October 6–8, 2020.

In this work, the concept of sound quality refers to any aspect of the sound not taken into account by L_{AE}. It includes well-defined aspects, which can be quantified by sound quality metrics (e.g., sharpness, tonality, etc.), but is not limited to those and may include variations in the spectral, temporal or spatial character of the sound that are less well-defined. Instead of studying sound quality directly, this work focuses on the differences in responses to sounds that are presented at the same L_{AE}. The main hypothesis is that when sounds are compared at the same L_{AE}, the variation in subjective responses reveals the efficacy of Sound Exposure Level in the rating of annoyance to helicopter noise. Furthermore, the variation in responses reveals significant qualitative aspects of helicopter noise not captured by L_{AE}.

Motivation and background

The motivation for this psychoacoustic test comes from previous work that suggests noise certification metrics (e.g., L_{AE}) do not fully describe human annoyance responses to helicopter noise and that noise characteristics, not just sound level, are important to their community response. A study of helicopter noise in Norway found that small helicopters were more annoying than a reference fixed wing aircraft of equal L_A (Ref. 3). A study in Switzerland suggested that helicopter landings were slightly more annoying than helicopter takeoffs (Ref. 4). The previous test in the current series (RoQM-I-2017) showed that annoyance to helicopter noise of equal loudness is a function of sound quality metrics, such as fluctuation strength, tonality and sharpness (Ref. 5). All of these tests suggest that helicopter sound quality is an important factor and deserves further study.

Further motivation arises from the need to assess L_{AE} as a perceptually relevant indicator that can be used to evaluate different noise mitigation strategies. Most pertinent to the current work, noise mitigation can be achieved either in the design phase when developing quieter rotors or in the operation phase when a pilot's maneuvers have a predictable effect on the radiated sound (Ref. 6). This psychoacoustic test evaluates L_{AE} as a perceptually relevant indicator of annoyance to helicopter noise originating from both simulations of optimized rotor blades and recordings of different flown maneuvers.

The simulations are a result of helicopter rotors designed for low noise. These sounds are generated through auralization, which is a technique for creating audible sound files from numerical data (Ref. 7). Here, it refers to the combined process of source noise synthesis and propagation to a ground observer. For these auralizations, sounds emitted by helicopter components other than the rotors, such as engine noise and airframe interactions, are omitted. The auralizations from Krishnamurthy et al. (Ref. 8) of AS350 helicopter main rotor flyovers are leveraged for the simulated sounds in this psychoacoustic test. In addition to auralizing a flyover of the original, or baseline, AS350 main rotor, Krishnamurthy et al. (Ref. 8) auralized two more flyovers: that of a main rotor that minimized L_{AE} on the ground and one that minimized Effective Perceived Noise Level (EPNL). EPNL, like L_{AE}, is an integrated noise metric but also includes a rudimentary penalty for tonality (Ref. 2). The optimization process changed the main rotor geometry while keeping the tail rotor constant. Simulated flyover sounds of all three AS350 rotor configurations (baseline, L_{AE}-optimized, and EPNL-optimized) are generated for this psychoacoustic test. The auralized sounds of the optimized rotor flyovers at a ground observer had L_{AE} and EPNL values roughly 12-16 dB lower than the original main rotor flyover. Further, the loudness and roughness sound quality metrics were calculated by Krishnamurthy et al. (Ref. 8) to predict the psychoacoustic response of the auralizations based on an annoyance model from Zwicker and Fastl (Ref. 9). The model predicted the optimized rotor flyovers to be less annoying, but as stated in Krishnamurthy et al. (Ref. 8), this predicted response can only be substantiated by psychoacoustic tests such as the one in this paper.

In addition to the synthesized sounds just mentioned, this psychoacoustic test also investigates human response to field recordings of helicopters performing various maneuvers, some of which having been designed for low noise. The recordings used for the psychoacoustic test are selected from a joint flight test among NASA, the FAA and the U.S. Army that produced recordings of six helicopters, including different models, engine power/size, tail rotor technologies and number of main and tail rotor blades (Ref. 10). The flight test consisted of several different maneuvers, including level flight, climbs, descents, turns under different constraints and low noise approaches. Measurements were made on the ground by an array of 56 microphones, and data from this report have been made available to the public (Ref. 10). Simulations have shown that helicopter maneuvers can greatly affect the radiated noise (Ref. 6). For example, decelerating while entering a turn should be avoided, and noise sensitive areas should ideally be on the inside of the turn and on the retreating side of the helicopter (Ref. 6). A psychoacoustic test with human subjects is needed to investigate how these predictions relate to annoyance.

The main contribution of this paper is the summary and analysis of collected human responses from a laboratory psychoacoustic test that evaluates the efficacy of Sound Exposure Level in the rating of annoyance to helicopter noise. The test subjects are presented with auralizations of noise-optimized rotor blades as well as field recordings of helicopter maneuvers. The auralizations include accurate modeling of the sound source and propagation to a ground observer, while the selection of helicopter maneuvers from a flight test faithfully reproduce the aural effect of various helicopter operations. The spatial impression of moving sources of both auralizations and recordings are accurately presented to test subjects in the Exterior Effects Room (EER) (Ref. 11) at the NASA LaRC. Finally, paired comparisons and annoyance responses make it possible to efficiently evaluate the efficacy of Sound Exposure Level.

TEST PREPARATION

Preparation for the RoQM-II psychoacoustic test consisted of: (1) performing auralizations of different rotor blade designs, (2) selection of field recordings from a recent flight test and (3) the design and execution of the psychoacoustic test in the Exterior Effects Room.

Auralizations

In a departure from Krishnamurthy et al. (Ref. 8), the auralization process for the AS350 helicopter rotor blade flyover test sounds directly uses Farassat's formulation 1A (F1A) (Ref. 12) to generate sound pressures near the rotors before propagation to a ground observer. F1A synthesis, which has been described previously (Refs. 13, 14), avoids audible artifacts that were caused by interpolation of sound pressure magnitudes and phases at discrete prediction points (Ref. 8). This is done by computing noise at the source in the time domain sample-by-sample using F1A synthesis at the instantaneous emission angles between the source and the receiver. Also, only sound pressures at emission angles that propagate to the ground observer are calculated, not the entire dense set of discrete points over the source hemisphere. A need to synthesize aperiodic sounds from maneuvers involving accelerations and attitude changes motivated the development of F1A synthesis, but the process is applied here to synthesize periodic sounds from straight and level flyovers. Due to updated rotor blade loading calculations, the blade loading data from (Ref. 8) is regenerated for this test. These data serve as input to the F1A calculations for the optimized and original main rotor flyover sounds. F1A synthesis is implemented with the NASA Auralization Framework (NAF) (Ref. 15), which simulates the propagation of sound from the source to a receiver.

Auralized flyovers of the AS350 helicopter included periodic sounds from both the main and tail rotors. For the L_{AE}-optimized rotor case, the main rotor geometry was changed, as seen in Figure 1. The tip sweep angles are noticeably different between the baseline and L_{AE}-optimized main rotors. A second noise optimization was done in terms of EPNL. The tip sweep angle of the EPNL-optimized rotor (not shown) is slightly different from that of the L_{AE}-optimized main rotor. Although the tail rotor geometry was kept constant in the optimization process, the trim settings of the tail rotor had to be adjusted when flown with an optimized main rotor. Although the loading conditions on the tail rotor change slightly during this process, the changes are not expected to have an impact on the perception of the tail rotor noise.

The NAF simulated the flyovers of the AS350 helicopter main and tail rotors to generate sound, or auralizations, at a ground observer that was then played to test subjects. Flyovers were straight and level with a constant speed of 47.59 m/s (92.5 kn). Rotors flew along a centerline path directly over a ground observer, which was flush with rigid ground, at an altitude of 150 m (492 ft). The NAF generated one minute long auralizations, inclusive of the initial propagation delay from source to ground. Rotors were directly over the ground observer at approximately 31 seconds into the flyover.

(a) Baseline rotor

(b) L_{AE}-optimized rotor

Figure 1: Main rotor geometries used for auralizations (EPNL-optimized rotor not shown).

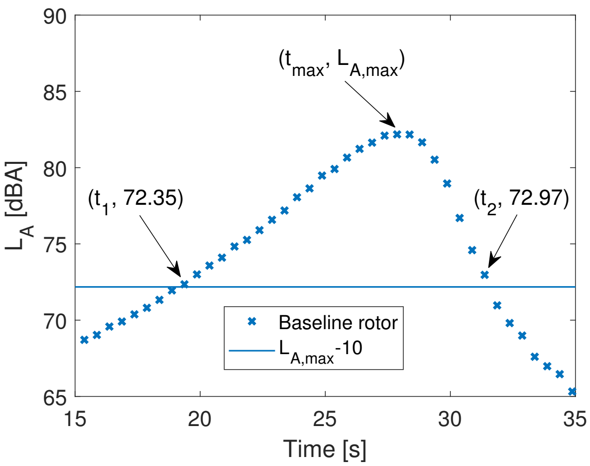
The relevant portion of the flyover for the baseline rotor is shown in Figure 2a. The A-weighted sound pressure level within 10 dBA of its peak is the portion above the horizontal line, which is also the portion used in the psychoacoustic test[1]. Three points during the flyover are noted that are used in the L_{AE} calculation in Eq. (1), which are $(t_1, L_A(k_1))$, $(t_{max}, L_{A,max})$ and $(t_2, L_A(k_2))$.

In Figure 2b, the L_A time histories for the auralizations used in the psychoacoustic test are shown (i.e., only the part that is within 10 dBA of $L_{A,max}$). In contrast to Figure 2a, the time histories begin at 0 s, because t_1 was subtracted from the time series for each auralization. This is done so that differences in duration and shape of the L_A profile are more evident. The two optimized rotor auralizations are almost identical in terms of L_A. They are about 7 s longer than the baseline but have a peak that is 1.67 dBA lower. These apparent differences in duration and amplitude are solely due to a change in the source noise, since all three auralizations were simulated at the same flight speed.

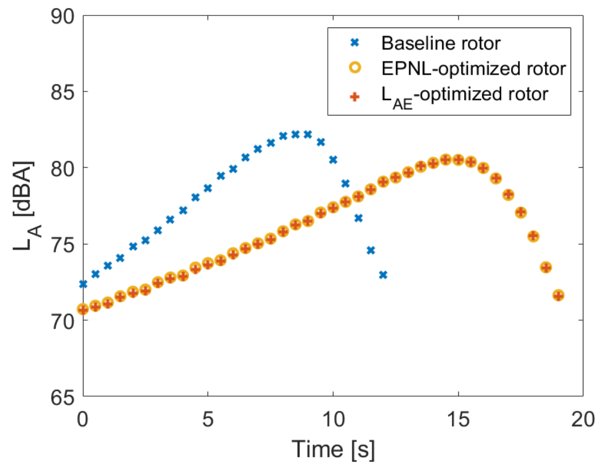
Flight Test Data

In addition to the auralizations just described, recordings from flight tests are also included in the psychoacoustic test. The reason for this is that actual flights contain temporal, spectral and spatial variations in the noise that are not easily simulated, variations that may be relevant in the evaluation of L_{AE}. The recordings that were considered for this test come from a recent flight test involving six different helicopters that included level flights, climbs, descents and different types of turns over a 52-microphone array (Ref. 10). The turns consist of different operational conditions, such as constant speed/torque or

[1] The levels for playback during the psychoacoustic test were lower than shown in order to avoid subject fatigue. See Section Psychoacoustic Test for details.



(a) The maximum (t_{max}) and boundaries of the 10 dBA down interval (t_1 and t_2) are indicated.



(b) The abscissa starts with the first instance within 10 dBA from the peak.

Figure 2: L_A time histories for auralizations used in the psychoacoustic test, normallized to $L_{AE} = 89.8\,\text{dBA}$[1].

acceleration/deceleration, which can affect not only the levels of noise but also the spectral content. Even the direction of the turn in relation to the advancing/retreating side of the helicopter may produce changes in noise signatures (Ref. 6).

The recorded stimuli chosen for this test are from flights of the Eurocopters AS350 and EC130, two helicopters of the same manufacturer with similar size and capabilities but with different tail rotor technologies. Photographs of the two helicopters are shown in Figure 3. Both helicopters have turbine engines with 3 main rotor blades, diameter of 10.69 m (35.07 ft) and a clockwise-rotating main rotor. The largest difference between the two helicopters is that the EC130 has a Fenestron instead of a tail rotor. Both helicopters were recorded at the Amedee Army Airfield in Lassen County, California (Ref. 10).

The recordings of the AS350 and EC130 helicopters that were selected from the Noise Abatement Flight Test are shown in Table 1. The flight, run and microphone numbers are shown, as well as the type of maneuver. The low noise approach (LNA) was developed over several days of the Noise Abatement Flight Test and made use of input from the manufacturer as well as the Fly Neighborly Guidelines from the Helicopter Association International (Ref. 10). This maneuver is used to compare the perceptual differences between the two helicopters. The AS350 steady and unsteady recordings were chosen after extensive, informal listening tests. Here, steady refers to the perception of the overall sound quality of the recording, meaning that although L_A varies throughout the flight, the sound quality is perceived to be quite constant with no sudden changes in temporal, spectral or spatial impression. In contrast, the unsteady recording was perceived to have several different sound characteristics throughout the flight. Sudden changes in temporal, spectral and spatial impression were perceived. Specifically, some portions were more impulsive, others were dominated by a broadband component, while tones were more prominent at other times.

(a) AS350

(b) EC130

Figure 3: Photographs of helicopters whose recorded flights were used as sound stimuli in the psychoacoustic test (Ref. 10).

The L_A time histories for the recorded flights listed in Table 1 are shown in Figure 4. As in Figure 2b, the time series is subtracted by t_1 for each flight, and all are normalized to $L_{AE} = 89.8\,\text{dBA}$. The low noise approaches for the two helicopters have a similar duration and $L_{A,max}$, and the EC130 reaches its peak L_A about 4 s earlier than the AS350. There is a slight increase at the end of the EC130 LNA, because the last

Table 1: Selected recordings from the Noise Abatement Flight Test (Ref. 10).

Stimulus	Flight	Run	Mic	Maneuver
AS350 LNA	291	294	29	low noise approach
EC130 LNA	298	280	29	low noise approach
AS350 steady	290	186	55	constant torque, level turn
AS350 unsteady	290	202	16	turn with acceleration through roll-in

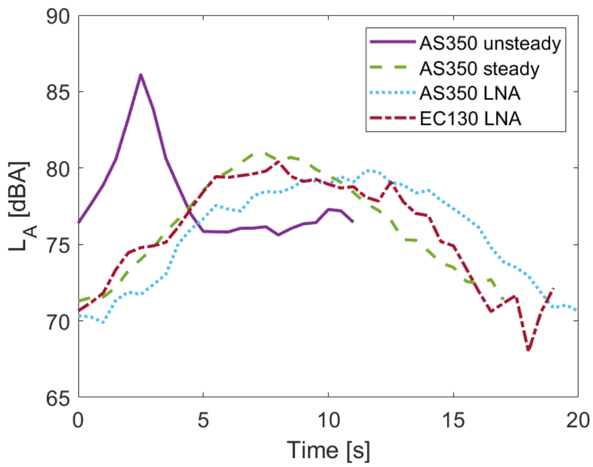


Figure 4: L_A time histories for recorded flights used in the psychoacoustic test, normalized to $L_{AE} = 89.8\,\mathrm{dBA}$[1]. The abscissa starts with the first instance within 10 dBA from the peak.

time sample within 10 dBA of the peak must be included in the L_{AE} calculation (Ref. 2). The steady AS350 flight has a similar profile to the low noise approaches but differs markedly with the unsteady flight. The unsteady flight, which is an accelerating turn, has a steep rise in L_A at the beginning, reaching its peak at around 2.5 s, has a roughly constant L_A after 5 s and a slight rise near the end.

Psychoacoustic Test

The psychoacoustic test was designed around paired comparisons (Ref. 16). The four pairs are detailed below, along with the particular research questions each pair addresses. Each research question (RQ) is a different test of the efficacy of L_{AE} in the rating of annoyance to helicopter noise.

Pair 1: Baseline rotor vs. L_{AE}-optimized rotor

RQ1a: When presented at the same L_{AE}, is there a perceived difference in annoyance between a baseline rotor and one optimized in terms of L_{AE}?

RQ1b: By how much should the L_{AE} of the L_{AE}-optimized rotor (relative to the baseline) be adjusted in order to give an annoyance response equal to that of the baseline rotor?

Pair 2: EPNL- vs. L_{AE}-optimized rotor

RQ2: Is there a perceived difference in annoyance for a rotor that is optimized in terms of EPNL instead of L_{AE}?

Pair 3: AS350 low noise approach vs. EC130 low noise approach

RQ3a: For a recorded low noise approach, is either the AS350 or EC130 perceived to be more annoying than the other when presented at the same L_{AE}?

RQ3b: By how much should the L_{AE} of the EC130 (relative to the AS350) be adjusted in order to give an annoyance response equal to that of the AS350 for a low noise approach?

Pair 4: AS350 unsteady flight vs. AS350 steady flight

RQ4a: For the AS350, is an unsteady flight perceived to be more annoying than a steady flight when presented at the same L_{AE}?

RQ4b: For the AS350, by how much should the L_{AE} of the unsteady flight (relative to the steady one) be adjusted in order to give an annoyance response equal to that of the steady flight?

RQ5: Are there different annoyance responses to full flyovers than there are when comparing short samples of the flyover at t_1, t_{max} and t_2?

The first research question for pairs 1-4 only involve comparing each pair of sounds at the same L_{AE}. If one sound is found to be more annoying than another, it does not quantify the difference in annoyance of the two sounds. To determine that, a different research question is asked, which is the second question for pairs 1, 3 and 4: by how much should the L_{AE} of sound B (relative to sound A) be adjusted in order to give an annoyance response equal to that of sound A? This relative change in L_{AE} is called the Equal Annoyance Point (EAP). To determine this relative level change, it was necessary to compare the two sounds at different levels. For pair 1, the relative levels were 0, ±5 and ±10 dBA. For pairs 3 and 4, the relative levels were 0, ±4 and ±8 dBA. It was determined from pilot testing that these levels covered a wide enough range such that at either extreme, most subjects would agree on which sound was more annoying.

To answer RQ5, short sounds centered at t_1, t_{max} and t_2 were extracted for pairs 1 and 4 and compared at different relative levels. For these short sounds, comparisons were only made between similar points within a flyover (e.g., t_1 of the baseline rotor was compared to t_1 of the L_{AE}-optimized rotor but not to t_{max} or t_2 of the L_{AE}-optimized rotor). The relative levels of the short sounds were the same as those of the long sounds, which were 0, ±5 and ±10 dBA for pair 1 and 0, ±5 and ±10 dBA for pair 4.

The selected comparisons are summarized in Table 2. After each pair was presented, the subject was asked to select the sound that was more annoying.

A total of 16 subjects were tested in groups of 4[2]. There were 30 unique paired comparisons of short sounds (used to answer RQ5). Each A-B comparison was also played in B-A order, doubling the number of comparisons to 60. The order was randomized once, so that each group heard the same order. The short sounds were played to each group over 2 back-to-back sessions. Similarly, there were 16 unique full flyover comparisons (used to answer RQs 1-4). Playing each pair in reverse order doubled the number to 32. These comparisons were randomized once and presented to each group in the same order. The full flyover pairs were also played over 2 back-to-back sessions to each group. Each group listened to all 92 comparisons, which were divided into 4 sessions. Groups 1 and 4 listened to the short sounds first, while Groups 2 and 3 listened to the full flyovers first.

The psychoacoustic test was performed in the EER at the NASA LaRC (Ref. 11). The EER has 27 satellite loudspeakers mounted on the walls and ceilings and 4 subwoofers placed in the corners. Vector-based amplitude panning gives a realistic impression of moving sources, which are given by the simulated trajectory for auralizations or the flown trajectory for the recordings through collected GPS data. The test subjects are positioned as if the center of the EER corresponds to the ground observer/microphone location.

The sampling rate used in the EER is 44.1 kHz. The auralizations were simulated at this sampling rate, but the recordings were up-sampled from 25 or 25.5 kHz. To eliminate clipping at the beginning or end of a stimuli, 2 s and 0.2 s tapers were added to the full flyover and short sound stimuli, respectively. The tapers also limit startling the subjects.

The test design required that the basic comparisons were presented to the subjects at equal L_{AE}. During pilot testing, a set of recordings from the flight test had an average L_{AE} of 89.8 dBA. This was used as the target level, L_T, for equalization of the recordings as well as the auralizations. Therefore, a normalization factor was calculated for each recording, which is given by

$$a = 10^{(L_T - L_{AE,i})/20} \qquad (2)$$

in which $L_{AE,i}$ was the initial sound exposure level of the auralization or recorded flyover. The normalization factor a was then multiplied by the pressure time history of the recording, yielding an L_{AE} value of $L_T = 89.8$ dBA. A similar normalization was done for the short sounds in terms of L_A in which the target L_A was the mean L_A of the two sounds.

In order to avoid subject fatigue, 89.8 dBA was not the desired playback level for the actual test. Instead, a level gain of −22.6 dBA was applied to all sounds, which resulted in an intended L_{AE} of 67.2 dBA for the full flyovers. To verify the intended level, a set of 26 normalized, full flyover recordings from the flight test were played in the EER while a Sound Level Meter at the center of the 4 seats measured $L_{A,max}$ and L_{AE}. Since the flights follow different trajectories, the $L_{A,max}$ is not expected to be the same, while the L_{AE} should be. The mean and standard deviation of the L_{AE} for the 26 normalized recordings as played back in the EER were 67.7 dBA and 0.5 dBA, respectively. The mean L_{AE} value is within 1 standard deviation of the intended playback level, which is considered reasonably accurate for the purposes of this test.

[2]Subjects were required to be at least 18 years of age and to not have significant hearing loss as shown by a pretest hearing screening. A gender balance of between one and two thirds female was also specified. The protocol for the psychoacoustic test was approved by the NASA Langley Institutional Review Board.

ANALYSIS TECHNIQUES

This section describes the techniques used to analyze the collected annoyance responses to the paired comparisons of different sound stimuli. Binomial tests of sounds at the same L_{AE} determine if one sound is more annoying than another. Probit models determine the EAP, that is, the point of subjective equality of annoyance. Monte Carlo simulations provide confidence intervals on EAP. Finally, perceptual adjustments of the L_A time history of flyover events provide an alternative method of determining the subjective equality of annoyance.

Binomial test

When subjects are asked, “Which is more annoying, sound A or sound B?”, their binary responses can be evaluated using a binomial test (Ref. 17). Since both sounds have the same L_{AE}, it is initially assumed that neither sound is more annoying than the other (i.e., the null hypothesis). If the responses indicate otherwise, a significant p-value (i.e., less than 0.05) suggests, although does not prove, that the null hypothesis should be rejected. This means that it would be highly unlikely that the responses resulted by chance alone and that there is something inherent about the sound, other than L_{AE}, that affected the subjects’ annoyance choice. In what follows, two-sided binomial tests are used, because there is no assumption made about which sound in each pair may be more annoying than the other. If however, there was an indication (e.g., public complaints) that a particular sound was more annoying than another, a one-sided binomial test could be used.

Equal Annoyance Point

The binomial test helps determine whether two sounds of equal L_{AE} are perceived to be equally annoying or not. If the null hypothesis can be rejected, the binomial test says nothing about how much more annoying one sound is compared to another. To answer this question, more sophisticated analyses are needed. The goal of the analysis described here is to

Table 2: Description of comparisons made in the psychoacoustic test.

Pair	Sound	Description	Auralization	Recording	Full flyovers	Different levels	t_1, t_{max} and t_2
1	A B	Baseline rotor L_{AE}-optimized rotor	yes	no	yes	yes	yes
2	A B	L_{AE}-optimized rotor EPNL optimized rotor	yes	no	yes	no	no
3	A B	AS350 LNA EC130 LNA	no	yes	yes	yes	no
4	A B	AS350 steady AS350 unsteady	no	yes	yes	yes	yes

determine the change in L_{AE} required such that both sounds are perceived to be equally annoying.

For two sounds A and B, assume sound A is presented at a constant L_{AE} and that B is varied up or down in amplitude. When sound B has a higher L_{AE}, it is often, but not always, determined to be more annoying than sound A. Similarly, when sound B has a lower L_{AE}, it is more often judged to be less annoying. For such patterns, logistic regression is often employed. In this work, a probit model is fit to this binary response data using a maximum pseudo likelihood approach (Ref. 18).

In the probit model, the link function gives the relationship between the mean response and a linear combination of predictors. The inverse link function gives the probability of one sound being judged more annoying than the other and is given by

$$\Pr(B \succ A \,|\, \Delta L_{AE}, \beta_0, \beta_1) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{\beta_0 + \beta_1 \Delta L_{AE}}{\sqrt{2}}\right)\right] \tag{3}$$

in which $\Pr(B \succ A \,|\, \Delta L_{AE}, \beta_0, \beta_1)$ is the probability that sound B is more annoying than sound A, given ΔL_{AE}, β_0 and β_1. The relative difference in L_{AE} of sound B relative to sound A is ΔL_{AE}, and erf is the error function. The regression coefficients estimated from the probit model are β_0 and β_1. Once an appropriate model is fit, the Equal Annoyance Point is defined as the value of ΔL_{AE} such that $\Pr = 0.5$, which occurs at $\Delta L_{AE} = -\beta_0/\beta_1$. Note that the EAP has the same units as L_{AE}.

Simple confidence interval for EAP It is important to evaluate the confidence interval for the EAP. If the responses vary greatly, it may be that a large range of values are possible for the EAP. If this range overlaps with 0 dBA, then the significance of EAP cannot be determined.

There are two methods that are used to determine the confidence interval on the EAP. The first uses the standard error, $\delta(\bullet)$, and estimates of the regression coefficients that are outputs of the probit model. The standard error of the EAP is given by

$$\delta(\text{EAP}) = |\text{EAP}|\sqrt{\left(\frac{\delta(\beta_0)}{\beta_0}\right)^2 + \left(\frac{\delta(\beta_1)}{\beta_1}\right)^2} \tag{4}$$

in which $\delta(\beta_0)$ and $\delta(\beta_1)$ are the standard error of the regression coefficients. The confidence interval on EAP is then $-\beta_0/\beta_1 \pm 1.96\,\delta(\text{EAP})$. This assumes that the errors are normally distributed and that the covariance between β_0 and β_1 is zero. It also assumes that the confidence interval is symmetric about the EAP. While this simple confidence interval is easy to calculate, it may be overly simplistic and the assumptions may not be valid in all cases. Because of this, a more advanced confidence interval is also used, which is described next.

Advanced confidence interval for EAP In order to get a more accurate estimate of the confidence interval, a more advanced approach is used, one that iteratively varies the parameters of the likelihood function and finds a distribution of the most likely values that satisfy the binary response data.

Instead of writing the probability in terms of the probit model regression coefficients, as in Eq. (3), it can also be written in terms of a normal cumulative distribution function, Φ, with mean, μ, and standard deviation, σ. This is given by

$$\Phi(\Delta L_{AE}, \mu, \sigma) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{\Delta L_{AE} - \mu}{\sigma\sqrt{2}}\right)\right] \quad . \tag{5}$$

This form of Φ[3] is used as input to a Markov Chain Monte Carlo (MCMC) simulation in which both μ and σ are varied according to a random walk using the Metropolis-Hastings algorithm. At each step, the new value of μ is simply the current estimate of the EAP. The change is accepted if the likelihood function is greater at the new step; otherwise, it is accepted with a probability of the ratio between the old and new values of the likelihood function. The likelihood function is the product of the likelihood of all responses for the given value of μ and σ, and the likelihood of a response is given by Φ in Eq. (5) for a correct response and $1-\Phi$ for an incorrect response. An initial "burn-in" phase of 1000 steps was used. After that, 100,000 steps of the algorithm produced a random walk around the most likely combination of μ and σ (i.e., the one in which Φ most closely matches the response data). The

[3] In Eq. (5), it should not be interpreted that the probability of sound B being more annoying than sound A is equal to the probability of the random variable ΔL_{AE} being less than or equal to some value. Equation (5) is not a cumulative distribution. It is a function of three variables that happens to take the same form as a cumulative distribution function.

resulting distribution of μ then gives a 95% confidence interval around the EAP, given by the quantiles bounded by 2.5% and 97.5%. The acceptance rate is the total steps accepted divided by the total steps and should be close to 0.5 for binary data (Ref. 19). Ten successive MCMC simulations for the pair 3 full flyovers resulted in a standard deviation of 0.15 dBA when calculating the width of the confidence interval. For a detailed description of this approach, see (Ref. 20).

Three-point perceptually-adjusted L_{AE} (TPPAS)

The previous discussion of the Equal Annoyance Point involved finding an appropriate fit of a probit model using full flyovers as sound stimuli. Another approach is to modify the L_A time history of a flyover by adjusting L_A at times t_1, t_{max} and t_2 based on subjects' annoyance responses. Basically, the EAP is found for short sounds (approximately 1 s) centered about these three times, adjusting the L_A value at these times and interpolating the rest of the L_A time history adjustments in between the 10 dBA down points. Recalculating L_{AE} based on this adjusted L_A time history gives the three-point perceptually-adjusted L_{AE} (TPPAS).

The TPPAS approach was pioneered during the NASA Environmentally Responsible Aviation project in which the efficacy of EPNL to accurately reflect human responses to noise across a wide range of vehicle designs was investigated (Ref. 20). In that study, the auralizations were longer than 45 s, so it was thought impractical to ask subjects to compare such long noise stimuli. Therefore, the motivation for using this approach is that the time needed for subjects to compare full flyovers can be made much shorter, enabling more flyover comparisons without fatiguing test subjects. If two full flyovers were 20 s, one comparison would require 40 s of listening time. If three short sounds of 1 s were compared instead, subjects would need only 6 s, meaning that subjects could make 6 comparisons in the same time needed to make comparisons of two full flyovers. Another advantage is that for a longer flyover, the short comparisons would still only require 6 s.

Finding TPPAS for a given pair of full flyovers starts with comparing the flyovers when L_A is at its maximum (t_{max}) and when it is 10 dBA lower than the maximum (t_1 and t_2). Let t_{1A} be the first time during the flyover of sound A that L_A is within 10 dBA of its maximum, and let t_{1B} be the first time during the flyover of sound B that L_A is within 10 dBA of its maximum. Binary responses corresponding to subjects' annoyance of sound A at t_{1A} compared to sound B at t_{1B} are collected. Then an MCMC simulation gives a distribution of 100,000 samples of the most likely adjustment, ΔL_A, needed to make the two sounds at t_1 equally annoying. A random resampled (i.e., bootstrapped) dataset is then generated from the MCMC result in order to make later statistical estimates of the MCMC result. This is repeated for t_{max} and t_2.

The L_A time history is adjusted using the bootstrap datasets at the three times of the flyover. For $t \leq t_1$, ΔL_A for t_1 is added to the original L_A. For $t_1 < t \leq t_{max}$, an interpolation between ΔL_A at t_1 and t_{max} is added to the original L_A. For $t_{max} < t \leq t_2$, an interpolation between ΔL_A at t_{max} and t_2 is added to the original L_A. For $t > t_2$, ΔL_A at t_2 is added to the original L_A. After adjusting the entire L_A time history, a new L_{AE} is calculated and represents a possible value for the EAP. This is done for all 100,000 resamples, resulting in a distribution of likely EAPs. It is then straightforward to get the 95% CI from this distribution by excluding the most extreme 2.5% of the values on either side.

RESULTS

Sounds of equal L_{AE} (RQs 1a, 2, 3a and 4a)

Table 3 shows the results of two-sided binomial tests for four pairs of sounds. All four pairs were presented to the subjects at the same L_{AE}. The number of responses, N, was 32 for each pair. The percentage (%) is the number of responses in which sound B for that pair was judged more annoying than sound A, divided by the total number of responses for that pair. The p-value, p, is also listed, in which $p < 0.05$ is considered statistically significant and that the null hypothesis (sounds A and B are equally annoying) should be rejected.

For pair 1, the null hypothesis cannot be rejected, meaning that when played at the same L_{AE}, it cannot be ruled out that the L_{AE}-optimized and baseline rotors are equally annoying. This means that if the sounds were not normalized in terms of L_{AE} and, instead, were played at their original absolute levels, the perceptual difference in annoyance would likely be comparable to the designed reduction in L_{AE}.

The null hypothesis also cannot be rejected for pair 2, meaning that when the rotor is optimized in terms of L_{AE} or EPNL, one is not significantly more annoying than the other. It means that either metric is suitable as a design criteria for this situation and that there are no apparent changes in sound quality due to this choice that would have a significant impact on annoyance for auralizations of periodic rotor sounds. This result may differ, however, for a more complicated sound, e.g., one that also includes a broadband component. It could be that minimizing one metric would reduce tonal noise while optimization based on another metric would focus more on reducing broadband noise. This comparison was outside the scope of the current psychoacoustic test. Nevertheless, for auralizations with similar noise components, design optimizations based on either L_{AE} or EPNL were not perceived to be significantly different in terms of annoyance.

In contrast to pairs 1 and 2, the null hypothesis should be rejected for pair 3. Since both helicopters were at the same L_{AE} and performed the same maneuver, it is likely that there is a difference in sound quality that makes the EC130 more annoying than the AS350 and that L_{AE} did not capture this difference in perception. It is unlikely that the difference in perception came from the L_A time history, because both profiles, including $L_{A,max}$, are quite similar (see Figure 4). Since the two helicopters come from the same manufacturer and are similar in engine size and load capacity, perhaps this is a result of the Fenestron on the EC130. It indicates that different tail rotor

Table 3: Results of binomial tests of four pairs of flyovers. LNA: low noise approach.

Pair	Sound	Description	%	p	N
1	A B	Baseline rotor L_{AE}-optimized rotor	56	0.60	32
2	A B	L_{AE}-optimized rotor EPNL optimized rotor	56	0.60	32
3	A B	AS350 LNA EC130 LNA	72	0.02	32
4	A B	AS350 steady AS350 unsteady	56	0.60	32

technologies may have a significant impact on sound quality and perception. In a previous test, sound quality metrics such as sharpness, tonality and fluctuation strength were found to be indicators of annoyance to helicopter sounds that were normalized in terms of loudness (Refs. 5,21). Some of the design features of the Fenestron, such as the increase in blade count or the uneven spacing of the fan blades, may create a negative shift with respect to these sound quality metrics or other qualitative aspects of the sound. This result shows that L_{AE} may fail to capture significant qualitative aspects of annoyance responses when comparing two different helicopters.

For pair 4, the null hypothesis cannot be rejected. When only comparing the two sounds at the same L_{AE}, there was close to an even probability of responses. From this simple comparison, it cannot be ruled out that the steady and unsteady maneuver are equally annoying.

Equal Annoyance Point (RQs 1b, 3b and 4b)

The binomial test results in Table 3 only consider the comparisons of sounds that were presented at the same L_{AE}. While the conclusions are indicative of the efficacy of L_{AE} in annoyance ratings of helicopter noise, more insight is gained by comparing the sounds for various relative L_{AE} values. In particular, probit fits lead to an estimate of the Equal Annoyance Point and its confidence interval.

Figure 5 summarizes the subject responses for pairs 1, 3 and 4 for various relative gains of sound B relative to sound A and shows the probability that sound B is more annoying than sound A. (Pair 2 is not shown because the paired comparisons of the L_{AE}- and EPNL-optimized rotors were not played for subjects at different relative L_{AE}. The probit fits are also shown.) The intersection of each probit fit with $\mathrm{Pr} = 0.5$ gives the EAP for each pair. Since the EAPs happen to be negative, it suggests that the L_{AE} of sound B should be a few dBA less than that of sound A (for each pair) in order for the two sounds to be equally annoying. However, to fully understand the significance of EAP, its confidence interval must also be considered.

The EAP for pairs 1, 3 and 4, found through probit models are shown in Table 4. The confidence intervals using Eq. (4) and MCMC simulations are also shown. Only for the baseline vs. L_{AE}-optimized rotor pair does the confidence interval contain 0, indicating that the null hypothesis cannot be rejected for pair 1. The null hypothesis should be rejected for the other two pairs. From these results on the confidence intervals, it is concluded that, when played at the same L_{AE}:

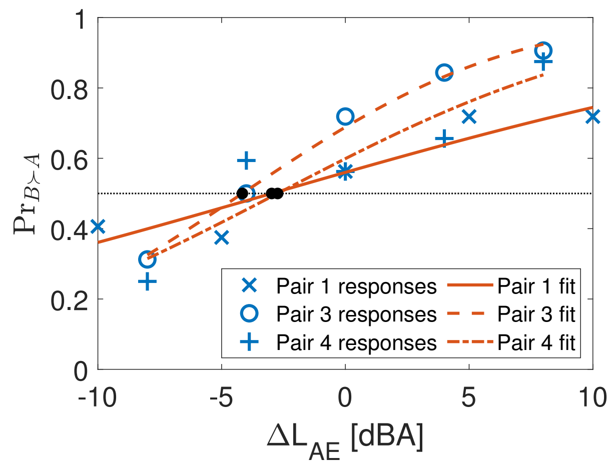


Figure 5: Probability that sound B is more annoying than sound A for pairs 1, 3 and 4. Intersections of the probit fits with the horizontal line show the Equal Annoyance Point for each pair (solid circles).

1. the L_{AE}-optimized rotor design is not perceived to be significantly different in terms of annoyance than the baseline rotor
2. the EC130 low noise approach is perceived to be significantly more annoying than the AS350 low noise approach and
3. the unsteady AS350 maneuver is perceived to be significantly more annoying than the steady AS350 maneuver.

The results of the probit models for pairs 1 and 3 in Table 4 agree with the binomial results in Table 3, confirming the conclusions that the null hypothesis cannot be rejected for pair 1 and should be rejected for pair 3. On the other hand, the results of the probit model for pair 4 differs from the binomial result. This can be understood by inspecting the responses shown in Figure 5. While the probability is 0.56 at 0 dBA, it is slightly higher (0.59) at -4 dBA. This shifts the curve fit and raises the modeled probability to 0.60 at 0 dBA, which is enough to keep it just outside the EAP confidence interval. In this way, the curve fit incorporates data taken at different ΔL_{AE} and takes into account more information than at a single relative level. Using this added information, it is concluded that the unsteady flight is perceived to be slightly more annoying than the steady flight.

To be clear, the assumptions using Eq. (4) are that the standard error of both probit regression parameters have no covariance and that the standard error on EAP is normally distributed. To check this, the more advanced estimation of the confidence interval using MCMC simulations is also done. The results are also presented in Table 4 and are discussed next.

Table 4: Equal Annoyance Point (EAP) found from probit fits to binary response data for full flyovers. The confidence intervals (CI) on EAP using Eq. (4) and Monte Carlo simulations are also shown.

Pair	EAP	CI	
		Eq. (4)	MCMC
1	-2.98	[-7.26, 1.31]	[-5.33, 0.57]
3	-4.16	[-6.55, -1.78]	[-7.93, -2.39]
4	-2.74	[-5.26, -0.21]	[-5.16, -0.56]

Further inspection of Figure 5 shows that the binary response data for pairs 1 and 4 are not monotonic while the data for pair 3 are. If all the data were as expected, higher L_{AE} for sound B would always lead to a higher probability that sound B was more annoying than sound A. However, random error (most pronounced near $\mathrm{Pr} = 0.5$) in the collection of responses can lead to nonmonotonicity. For this response data, there are many combinations of μ and σ that are just as likely, which can make the MCMC simulations unstable.

One of the problems is that for a large standard deviation, the slope of Φ around the EAP is very shallow, resulting in a large range of possible EAP values. For pairs 1 and 4, this leads to large acceptance rates (i.e., near 1) and unrealistically large predictions of the confidence interval on EAP. On the other hand, the smooth data collected for pair 3 led to a good acceptance rate (i.e., near 0.5) and a good estimate of the confidence interval on EAP.

Therefore, the results of the MCMC simulations here start with pair 3, assuming no prior information about the mean and standard deviation of Φ. Then, a log-normal distribution is fit to the resulting distribution of σ. This fit is applied to the response data for pairs 1 and 4 as a prior (in the Bayesian inference sense, see (Ref. 18), p. 392), restricting the likely values of σ. This restriction leads to narrower, more reasonable estimations of the confidence interval on EAP than if no prior would have been used[4].

The results of the probit model fit, as well as the MCMC simulation, are shown in Figure 6 for pair 3, the comparison between the EC130 and AS350 low noise approaches. The probit fit gives an estimate of the EAP as -4.16 dBA. The estimate of the confidence interval on EAP is found through the MCMC simulation. While the probit model gives the best fit of Φ that matches the binary response data, the MCMC simulation gives a large number of likely combinations of the mean, μ, and standard deviation, σ, of Φ that match the data (see Figure 6b). The resamples tend to accumulate around the EAP, which is shown more clearly in the histogram of resamples for μ in Figure 6d. Finally, the 95% CI for EAP is given by the quantiles of the resamples on μ, bounded by 2.5% and 97.5%, which is shown in Figure 6d; it is also plotted on the probit fit in Figure 6c and shown in Table 4. The CI is not symmetric about the EAP and is 0.77 dBA wider than that predicted by Eq. (4). Since the CI does not contain 0, this result agrees both with the binomial test for $\Delta L_{AE} = 0$ and for the simple CI given by Eq. (4), which is that the EC130 is perceived to be more annoying than the AS350 when performing the same low noise approach and when presented at the same L_{AE}.

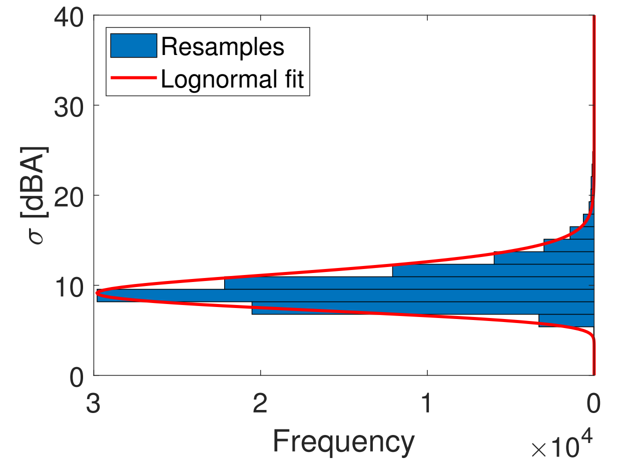
Although this comparison was made to characterize the efficacy of L_{AE} in rating annoyance of two sounds at the same L_{AE}, another practical question could be whether the two recordings differ in annoyance at their absolute levels as recorded. The absolute levels of the low noise approaches that were flown by the AS350 and EC130 were 89.15 and 89.51 dBA, respectively, a relative difference of 0.36 dBA. Since this ΔL_{AE} is well outside the ranges for the EAP confidence intervals found in Table 4 for pair 3, it is concluded that the EC130 is still more annoying than the AS350 when both fly a low noise approach under the same conditions, even at their absolute levels.

The other information found from the MCMC simulation on pair 3 is the distribution of σ. The histograms of the likely values of σ are shown in Figure 6a, which closely follows a lognormal distribution. This type of distribution for σ is expected, because (1) the standard deviation cannot be negative and (2) the likelihood should decrease for large values of σ. These expectations were not met when analyzing the responses to pairs 1 and 4. Therefore, the best fit to a lognormal distribution using responses for pair 3 is used as a prior to calculate the likelihood at each iteration in the MCMC simulation for pairs 1 and 4, which leads to accurate predictions of the CI for EAP.

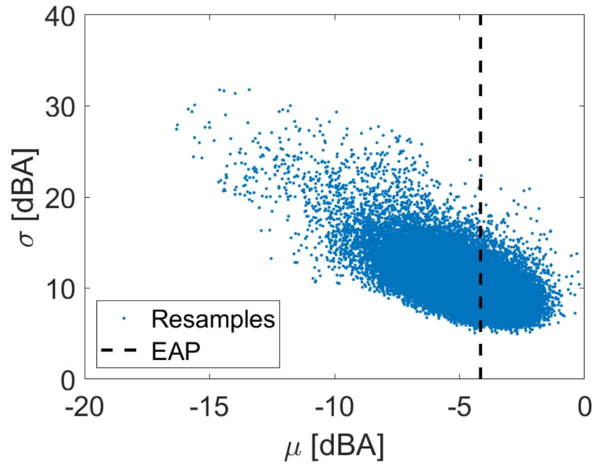
The confidence interval for EAP for pair 1 is shown in Figure 7 and Table 4. The results from the MCMC simulations agree with the simple estimation from Eq. (4); the confidence interval for pair 1 contains 0. Even though the CI using MCMC for pair 1 is 2.67 dBA narrower, it still contains 0 because of its asymmetry. The results from the MCMC simulations, therefore, support the conclusions found using the simple estimate of the CI. Specifically, the L_{AE}-optimized rotor is not significantly more annoying than the baseline rotor when presented at the same L_{AE}.

Nevertheless, the estimate of the EAP for the L_{AE}-optimized rotor compared to the baseline rotor is -2.98 dBA, indicating that perception gravitates toward being less annoying for the baseline rotor than for the L_{AE}-optimized rotor. Although not conclusive, one possible cause for this is that there is an aspect of the sound in the optimized rotor that is more annoying than the baseline rotor. This could be due to the fact that only the main rotor was optimized (the tail rotor geometry was left unchanged). Reducing the L_{AE} of the main rotor while leaving the tail rotor mostly constant, then normalizing in terms of L_{AE} means that the tail rotor noise is slightly elevated compared to the main rotor in the L_{AE}-optimized auralization. Since the tail rotor noise is a harmonic tone complex that starts at higher frequencies and hence has larger spacing between tones, this could cause a slightly higher annoy-

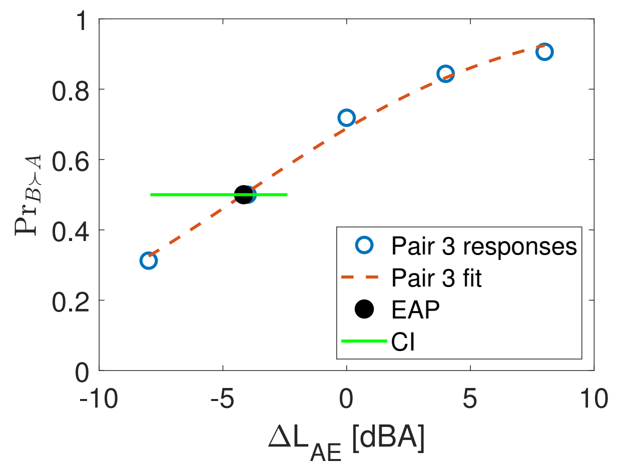
[4] Applying a log-normal distribution as a prior for the MCMC simulation results in a posterior distribution on μ, and hence a Bayesian high-density interval around the EAP (Ref. 22). However, the term *confidence interval* is still used here due to its conceptual simplicity.



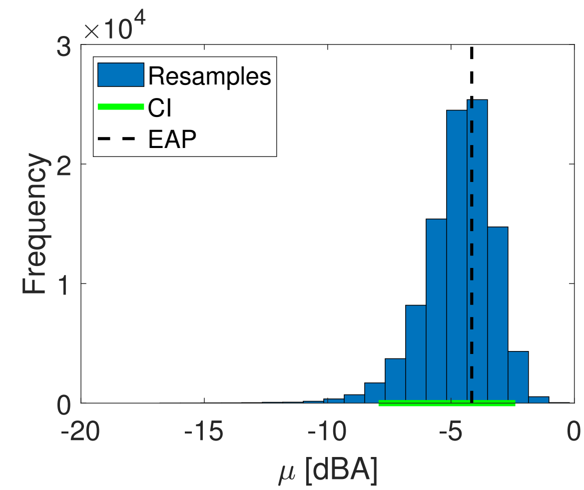
(a) Histogram of MCMC resamples of σ, the standard deviation of Φ, and lognormal fit. Ordinate is congruent with that in (b).



(b) Resamples from MCMC simulations and EAP from probit fit.



(c) Probability that the EC130 is more annoying than the AS350 for a low noise approach.



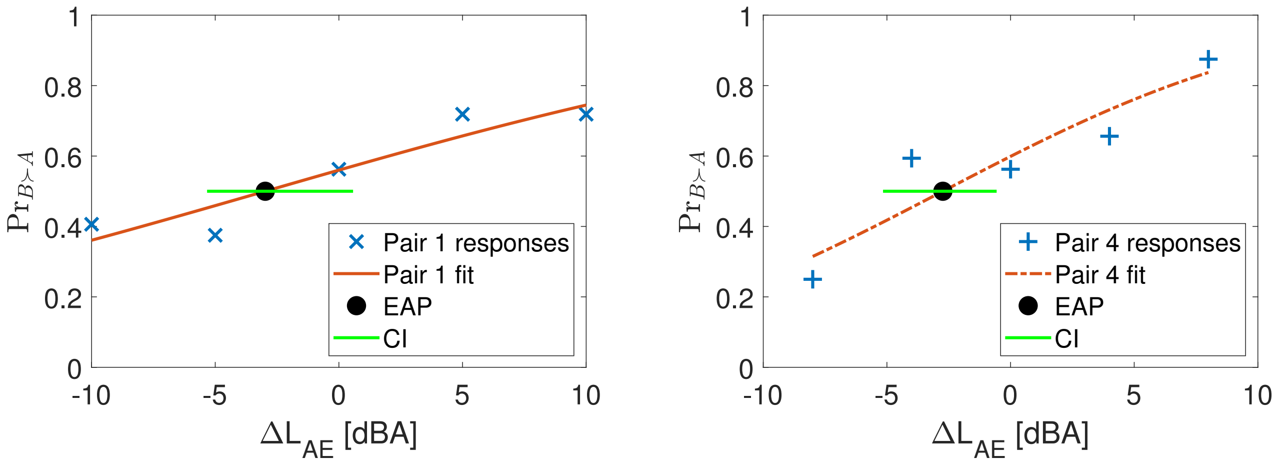
(d) Histogram of MCMC resamples of μ, the mean of Φ, and 95% confidence interval. Abscissa is congruent with that in (b).

Figure 6: An MCMC simulation gives the most likely values of the mean, μ, of Φ that fits the response data for pair 3. The histogram (i.e., marginalization) gives an estimate of the confidence interval on EAP. A lognormal fit on the marginalization of results on σ is used as a prior for MCMC simulations for other pairs.

ance response even though the main rotor noise was reduced. The results suggest that annoyance might be further reduced if both main and tail rotor geometries are optimized simultaneously. Another possible cause for the annoyance rating gravitating slightly higher for the optimized rotor is that the duration of the 10 dBA-down time interval is 50% longer than that of the baseline rotor (see Figure 2b). Through written responses, several subjects reported that sounds that appeared to loiter (i.e., longer sounds) were found to be more annoying, in general. There is no penalty in Eq. (1) for an increase in the quantity $t_2 - t_1$. It means that for some subjects, this was another qualitative aspect that L_{AE} failed to capture.

The difference in the absolute (i.e., before normalization) L_{AE} of the L_{AE}-optimized rotor relative to the baseline rotor auralization was -4.22 dBA, which is contained within the 95% confidence interval of the EAP. It means that, from a statistical point of view, it cannot be ruled out that the optimized and baseline rotors are equally annoying when played at their absolute levels. An overly simplistic interpretation of these results would be that the optimized rotor should be designed with a reduction that falls outside the 95% confidence interval of the EAP. A better, more nuanced interpretation is that the experimental variance was higher in this test than in a previous psychoacoustic test of similar design (Ref. 20), which led to a power of the test that was lower than desired and confidence intervals that were wider than expected. Although there was a preference for the baseline rotor, the designed reduction in L_{AE} was greater than this preference.

The confidence interval for EAP for pair 4 is also shown in Figure 7 and Table 4. The results from the MCMC simulations agree with the simple estimation from Eq. (4), i.e., the confidence interval for pair 4 does not contain 0. Therefore, the



(a) Probability that the L_{AE}-optimized rotor is more annoying than the baseline rotor.

(b) Probability that the unsteady AS350 flight is more annoying than the steady AS350 flight.

Figure 7: Probit models and EAPs with CIs.

results from the MCMC simulations support the conclusion found using the simple estimate of the CI, which is that the unsteady AS350 flight is perceived to be slightly more annoying than the steady one when presented at the same L_{AE}. This suggests that L_{AE} may not be sufficient to quantify the perceived difference between two maneuvers flown by the same vehicle and that the induced changes in sound quality are an important factor in the rating of annoyance to helicopter noise.

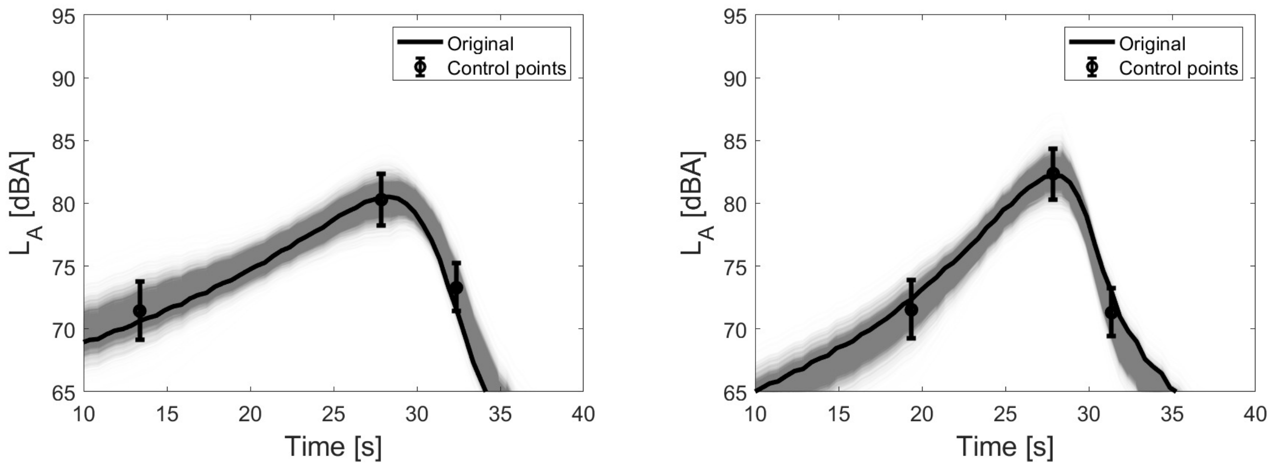
A possible nonacoustic cause for the increase in annoyance to the unsteady sound could be that some subjects reported that unsteadiness in the sound was interpreted as an unsteadiness in the flight or control of the vehicle, which signaled an elevated safety risk. This supports other studies in which fear was determined to be a moderating factor on an individual's annoyance to noise (Ref. 23).

Three-point perceptually adjusted L_{AE} (RQ 5)

Calculating the Equal Annoyance Point and its confidence interval by comparing full flyovers is one way to evaluate the efficacy of L_{AE} as an appropriate metric in the rating of annoyance to helicopter noise. The three-point perceptually-adjusted L_{AE} (TPPAS) approach is another way to evaluate it. As explained in Section ANALYSIS TECHNIQUES, TPPAS generates perceptually-adjusted L_A time histories based on three control points at t_1, t_{max} and t_2. The response data are collected in a similar way to the full flyover, except that the stimuli are short parts of the flyover centered around the three control points. Corresponding control points from sounds A and B were compared at different relative L_A and analyzed with MCMC simulations, which gave a range of possible adjustments to the L_A time history that make the short sound of A equally annoying to B. One adjustment for each control point, plus interpolation, produced a perceptually-adjusted L_A time history. The perceptually adjusted L_{AE} was then calculated from this adjusted L_A time history using Eq. (1). Doing this for all 100,000 outputs of the MCMC simulations gave a distribution of likely adjustments to L_{AE} to make sounds A and B equally annoying.

Figure 8 shows the perceptually adjusted L_A time histories for the L_{AE}-optimized vs. baseline rotor case (pair 1). The three control points at t_1, t_{max} and t_2 were found from MCMC simulations on the binary response data comparing short sounds centered at these control points. The average adjustment at each control point is represented by circles within each error bar. The means and 95% confidence intervals are given in Table 5. The error bars in Figure 8a show the adjustments necessary to make the short sounds of the L_{AE}-optimized rotor equally annoying to the baseline rotor short sounds in 95% of the cases. By reciprocity, the adjustments needed to make the baseline rotor sounds equally annoying to the L_{AE}-optimized rotor are opposite, which are shown in Figure 8b, and the adjustments in L_A are now relative to the baseline rotor L_A time history. The 95% confidence intervals at all three control points contain 0, meaning that these short sounds were not significantly more or less annoying than when sounds A and B were played at the same L_A. As a result, the adjusted L_A time histories significantly overlap with the originals. As will be discussed later in Figure 10a, the results of TPPAS and those from the full flyovers both show that the L_{AE}-optimized rotor is not significantly more or less annoying than the baseline rotor when played at the same L_{AE}.

Figure 9 shows the perceptually adjusted L_A time histories for the unsteady vs. steady AS350 recorded flights (pair 4) using the TPPAS approach. The mean L_A adjustment and confidence interval for the control points comparisons are shown in Table 5. Here, the stimuli from the 10 dBA down points are not perceived to be very different in terms of annoyance, but the comparison at t_{max} is. The short sound centered at t_{max} for the unsteady flight was perceived to be less annoying than the short sound centered at t_{max} of the steady flight. Since



(a) Adjustment of the L_{AE}-optimized rotor L_A time history: pair 1, TPPAS. **(b) Adjustment of the baseline rotor L_A time history: pair 1, TPPAS.**

Figure 8: Perceptually-adjusted L_A time histories for pair 1[1]. "Original" indicates the L_A time history used in the full flyover comparison. Control points are at the maximum (t_{max}) and 10 dBA down points (t_1 and t_2). Thin curves are the adjusted L_A time histories, which are interpolated from the distributions found at the control points.

the sounds at t_1 and t_2 were not perceived to be significantly different in terms of annoyance, the difference at t_{max} results in interpolating almost the entire L_A time history of the unsteady flight to be higher. Reciprocally, most of the L_A time history of the steady flight is adjusted lower in order to match the responses of the unsteady flight.

Each light trace in Figure 8 gives a possible value of ΔL_{AE} to make the L_{AE}-optimized rotor equally annoying as the baseline rotor. All the light traces give a distribution of possible values for ΔL_{AE}. The distributions found from TPPAS and those found from the full flyovers are shown in Figure 10a. All three confidence intervals overlap with 0, meaning that the L_{AE}-optimized rotor is not significantly more or less annoying than the baseline rotor when played at the same L_{AE}. The distributions found by adjusting the baseline sound and L_{AE}-optimized sound are almost identical, meaning that this reciprocal relationship holds for this pair of sounds.

Although all three confidence intervals in Figure 10a overlap with 0, there are some important differences. Firstly, the distribution is wider for the full flyover comparison than with TPPAS, indicating a higher variation in responses for full flyovers. In contrast, the confidence intervals are narrower for each control point comparison as well as for the overall TPPAS distribution. It means that there is less variation in responses when subjects compare shorter sounds than when they compare full flyovers, suggesting focusing on simpler tasks produces a more well-controlled response from subjects. On the other hand, responding to an entire flyover, subjects focus on different parts of the flyover and give more varied responses. A second difference is that the distribution of the full flyover tends negative while the TPPAS results are centered closer to 0, indicating slightly different EAPs between the two methods.

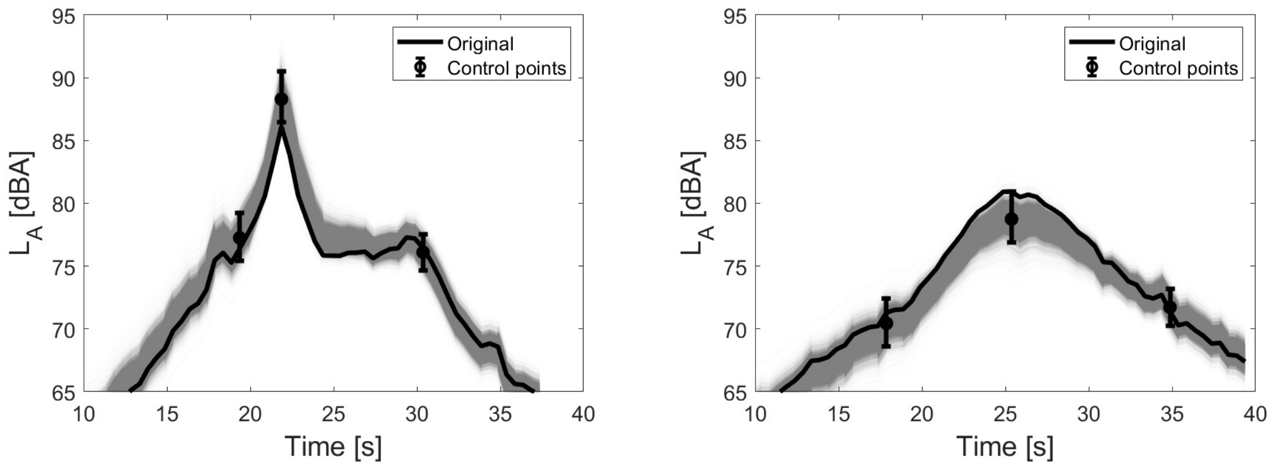
Table 5: Mean adjustment (with confidence intervals) of L_A for the TPPAS comparisons at three control points, found from Monte Carlo simulations. Adjustments are to make sound B equally annoying to sound A for each pair (see Table 3).

	Pair	
Control point	1	4
t_1	0.81 [-1.48, 3.18]	0.84 [-0.96, 2.74]
t_{max}	-0.20 [-2.26, 1.84]	2.17 [0.34, 4.24]
t_2	1.68 [-0.21, 3.63]	-0.37 [-1.87, 1.10]

Similar to pair 1, the distributions for pair 4 using the TPPAS approach are narrower than the distribution on the EAP for the full flyover comparison, as shown in Figure 10b. Again, this indicates that it is easier and more accurate for subjects to give responses to shorter sounds, because the focused task is simpler.

The distributions for pair 4 do not give the same conclusions on which sound is more annoying. The distribution with the full flyovers is negative, and the confidence interval does not overlap 0 (see Table 4), which mean that the unsteady flight is significantly more annoying than the steady flight. The confidence interval for adjusting the unsteady AS350 sound with the TPPAS approach does overlap 0, which would mean that neither flight is significantly more or less annoying than the other. The distribution for adjusting the steady AS350 sound with the TPPAS approach does not overlap 0 and is positive, which would mean that the steady flight is more annoying than the unsteady flight.

These three different outcomes is an interesting result that gives further insight into the efficacy of L_{AE}. In particular, it is evident that neither Eq. (4) nor the TPPAS approach takes



(a) Adjustment of the unsteady AS350 flight: pair 4, TPPAS. **(b) Adjustment of the steady AS350 flight: pair 4, TPPAS.**

Figure 9: Perceptually-adjusted L_A time histories for pair 4[1]. "Original" indicates the L_A time history used in the full flyover comparison. Control points are at the maximum (t_{max}) and 10 dBA down points (t_1 and t_2). Thin curves are the adjusted L_A time histories, which are interpolated from the distributions found at the control points.

into account the fact that t_{max}-t_1 is only around 2.5 s for the unsteady flight, an example of the onset rate being a factor of annoyance (Ref. 24). Evidence of this is in the different responses that are found when subjects are presented with the entire flyover. The fast rise in L_A suggests that comparing sounds at t_1 and t_{max} separately, as with TPPAS, does not give the subjects any indication that those sounds are spaced so close together in time. Likewise, a 10 dBA difference in 2.5 s does not appear in the calculation for L_{AE}.

CONCLUSIONS

The main results of a psychoacoustic test related to the efficacy of L_{AE} in the rating of annoyance to helicopter noise are presented. Auralizations based on optimized rotor designs for an AS350 as well as recorded flights of an AS350 and EC130 were judged by human subjects in the Exterior Effects Room at NASA Langley Research Center.

It was found that optimizing rotor designs in terms of L_{AE} or EPNL does not give significant differences in the perception of annoyance. To within the resolution of the test, a rotor design optimized in terms of L_{AE} was not judged to be significantly different than the baseline rotor when played at the same L_{AE}, which means that L_{AE} can be an effective metric to minimize in the design phase of low noise rotors. However, an EAP of around −3 dBA suggests that both main and tail rotors should be optimized to not introduce unwanted changes in sound quality. The optimized rotor resulted in an auralized flyover within 10 dBA of its peak that was around 50% longer than the baseline rotor auralization, which could have contributed to the slightly higher annoyance responses for the optimized rotor and exhibits a deficiency in using L_{AE} as an indicator of annoyance to helicopter noise.

In addition to comparing auralized sounds based on different rotor designs, the psychoacoustic test also used recordings of helicopter maneuvers in order to evaluate L_{AE} as an indicator of annoyance. When a low noise approach was played at the same L_{AE}, the EC130 was judged to be more annoying than the AS350. This suggests that L_{AE} may not be a good indicator of annoyance when comparing different helicopters and that other temporal, spectral or spatial components not contained in the L_{AE} calculation are important. In particular, different tail rotor technologies may be an important consideration. Sound Exposure Level also did not fully capture annoyance responses when comparing different maneuvers for the same helicopter; for equal L_{AE}, a flight with variations in sound quality and a steep rise in L_A was found to be more annoying than a flight with more constant sound quality.

Author contact: Matthew Boucher matthew.a.boucher@nasa.gov

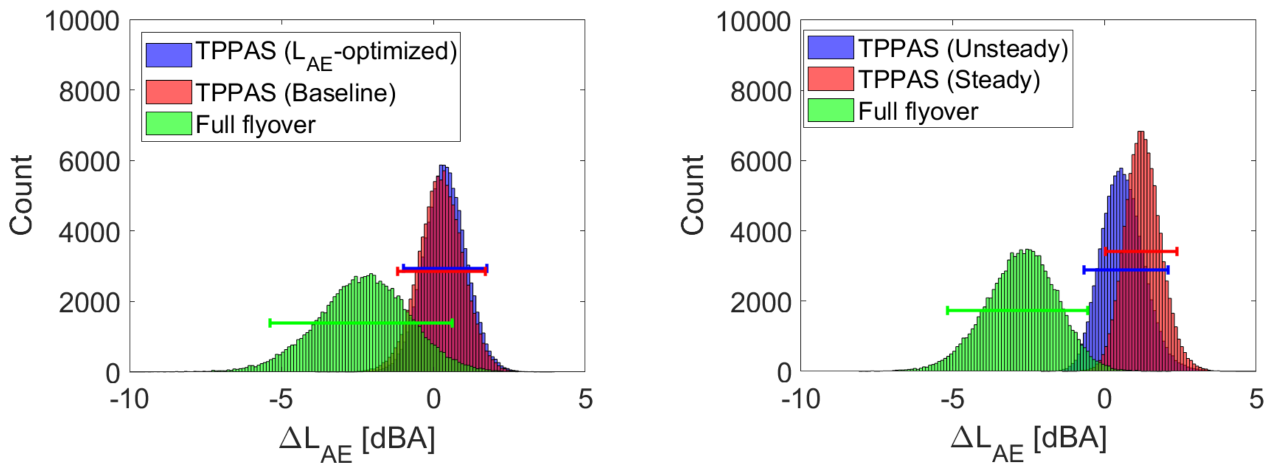
ACKNOWLEDGMENTS

The authors are grateful for the support from the Revolutionary Vertical Lift Technology project of the NASA Advanced Air Vehicles Program, Erin Thomas for the recruiting of test subjects and performing audiograms, several colleagues for participating in pilot studies, as well as to Menachem Rafaelof, Brian Tuttle and Matthew Hayes for assistance in executing the test.

REFERENCES

1. V. Mestre, S. Fidell, R.D. Horonjeff, Schomer P, A. Hastings, B.G. Tabachnick, and F.A. Schmitz. *Assessing Community Annoyance of Helicopter Noise*. Transportation Research Board, nov 2017.

2. International Civil Aviation Organization. Annex 16 to the convention on international civil aviation environ-



(a) Comparison of EAP for the full flyover and TPPAS for the L_{AE}-optimized rotor case (pair 1).

(b) Comparison of EAP for the full flyover and TPPAS for the unsteady/steady AS350 flights (pair 4).

Figure 10: Distributions of bootstrapped L_{AE} values resulting from the TPPAS method, compared to the distribution of EAP for full flyovers.

mental protection volume i: Aircraft noise. Technical report, ICAO, 2014.

3. T. Gjestland. Assessment of helicopter noise annoyance: a comparison between noise from helicopters and from jet aircraft. *Journal of Sound and Vibration*, 171(4):453–458, 1994.

4. A. Taghipour, R. Pieren, and B. Schäffer. Short-term annoyance reactions to civil helicopter and propeller-driven aircraft noise: A laboratory experiment. *The Journal of the Acoustical Society of America*, 145(2):956–967, 2019.

5. S. Krishnamurthy, A. Christian, and S.A. Rizzi. Psychoacoustic test to determine sound quality metric indicators of rotorcraft noise annoyance. In *Inter-Noise: Impact of Noise Control Engineering*, Chicago, 2018.

6. E. Greenwood. Helicopter flight procedures for community noise reduction. In *American Helicopter Society 73rd Annual Forum*, Fort Worth, TX, 2017.

7. M. Vorländer. *Auralization: Fundamentals of Acoustics, Modeling, Simulation, Algorithms and Acoustic Virtual Reality*. Springer-Verlag GmbH, 2007.

8. S. Krishnamurthy, S.A. Rizzi, D.D. Boyd, and A.R. Aumann. Auralization of rotorcraft periodic flyover noise from design predictions. In *American Helicopter Society 74th Annual Forum*, Phoenix, AZ, 2018.

9. E. Zwicker and H. Fastl. *Psychoacoustics: Facts and Models*. Springer, 2013.

10. M. Watts, E. Greenwood, C.D. Smith, and J.H. Stephenson. Noise abatement flight test data report. Technical Report TM-2019-220264, NASA, March 2019.

11. K. Faller, S.A. Rizzi, and A.R. Aumann. Acoustic performance of a real-time three-dimensional sound-reproduction system. Technical Report TM-2013-218004, NASA, June 2013.

12. F. Farassat and G.P. Succi. The prediction of helicopter rotor discrete frequency noise. *Vertica*, 7:309–320, 1983.

13. S. Krishnamurthy, B.C. Tuttle, and S.A. Rizzi. Auralization of unsteady rotor noise using a solution to the ffowcs williams-hawkings equation. In *Vertical Flight Society 75th Annual Forum*, Philadelphia, 2019.

14. S. Krishnamurthy, B.C. Tuttle, and S.A. Rizzi. A synthesis plug-in for steady and unsteady loading and thickness noise auralization. In *AIAA Aviation*, Reno, NV, 2020.

15. A.R. Aumann, B.C. Tuttle, W.L. Chapin, and S.A. Rizzi. The nasa auralization framework and plugin architecture. In *InterNoise 2015*, San Francisco, CA, 2015.

16. H.A. David. *The Method of Paired Comparisons*. Charles Griffin & Company, Ltd., 2nd edition, 1988.

17. G.E.P. Box, W.G. Hunter, and J.S. Hunter. *Statistics for experimenters*. John Wiley & Sons, New York, 1978.

18. A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007. p. 118.

19. G.O. Roberts, A. Gelman, and W.R. Gilks. Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of Applied Probability*, 7(1):110–120, 1997.

20. S.A. Rizzi and A. Christian. A psychoacoustic evaluation of noise signatures from advanced civil transport aircraft. In *22nd AIAA/CEAS Aeroacoustics Conference*, Lyon, France, 2016. American Institute of Aeronautics and Astronautics.

21. M. Boucher, S. Krishnamurthy, A. Christian, and S.A. Rizzi. Sound quality metric indicators of rotorcraft noise annoyance using multilevel regression analysis. *The Journal of the Acoustical Society of America*, 145(3):1899–1899, 2019.

22. J.K. Kruschke. *Doing Bayesian Data Analysis*. Elsevier LTD, Oxford, 2015.

23. M. Basner, C. Clark, A. Hansell, J.I. Hileman, S. Janssen, K. Shepherd, and V. Sparrow. Aviation noise impacts: State of the science. *Noise Health*, 19(87):41–50, 2017.

24. K.J. Plotkin, K.W. Bradley, J.A. Milino, K.G. Helbing, and D.S. Fischer. The effect of onset rate on aircraft noise annoyance, volume 1, laboratory experiments. Technical Report AL–TR–1992–0093, Wyle Laboratories and Tech-U-Fit Corporation, 1992.