Perceptual evaluation of Sound Exposure Level in annoyance ratings to helicopter noise

Matthew A. Boucher

Research Engineer,

NASA Langley Research Center,

Hampton, VA, USA

Andrew W. Christian

Research Engineer,

NASA Langley Research Center,

Hampton, VA, USA

Siddhartha Krishnamurthy

Research Engineer,

NASA Langley Research Center,

Hampton, VA, USA

Stephen A. Rizzi

Senior Researcher for Aeroacoustics,

NASA Langley Research Center,

Hampton, VA, USA

matthew.a.boucher@nasa.gov, NASA Langley Research Center, Hampton, VA. Presented at the Vertical Flight Society's 76th Annual Forum & Technology Display, Virtual, May 19-21, 2020. Published in the Journal of the American Helicopter Society, DOI: 10.4050/JAHS.69.032006.

Abstract

A psychoacoustic test was performed to assess the effectiveness of Sound Exposure Level (SEL) for indicating changes in annoyance to helicopter noise. SEL was evaluated for flyover auralizations of optimized rotor designs and for flyover recordings of different helicopters and maneuvers. The test used paired comparisons of flyovers within $\mathbf{10\,dB}$ of the maximum A-weighted sound pressure level. For stimuli of equal SEL, annoyance responses showed whether or not SEL is a good indicator of annoyance. While this work does not seek to determine specific attributes contributing to annoyance that are not included in SEL, the magnitude of this offset is of primary interest. Specifically, annoyance responses to relative differences in SEL allowed the calculation of an Equal Annoyance Point. Reductions in SEL lead to reductions in annoyance as expected, but for certain cases, SEL can fail to capture perceptually significant features such as audible differences due to changes in tail rotor design or unsteadiness in the sound of the helicopter.

Nomenclature

a	normalization factor
β_0	regression coefficient representing the intercept parameter
β_1	regression coefficient representing the slope parameter
Δt	sampling interval of the A-weighted sound pressure level
dB	decibels
ΔL_{AE}	the relative difference in SEL between two sounds (SEL of sound B minus SEL of sound A)
$\delta(\bullet)$	standard error of the quantity $\bullet$
erf	the error function
F1A	Farassat formulation 1A
k_1	first sample in which $L_A(k)$ rises above $L_{A,max} - 10\,\mathrm{dB}$
k_2	last sample in which $L_A(k)$ falls below $L_{A,max} - 10\,\mathrm{dB}$
L_{AE}	Sound Exposure Level (measured in decibels)
L_A	A-weighted sound pressure level (measured in decibels)

$L_A(k)$	k-th sample of the A-weighted sound pressure level
$L_{AE,i}$	Sound Exposure Level of an auralization or recorded flyover before equalizing to a target level
L_T	target level for equalization of recordings and auralizations
$L_{A,max}$	maximum A-weighted sound pressure level (measured in decibels)
μ	mean of the Normal distribution
Φ	Normal cumulative distribution function
$\Pr$	probability
σ	standard deviation of Φ
t_1	first instance in time at which the A-weighted SPL rises above $L_{A,max} - 10\,\mathrm{dB}$
t_{max}	time at which maximum A-weighted SPL occurs
t_2	last instance in time at which the A-weighted SPL falls below of $L_{A,max} - 10\,\mathrm{dB}$

Introduction

All helicopters must meet certain noise certification requirements. For light helicopters having a maximum certificated takeoff weight of not more than 7,000 pounds, the certification metric used is the Sound Exposure Level (SEL), which is an integration of the A-weighted sound energy contained in a noise event (Ref. 1 and 2). Despite successful noise certification, complaints related to helicopter noise still persist (Ref. 3). Highly variable operations, low altitude flights over communities and qualitative aspects of the sound are all hypothesized to be factors contributing to higher annoyance to helicopters than to fixed-wing aircraft. The implication is that A-weighted, integrated noise metrics do not capture the complete human annoyance response to helicopter noise.

The test presented in this work seeks to answer whether helicopter noise mitigation strategies (such as optimized rotor blade geometry and specified maneuvers) based on SEL lead to reductions in annoyance. This paper presents the 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 annoyance (1) when rotor blades are optimized for low noise and (2) due to unsteady sounds common to helicopter maneuvers.

To calculate the SEL, the A-weighted sound energy of a noise event is integrated over a given time interval. For use as a noise certification metric for light helicopters, SEL 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, SEL can be approximated by

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

in which L_{AE} is SEL and $L_A(k)$ is the k-th sample of the A-weighted sound pressure level, L_A, and $\Delta t = 0.5\,\mathrm{s}$. For a reference duration of $T_0 = 1\,\mathrm{s}$, SEL represents the total A-weighted energy of the flyover event. For certification purposes, as well as to evaluate the efficacy of SEL in this work, the flyover event is defined as the portion of the A-weighted SPL that is within $10\,\mathrm{dB}$ of its maximum, $L_{A,max}$, such that k_1 is the first sample for which $L_A(k)$ rises above $L_{A,max} - 10\,\mathrm{dB}$ and k_2 is the last sample for which $L_A(k)$ falls below $L_{A,max} - 10\,\mathrm{dB}$.

In this work, the concept of sound quality refers to any aspect of the sound not taken into account by SEL. 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 metrics directly, this work focuses on annoyance differences between sounds that are presented at the same SEL. The main hypothesis is that when subjects compare sounds at the same SEL, the variation in subjective responses reveals the efficacy of Sound Exposure Level in the rating of annoyance to helicopter noise. A large variation in responses for equal SEL indicates there are significant qualitative aspects of helicopter noise not captured by SEL.

Motivation and background

The motivation for this psychoacoustic test comes from previous work that suggests noise certification metrics (e.g., SEL) do not fully describe human annoyance responses to helicopter noise and that noise characteristics, other than just sound level, are important factors. A study of helicopter noise in Norway found that small helicopters were more annoying than a reference fixed wing aircraft with the same A-weighted equivalent continuous sound level (Ref. 4). A study in Switzerland suggested that helicopter

landings were slightly more annoying than helicopter takeoffs of equal SEL (Ref. 5). 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 (slow modulations with maximum sensation around 4 Hz modulation frequency), tonality (effect of sound energy concentrated in a very narrow frequency band) and sharpness (sharper sounds have more higher frequency content) (Refs. 6–8). All these tests suggest that the relationship between noise certification metrics, such as SEL, and annoyance to helicopter noise deserves further study.

Another motivation arises from the need to assess SEL as a perceptually relevant indicator to evaluate noise mitigation strategies that occur in the design phase. Auralization, a technique for creating audible sound files from numerical data (Ref. 9), can be used for this design phase assessment. In room acoustics auralizations, an anechoic sound source is convolved with a room impulse response to create a virtual impression of the sound source in different acoustical environments (Ref. 10). However, when the sound source does not yet exist, as is the case with a new aircraft, or only exists in the form of a spectrum that cannot be listened to, auralization refers to the combined process of source noise synthesis and propagation to an observer location (Ref. 11). Reproduction of a field recording that does not involve any modification of the source noise or propagation effects is not considered an auralization.

Krishnamurthy et al. (Ref. 12) generated auralizations of different AS350 helicopter main rotor geometries, including the original AS350 main rotor and two optimized main rotors: one that minimized SEL on the ground and one that minimized Effective Perceived Noise Level (EPNL). Like SEL, EPNL 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 geometry constant. The auralized sounds of the optimized rotor flyovers at a ground observer had SEL and EPNL values roughly 12–16 dB lower than that of the original main rotor flyover. Further, the loudness (DIN standard 45631/A1) and roughness (modulation analysis from HEAD Acoustics ArtemiS Suite v12.2) sound quality metrics were found by Krishnamurthy et al. (Ref. 12) to predict the psychoacoustic response of the auralizations based on the psychoacoustic annoyance model from Zwicker and Fastl, which includes the effects of loudness, sharpness, fluctuation strength and roughness (Ref. 13). The model predicted the optimized rotor flyovers to be less annoying, but as stated in Krishnamurthy et al., this predicted response can only be substantiated

by psychoacoustic tests such as the one in this paper.

Further motivation for this psychoacoustic test comes from the need to assess SEL as a perceptually relevant indicator to evaluate vehicle operations and maneuvers that mitigate noise. Simulations have shown that helicopter maneuvers can greatly affect the radiated noise (Ref. 14). 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. 14). A psychoacoustic test with human subjects is needed to investigate how these predictions relate to annoyance.

This paper summarizes the RoQM-II laboratory psychoacoustic test, presents detailed analyses of the collected human responses and evaluates the efficacy of Sound Exposure Level in the rating of annoyance to helicopter noise. The helicopter noise presented to test subjects in the RoQM-II test includes (1) auralizations with optimized rotor blades, which include accurate modeling of the sound source and propagation to a ground observer and (2) ground level, sound pressure time history recordings of different flown maneuvers from a flight test, which faithfully reproduce the perceptual effect due to noise from 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. 15) at NASA LaRC. Finally, paired comparisons and annoyance responses make it possible to efficiently evaluate the efficacy of Sound Exposure Level.

Test Preparation, Stimuli and Execution

An overview of the test protocol for the RoQM-II psychoacoustic test consisted of the generation/collection of sound stimuli, subject recruitment and test design, and is briefly summarized.

The sound stimuli consisted of three auralizations of main and tail rotor noise from an AS350 helicopter, as well as four recordings of AS350 and EC130 helicopters taken from a flight test. These seven stimuli (full flyovers) ranged between 11 and $21\,\mathrm{s}$ in duration and were based on the portion of the A-weighted SPL within $10\,\mathrm{dB}$ of its maximum. In addition, short stimuli ($1\,\mathrm{s}$ in duration) centered at the beginning and ending of the full flyovers, as well as the time of maximum A-weighted SPL, were used. To eliminate popping transients and avoid startling subjects, $2\,\mathrm{s}$ and $0.2\,\mathrm{s}$ fade-ins/-outs were added to

the full flyover and short sound stimuli, respectively. All sound stimuli were reproduced in the NASA Langley Exterior Effects Room (27 satellite loudspeakers and 4 subwoofers), an acoustically treated room which presents calibrated audio to four locations where test subjects are seated. The loudspeakers are placed around the subjects (including above, front/back and left/right) so that moving sound sources are authentically reproduced.

A total of 16 test subjects (10 female and 6 male) were recruited from the community surrounding Hampton, VA. Four subjects were tested at one time. The test was divided into four sessions that were between 9 and 14 minutes in duration each. The test was designed around paired comparisons in which the test subjects indicated which sound, A or B, was the more annoying sound. Comparisons in A-B order were repeated in B-A order. Fixed stimulus levels were used, and some pairs of sounds were compared at different relative SEL levels, from which psychometric functions were generated.

More specific details of the test are given next, which 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

Three auralized flyover sounds of different rotor geometries were used as test stimuli: (1) baseline, which used the original AS350 main and tail rotor geometries, (2) SEL-optimized and (3) EPNL-optimized. Appendix A of Ref. 12 details the optimization process that changed the baseline rotor tip dihedral angle, tip sweep angle, and the tip chord to arrive at the optimized rotor geometries. The optimization used the NonSorting Genetic Algorithm II within OpenMDAO (Ref. 16). When optimizing for SEL, the objective function was the area enclosed by the $70\,\mathrm{dB}$ SEL contour and predicted a 58% reduction in the ground exposure area. When optimizing for EPNL, the objective function was the averaged EPNL over three microphone positions used for noise certification under FAR 36 Appendix H, which resulted in a $5\,\mathrm{EPNdB}$ reduction. In both cases, the flyover was at an altitude of $2000\,\mathrm{ft}$ ($609.6\,\mathrm{m}$) at 92.5 knots ($47.59\,\mathrm{m/s}$) true air speed.

For the SEL-optimized rotor case, the main rotor geometry was changed, as seen in Fig. 1. The

(a) Baseline rotor **(b) SEL-optimized rotor**

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

tip sweep angles were noticeably different between the baseline and SEL-optimized main rotors. The geometry of the EPNL-optimized rotor (not shown) was similar to the SEL-optimized rotor but with a slightly larger tip sweep angle. Each main rotor was flown with the simulated AS350 tail rotor. Auralized flyovers of the AS350 helicopter included only periodic loading and thickness noises that were synthesized from both the main and tail rotors harmonics.* Sounds emitted by helicopter components other than the rotors, such as engine noise and airframe interactions, were omitted. The tail rotor geometry was kept constant in the optimization process, but the trim settings of the tail rotor had to be adjusted when flown with each optimized main rotor. Changes in the sound of the tail rotor due to these trim and loading differences were imperceptible.

The auralization process for the AS350 helicopter rotor blade flyover test sounds used Farassat's formulation 1A (F1A) (Ref. 18) to generate sound pressures near the rotors before propagation to a ground observer. F1A synthesis, which has been described previously (Refs. 19, 20), avoids audible artifacts caused by interpolation of sound pressure magnitudes and phases at discrete prediction points (Ref. 12). This was done by computing noise at the source in the time domain sample-by-sample only at the instantaneous emission angles between the source and the receiver. 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. Blade loading data were generated for each rotor geometry and served as input to the F1A calculations for the optimized

*For a primer on the source-noise mechanisms involved in rotorcraft flight, see the publicly-available NASA RP-1258, especially Ch. 2 (Ref. 17).

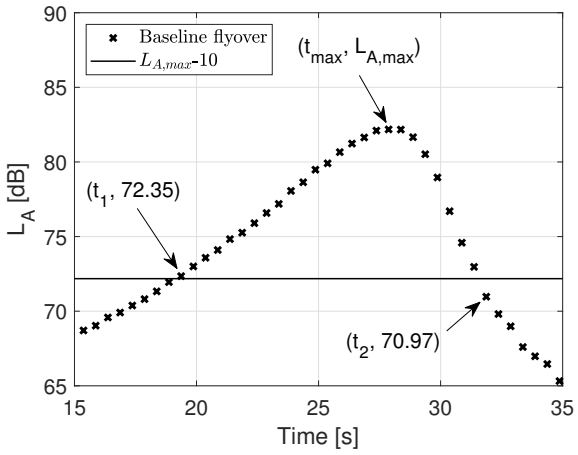


Fig. 2: A-weighted SPL time history for the baseline flyover auralization used in the psychoacoustic test, normalized to $L_{AE} = 89.8\,\mathrm{dB}$. The maximum ($t_{max}$) and boundaries of the $10\,\mathrm{dB}$ down interval (t_1 and t_2) are indicated. (Actual playback levels were lower. See Section Test Design and Execution.)

and baseline flyover sounds. F1A synthesis was implemented with the NASA Auralization Framework (NAF) (Ref. 21), which also simulated the propagation of sound from the source to a receiver.

The NAF simulated the flyovers of the AS350 helicopter main and tail rotors to generate sound at a ground observer that was then played to test subjects. Flyovers were straight and level at an altitude of $150\,\mathrm{m}$ ($492\,\mathrm{ft}$) with a constant speed of $47.59\,\mathrm{m/s}$ ($92.5\,\mathrm{kn}$). Rotors flew along a centerline path directly over a ground observer, which was flush with rigid ground. 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.

The A-weighted sound pressure level of a segment of the baseline rotor flyover is shown in Fig. 2. The horizontal line may be used to visualize the portion of the time history within $10\,\mathrm{dB}$ of the maximum level, corresponding to the portion used in the psychoacoustic test[†]. Three points that are used in the SEL calculation in Eq. (1) are noted in the figure: $(t_1,\ L_A(k_1))$, $(t_{max},\ L_{A,max})$ and $(t_2,\ L_A(k_2))$.

In Fig. 3a, the A-weighted SPL time histories for the baseline and SEL-optimized auralizations used in the psychoacoustic test are shown (i.e., only the part that is within $10\,\mathrm{dB}$ of the maximum A-weighted

[†]The levels for playback during the psychoacoustic test were lower than shown in order to avoid subject fatigue. See Test Design and Execution section for details.

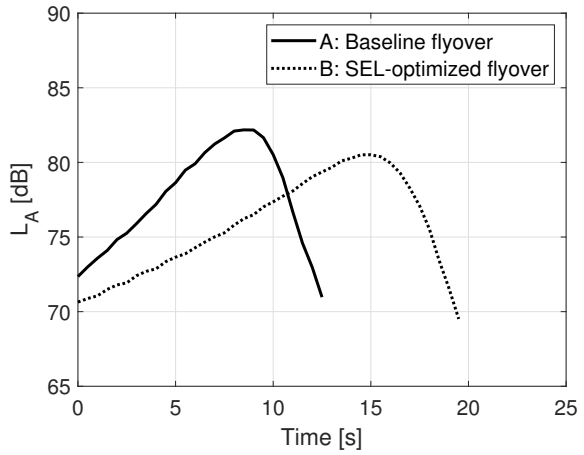
SPL). The baseline and optimized flyovers have been normalized to have the same SEL (SEL of both optimized flyovers were $4.2\,\mathrm{dB}$ lower than the baseline flyover before normalization). The time histories begin at $0\,\mathrm{s}$, because t_1 was subtracted from the time series for each auralization. As a result, differences in duration and shape of the A-weighted SPL profile are more evident. The EPNL-optimized and SEL-optimized time histories are virtually identical, as shown in Fig. 3b. They are about $7\,\mathrm{s}$ (or about 50%) longer than the baseline and have a maximum A-weighted SPL that is $1.67\,\mathrm{dB}$ lower. These apparent differences in duration and amplitude are solely due to a difference in the noise characteristics between the baseline rotor and the optimized rotors, since all three auralizations were simulated at the same flight speed.

Sound files of the auralizations (as well as all other acoustic stimuli) used in the psychoacoustic test are publicly available and can be downloaded in Waveform Audio File Format (WAV) from Ref. 22.

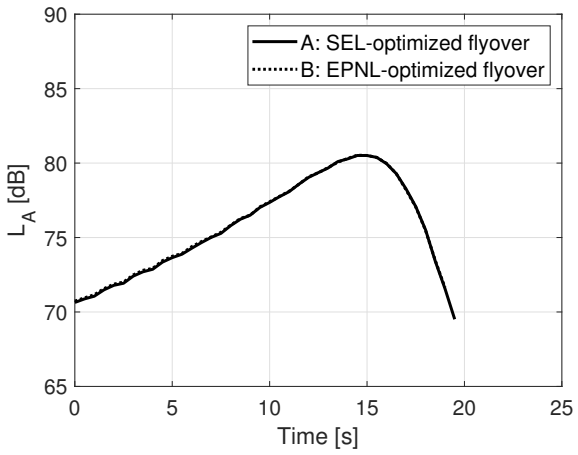
Flight Test Data

In addition to the auralizations just described, recordings from actual flight tests are also included as stimuli in the psychoacoustic test. The reason for this is that flights contain temporal, spectral and spatial variations in the noise that are not easily simulated, and these variations may be relevant in the evaluation of SEL. The recordings for this test were drawn from a Noise Abatement Flight Test conducted jointly by NASA, the FAA and the U.S. Army. The flight test involved six different helicopters and included level flights, climbs, descents and different types of turns over a 52-microphone array (Ref. 23). The turns consist of different operational conditions, such as constant speed/torque or acceleration/deceleration, that can affect not only the sound levels 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 sound level and/or spectral content (Ref. 14).

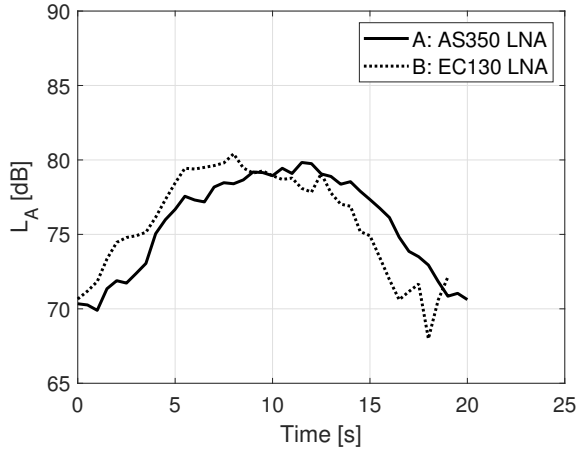
The recorded stimuli chosen for this test are from flights of the Eurocopters AS350 and EC130, two helicopters of the same manufacturer with similar size and capabilities but with different tail rotor technologies (conventional or ducted). Photographs of the two helicopters are shown in Fig. 4. Both helicopters have 3 clockwise-rotating main rotor blades with a diameter of $10.69\,\mathrm{m}$ ($35.07\,\mathrm{ft}$) and are powered by gas



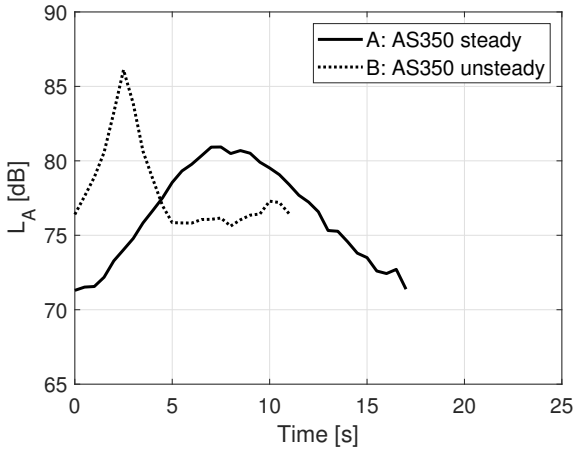
(a) Pair 1: Baseline and SEL-optimized auralizations



(b) Pair 2: SEL- and EPNL-optimized auralizations



(c) Pair 3: AS350 and EC130 low noise approach recordings



(d) Pair 4: AS350 steady and unsteady recordings

Fig. 3: A-weighted SPL time histories for full flyover stimuli used in the psychoacoustic test, normalized to $L_{AE} = 89.8\,\mathrm{dB}$. (Actual playback levels were lower. See Section Test Design and Execution.) The abscissa starts with the first instance within $10\,\mathrm{dB}$ from the maximum.

turbine engines. The largest difference between the two helicopters is that the EC130 has a Fenestron (ducted tail rotor) while the AS350 has a more conventional tail rotor. Both helicopters were recorded at the Amedee Army Airfield in Lassen County, California (Ref. 23). The AS350 and EC130 had takeoff

(a) AS350 **(b) EC130**

Fig. 4: Photographs of helicopters whose recorded flights were used as sound stimuli in the psychoacoustic test (Ref. 23).

gross weights (TOGW) of 4,247 (1,926.4) and 4,529 (2,054.3) pounds (kg), respectively.

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 with input from the manufacturer and consideration of the Fly Neighborly Guidelines from the Helicopter Association International (Ref. 23). 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 the A-weighted SPL 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 distinct regions of different source characteristics throughout the flyover. Sudden changes in temporal, spectral and spatial impression were perceived as the sound shifted between different portions of the flyover containing more impulsive blade-vortex interaction (BVI) noise, broadband noise, and tonal noise apparently dominated by emissions from

Table 1: Selected recordings from the Noise Abatement Flight Test (Ref. 23).

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

the tail rotor and turbine.

The A-weighted SPL time histories for the recorded flights listed in Table 1 are shown in Figures 3c and 3d. As in Fig. 3a, t_1 is subtracted from the time series for each flight, and all are normalized to $L_{AE} = 89.8\,\mathrm{dB}$. The low noise approaches (Fig. 3c) for the two helicopters have a similar duration and maximum A-weighted SPL, and the EC130 reaches its maximum A-weighted SPL about $4\,\mathrm{s}$ earlier than the AS350. The steady AS350 flight has a similar profile to the low noise approaches but differs markedly with the unsteady flight (Fig. 3c). The unsteady flight is an accelerating turn and has a steep rise in A-weighted SPL at the beginning, reaching its maximum at around $2.5\,\mathrm{s}$. It has a roughly constant A-weighted SPL after $5\,\mathrm{s}$ and a slight rise near the end.

Sound files of the recordings (as well as all other acoustic stimuli) used in the psychoacoustic test are publicly available and can be downloaded in Waveform Audio File Format (WAV) from Ref. 22.

Test Design and Execution

The psychoacoustic test was designed using paired comparisons (Ref. 24). The four pairs are listed in Table 2 and detailed below, along with the particular research questions each pair addresses. All acoustic stimuli used in the psychoacoustic test are publicly available and can be downloaded from Ref. 22. Each research question (RQ) is a different test of the efficacy of SEL in the rating of annoyance to helicopter noise.

Pair 1: Baseline flyover vs. SEL-optimized flyover

Table 2: Stimuli used for each pair of sounds in the psychoacoustic test. Auralizations are in italics; recordings are in regular font.

Pair	A	B
1	*Baseline*	*SEL-optimized*
2	*SEL-optimized*	*EPNL-optimized*
3	AS350 LNA[a]	EC130 LNA[a]
4	AS350 steady	AS350 unsteady

[a] low noise approach

RQ1a: When presented at the same SEL, is there a perceived difference in annoyance between a baseline flyover and one where the main rotor is optimized in terms of SEL?

RQ1b: By how much should the SEL of the SEL-optimized flyover be adjusted (relative to the baseline) in order to give an annoyance response equal to that of the baseline?

Pair 2: EPNL-optimized flyover vs. SEL-optimized flyover

RQ2: When presented at the same SEL, is there a perceived difference in annoyance for a flyover where the main rotor is optimized in terms of EPNL instead of SEL?

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

RQ3a: When presented at the same SEL, is there a perceived difference in annoyance between an AS350 low noise approach and an EC130 low noise approach?

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

Pair 4: AS350 unsteady flight vs. AS350 steady flight

RQ4a: When presented at the same SEL, is there a perceived difference in annoyance between a steady flight and an unsteady flight?

RQ4b: By how much should the SEL of the unsteady AS350 flight be adjusted (relative to the steady one) in order to give an annoyance response equal to that of the steady flight?

RQ5: What can perceptual differences of short samples of the flyover at t_1, t_{max} and t_2 tell us about SEL?

The first research question for pairs 1-4 involved comparing each pair of sounds at the same SEL and identified a qualitative difference in annoyance. To quantify this difference, the second question for pairs 1, 3 and 4 was asked: by how much should the SEL of sound B be adjusted (relative to sound A) in order to give an annoyance response equal to that of sound A? This relative change in SEL 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 $0,\ \pm5$ and $\pm10\,\mathrm{dB}$. For pairs 3 and 4, the relative levels were $0,\ \pm4$ and $\pm8\,\mathrm{dB}$. 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.

Research Question 5 addresses the fact that there is little perceptual information contained in Eq. (1) other than A-weighted SPL. Answering RQ5 involved re-calculating SEL after adjusting the A-weighted SPL time history based on perceptual adjustments at t_1, t_{max} and t_2. Short sounds centered at these points of the flyover 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 flyover was compared to t_1 of the SEL-optimized flyover but not to t_{max} or t_2 of the SEL-optimized flyover). The relative levels of the short sounds were the same as those of the long sounds ($0,\ \pm5$ and $\pm10\,\mathrm{dB}$ for pair 1 and $0,\ \pm4$ and $\pm8\,\mathrm{dB}$ for pair 4). The perceptually adjusted SEL will be used to identify other aspects that may not be represented in Eq. (1).

For RQ1b, RQ3b, RQ4b, and RQ5, the method of constant stimuli was used in order to test more than one subject at a time. Typically in the literature, adaptive methods are used in these types of experiments due to their efficiency and ability to concentrate stimuli levels near the desired threshold (EAP, in this case). However, for this experiment, adaptive methods are not preferred since they only work for one subject at a time. The method of constant stimuli is not overly inefficient relative to adaptive methods when the upper and lower bounds of stimuli levels can be properly estimated. As a result, the potential gain

in testing efficiency for adaptive methods should not outweigh the fact that four subjects can participate at once when constant stimuli are used (Ref. 25).

The selected comparisons of full flyovers and short stimuli are summarized in Table 3. There were 16 unique full flyover paired comparisons (used to answer RQs 1-4) and 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 full flyover comparisons to 32 and the number of short sound comparisons to 60. For both the full flyover and short sound comparisons, the order was randomized once, so that each group of test subjects heard the same order. A total of 16 subjects was tested in four groups with four subjects per group. The full flyover and short sound comparisons were each played to every group over two back-to-back sessions, for a total of four sessions for each group. Each group listened to all 92 comparisons over the four sessions. Groups 1 and 4 listened to the 2 sessions of short sounds first, while Groups 2 and 3 listened to the 2 sessions of full flyovers first. The sessions containing full flyovers lasted about 14 minutes, and the sessions containing short sounds lasted about nine minutes. After session four, the subjects were also given the opportunity to provide written responses about the test stimuli or any other aspect of the study. Along with familiarization and practice sessions, as well as pre- and post-test audiograms, the total time spent by each subject was just over two hours. No response data (e.g., outliers) were removed; all response data were used in the analysis.

The psychoacoustic test was performed in the EER at NASA LaRC (Ref. 15). The EER has 27 satellite loudspeakers mounted on the walls and ceiling and four subwoofers placed in the corners. Vector base 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 background noise during the test was at or below the NC-15 noise criterion curve (Ref. 26).

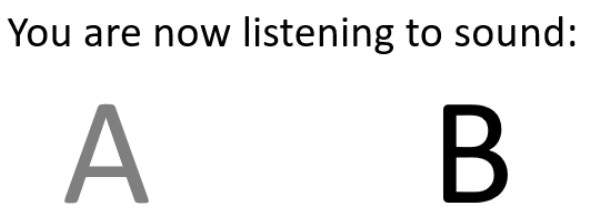
The sampling rate used in the EER is $44.1\,\mathrm{kHz}$. The auralizations were simulated at this sampling rate, but the recordings were up-sampled from 25 or $25.5\,\mathrm{kHz}$. To eliminate popping transients and avoid startling subjects, $2\,\mathrm{s}$ and $0.2\,\mathrm{s}$ fade-ins/-outs were added to the full flyover and short sound stimuli, respectively.

Test subjects responded to the sound stimuli via computer tablets. When comparing sounds A and B, a

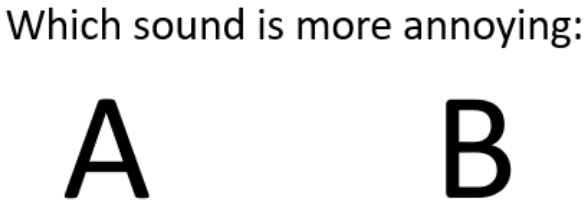
Table 3: Description of comparisons made in the psychoacoustic test. Auralizations are in italics; recordings are in regular font.

Pair	Sound	Description	Full flyovers	Different levels	Short stimuli t_1, t_{max} and t_2
1	A	*Baseline flyover*	yes	yes	yes
	B	*SEL-optimized flyover*			
2	A	*SEL-optimized flyover*	yes	no	no
	B	*EPNL-optimized flyover*			
3	A	AS350 LNA[a]	yes	yes	no
	B	EC130 LNA[a]			
4	A	AS350 steady	yes	yes	yes
	B	AS350 unsteady			

[a] low noise approach



(a) Gray letter "A" indicates that sound A is currently being played.



(b) After both sounds are played, the above question is displayed.

Fig. 5: Test instructions presented to the subjects via computer tablet.

box appeared on the tablet as shown in Fig. 5a. A gray letter indicated the sound that was currently being played. After each pair of sounds was presented, the subject was asked to select the sound that was more annoying, as shown in Fig. 5b.

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, resulting in 10 female and 6 male subjects. The protocol for the psychoacoustic test was approved by the NASA Langley Institutional Review Board.

The test design required that the basic comparisons were presented to the subjects at equal SEL. During pilot testing, a set of recordings from the flight test had an average SEL of $89.8\,\mathrm{dB}$. 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

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

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 SEL value of $L_T = 89.8\,\mathrm{dB}$. A similar normalization was done for the short sounds in terms of A-weighted SPL in which the target was the mean A-weighted SPL of the two sounds and $a = 10^{(L_T - L_{A,i})/20}$.

In order to avoid subject fatigue, an intended SEL of $67.2\,\mathrm{dB}$ was used 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 subject seats measured SEL. The resulting mean and standard deviation of the SEL for the 26 normalized recordings were $67.7\,\mathrm{dB}$ and $0.5\,\mathrm{dB}$, respectively. The mean SEL value is within 1 standard deviation of the intended playback level, which is considered sufficiently accurate for the purposes of this test. Adjustments to relative levels described previously were applied to the intended SEL, meaning a relative level of $\Delta L_{AE} = 2\,\mathrm{dB}$ corresponds to sound B at $69.2\,\mathrm{dB}$ and sound A at $67.2\,\mathrm{dB}$.

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 SEL 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 the EAP. Finally, perceptual adjustments at three points of the A-weighted SPL time history of flyover events provide insight into

perceptual cues not accounted for by SEL.

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. 27). In this case, the "successes" are considered the number of times sound B is judged more annoying than sound A. Since both sounds have the same SEL, 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 SEL, that affected the subjects' annoyance choice. A confidence interval on the percentage of successes is also calculated; when this confidence interval does not overlap 50%, sounds A and B caused a significantly different annoyance response. Two-sided binomial tests are used here, because there is no assumption made about which sound in each pair may be more annoying than the other.

Equal Annoyance Point (EAP)

The binomial test helps determine whether two sounds of equal SEL are perceived to be equally annoying or not. However, 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 determine the change in SEL 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 SEL and that B is varied up or down in magnitude. When sound B has a higher SEL, it is often, but not always, determined to be more annoying than sound A. Similarly, when sound B has a lower SEL, 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 likelihood approach (Ref. 28, p. 118).

In the probit model, the link function gives the relationship between the mean response and a linear

combination of predictors. The inverse of the 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 SEL 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, which are the intercept and slope parameters, respectively. 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 SEL.

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. More importantly, if the confidence interval on the EAP contains $0\,\mathrm{dB}$, then the EAP is not *significantly* different from $0\,\mathrm{dB}$, and from a statistical point of view, it cannot be determined if the two sounds are different in terms of annoyance.

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

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

in which $\delta(\beta_0)$ and $\delta(\beta_1)$ are the standard error of the regression coefficients. The confidence interval on EAP is then $-\beta_0/\beta_1 \pm 1.96 \times \delta(\mathrm{EAP})$. This model 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 a normally distributed EAP. While this 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 computed.

Advanced confidence interval for EAP

In order to get a more accurate estimate of the confidence interval, a more advanced approach is used. This more general method 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 (CDF), Φ, with mean, μ, and standard deviation, σ. This is given by

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

In Eq. (5), the normal CDF shape is used as a sigmoid shape to fit the data. Equation (5) 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.

The form of Φ shown in Eq. (5) 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. More details of the MCMC analysis technique are given in the Appendix, including an example random walk and how the likelihood of μ and σ are calculated. The output of the MCMC simulation is a distribution of possible values for μ and σ, referred to as resamples, where the 95% confidence interval around the EAP is given by the quantiles bounded by 2.5% and 97.5% of the μ resamples.

Three-point perceptually-adjusted SEL (TPPAS)

The three-point perceptually adjusted SEL (TPPAS) approach is an alternative way to calculate SEL, after making perceptual adjustments to the A-weighted SPL time history at the beginning, end and maximum (t_1, t_2 and t_{max}, respectively) of a complete flyover. While it is not meant to replace SEL or full flyover comparisons, it is used to gauge the importance of noise characteristics not accounted for by

SEL. Since the calculation of SEL includes little perceptual information beyond A-weighting, the TPPAS approach augments the SEL calculation by incorporating perceptual responses collected from the psychoacoustic test. This section outlines the main steps of the TPPAS approach (with greater detail given in the Appendix) and describes how it can be used to evaluate SEL.

The TPPAS approach was first developed 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 commercial fixed-wing concept vehicles was investigated (Ref. 29). In that study, the auralizations were longer than $45\,\mathrm{s}$, so it was thought impractical to ask subjects to compare such long noise stimuli. Using this approach, it was concluded that reductions in EPNL correspond with a reduction in annoyance, but the "perceived difference in EPNL...is significantly less than that indicated by the metric" (Ref. 29).

The TPPAS approach starts with collecting annoyance comparisons to short sounds (approximately $1\,\mathrm{s}$ in duration) centered around t_1, t_{max} and t_2 from two flyovers (A and B). These comparisons are made at different relative A-weighted SPLs between the two stimuli. Next, an MCMC simulation for each control point results in distributions of likely adjustments of the A-weighted SPL to be made at the three control points. Since these resamples are assumed to be a random sampling, they are used as bootstrapping samples. Taking one sample at each of the three control points and interpolating the adjustment between t_1, t_{max} and t_2 gives an adjusted A-weighted SPL time history, and Eq. (1) is used to give a possible adjusted SEL. Doing this for all resamples gives a distribution of adjusted SEL values. These adjustments are made to the A-weighted SPL time histories of both signals A and B, which are equal and opposite.

The TPPAS approach serves as another way to evaluate the efficacy of SEL in the rating of annoyance to helicopter noise. If the confidence interval of the adjusted SEL does not overlap with the original SEL, it means that important perceptual effects exist that are not accounted for in Eq. (1). Furthermore, adjustments can be made to the A-weighted SPL time history of either sound A or sound B. If differences arise from these adjustments, it indicates that important perceptual differences exist in between the control points, highlighting a shared deficiency between TPPAS and SEL.

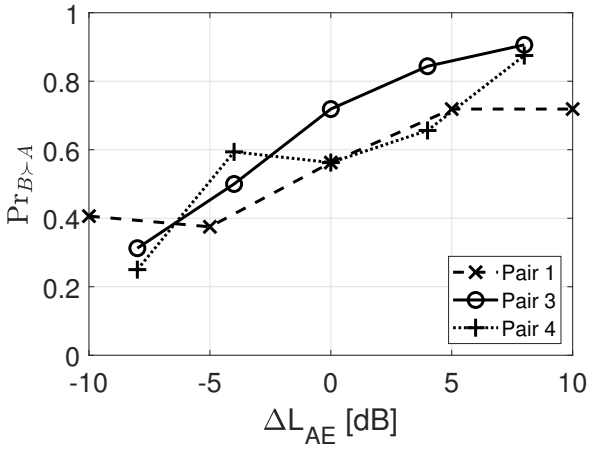


Fig. 6: Raw results from the full flyover comparisons. Pair 2 (not shown) was only tested at $\Delta L_{AE} = 0$, for which $\Pr(B \succ A) = 0.56$.

Results

Figure 6 shows the raw results of the psychoacoustic test in terms of the probability of sound B being judged more annoying than sound A for each pair of sounds. The general trend is that when sound B is played at a higher SEL, it is more likely to be judged more annoying, as expected. However, the data points are not guaranteed to be monotonically increasing due to the statistical nature of the subject responses.

Analyses for each pair consist of binomial test results for sounds played at the same SEL, a curve fit to the data when the two sounds are played at different relative levels, determination of the EAP and calculating the confidence interval on the EAP. The tables and figures discussed below aid in the analyses.

Table 4 shows the results of two-sided binomial tests for the four pairs of sounds. The binomial test only considers the responses when sounds A and B were played at the same SEL. The number of responses per pair, N, was 32. The percentage (%) is the number of responses in which sound B was judged more annoying than sound A, divided by the total number of responses for that pair. A confidence interval for the percentage is also given. The p-value, p, is considered statistically significant when $p < 0.05$ and indicates that the null hypothesis (sounds A and B are equally annoying) should be rejected.

The binomial test results in Table 4 only consider the comparisons of sounds that were presented at the

Table 4: Results of binomial tests of four pairs of flyovers when sounds A and B are presented at the same SEL. %: percentage of responses where sound B was judged more annoying than sound A with confidence interval $[\bullet, \bullet]$. p: p-value. N: number of responses. Auralizations are in italics; recordings are in regular font.

Pair	Sound	Description	%	p	N
1	A	*Baseline* flyover	56 [37, 73]	0.60	32
	B	*SEL-optimized flyover*			
2	A	*SEL-optimized flyover*	56 [37, 73]	0.60	32
	B	*EPNL-optimized flyover*			
3	A	AS350 LNA[a]	72 [51, 85]	0.02	32
	B	EC130 LNA[a]			
4	A	AS350 steady	56 [37, 73]	0.60	32
	B	AS350 unsteady			

[a] low noise approach

same SEL. While the conclusions are indicative of the efficacy of SEL in annoyance ratings of helicopter noise, more insight is gained by comparing the sounds for various relative SEL values. In particular, probit fits lead to an estimate of the Equal Annoyance Point (perceptually equivalent level) and its confidence interval. The binomial test gives a (vertical) confidence interval on the probability that sound B is more annoying than sound A when $\Delta L_{AE} = 0$, and the probit fit gives a (horizontal) confidence interval on the Equal Annoyance Point when sounds A and B are played at different relative SEL levels.

Figure 7 shows the same response data as Fig. 6 but with the addition of probit fits. (The SEL- and EPNL-optimized rotors, pair 2, were not played at different relative SEL, because the Equal Annoyance Point was assumed to be very small.) The intersection of each probit fit with $\Pr = 0.5$ gives the EAP for each pair. Since the EAPs happen to be negative, it suggests that the SEL of sound B should be a few dB less than that of sound A (for each pair) in order for the two sounds to be equally annoying. However, to

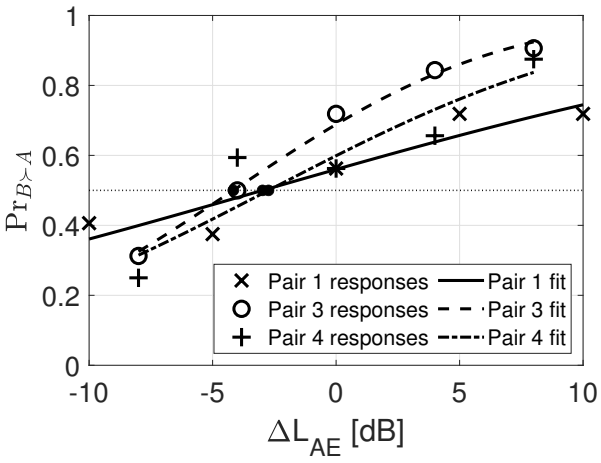


Fig. 7: Probability that sound B is more annoying than sound A for pairs 1, 3 and 4. Intersections of the probit fits with the horizontal line show the Equal Annoyance Point for each pair (solid circles).

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 5 and depicted graphically in Fig. 8. The confidence intervals using Eq. (4) are also shown. To be clear, the assumptions using Eq. (4) are that the standard errors of both probit regression parameters have no covariance and that the standard error on EAP is normally distributed. To check the validity of these assumptions, results are compared with the more advanced estimation of the confidence interval using MCMC simulations, which are presented in Table 5. The confidence intervals for pairs 1 and 4 are quite symmetric. However, the confidence interval for pair 3 is asymmetric, indicating that EAP may be as low as $-7.93\,\mathrm{dB}$. In general, the confidence intervals are quite wide, up to $8.6\,\mathrm{dB}$ for pair 1, Eq. (4), indicating a high variability among subjects.

The analysis will start with pair 3, the comparison between the AS350 and EC130 low noise approaches. Then, pair 4 is discussed, which describes the differences in annoyance between steady and unsteady recorded maneuvers of the AS350. This is followed by a discussion of the auralizations using the optimized rotor blades (Pairs 1 and 2). After the pairs are analyzed individually, the efficacy of SEL is discussed in terms of the TPPAS approach for pairs 1 and 4.

Table 5: Equal Annoyance Point (EAP) with confidence intervals in terms of SEL (dB) 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. Auralizations are in italics; recordings are in regular font.

				CI	
Pair	Sound A[a]	Sound B	EAP (ΔSEL)	Eq. (4)	MCMC
1	*Baseline*	*SEL-optimized*	-2.98	[-7.26, 1.31]	[-5.33, 0.57]
3	AS350 LNA[b]	EC130 LNA[b]	-4.16	[-6.55, -1.78]	[-7.93, -2.39]
4	AS350 steady	AS350 unsteady	-2.74	[-5.26, -0.21]	[-5.16, -0.56]

[a] reference

[b] low noise approach

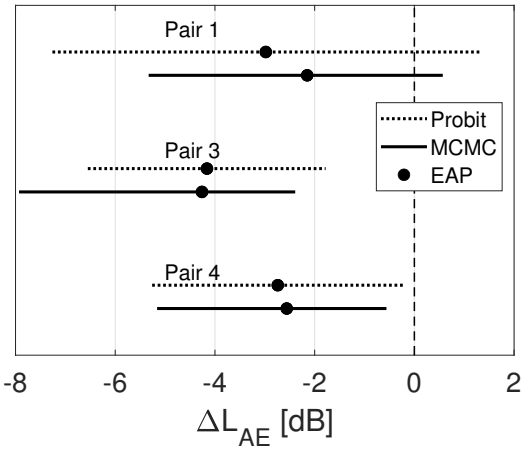


Fig. 8: Summary of full flyover Equal Annoyance Point and confidence intervals using probit regression and MCMC simulations.

Pair 3: AS350 vs. EC130 recordings of low noise approaches

The results regarding the recorded low noise approaches for the AS350 and EC130 helicopters are presented here.

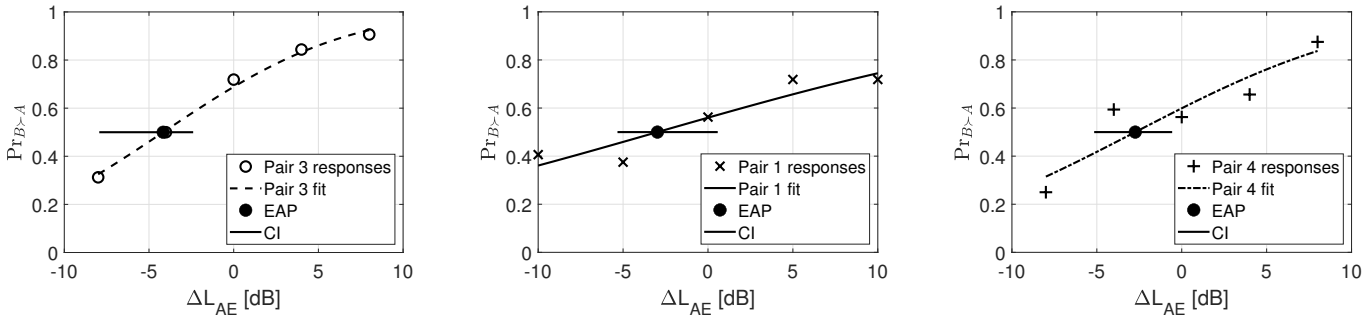
When only considering the case when the AS350 and EC130 are played at the same SEL, a binomial

test shows that the percentage of responses choosing the EC130 low-noise approach flyover to be more annoying than the AS350 flyover was significant. From Table 4, the null hypothesis (that both sounds are equally annoying) should be rejected (p-value$<$0.05 and confidence interval on percentage does not overlap 50%). Since both helicopter flyovers were played at the same SEL and performed the same maneuver, it is likely that the EC130 is more annoying than the AS350 and that SEL did not capture the difference in perception.

Considering the cases when the AS350 and EC130 were also played at different relative SEL levels (see Table 5), the results from the probit model show that the null hypothesis should be rejected for pair 3, because $\Delta L_{AE} = 0$ is outside the confidence interval of the EAP. From this result, it is concluded that when played at the same SEL, the EC130 low noise approach is perceived to be significantly more annoying than the AS350 low noise approach. This conclusion agrees with the result of the binomial test.

An MCMC simulation was run, and its results are compared to the results from the binomial test and when using the probit model. See the Appendix for more details. The CI of the EAP is also plotted on the probit fit in Fig. 9a and shown in Table 5. The CI is not symmetric about the EAP and is $0.77\,\mathrm{dB}$ 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 with the simple CI given by Eq. (4), indicating 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 SEL.

It is unlikely that the difference in perception between the two helicopters came from the A-weighted SPL time history, because both profiles, including the maximum A-weighted SPL, are quite similar (see Fig. 3c). The two helicopters also come from the same manufacturer and are similar in engine size and load capacity. The perceptual difference may be due to differences in qualitative aspects of the sounds as a result of the Fenestron on the EC130. It indicates that different tail rotor 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. 6–8). 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. Further analysis confirmed



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

(b) Probability that the SEL-optimized flyover is more annoying than the baseline flyover.

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

Fig. 9: Probit models and EAPs with confindence intervals for full flyover comparisons.

that the EC130 has higher levels of fluctuation strength and tonality relative to those of the AS350. This result shows that SEL may fail to capture significant qualitative aspects of annoyance responses when comparing two different helicopters.

Although playback levels were at an average SEL of $67.2\,\mathrm{dB}$ (see Section Test Design and Execution), a practical question is 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 $89.51\,\mathrm{dB}$, respectively, a relative difference of $\Delta L_{AE} = 0.36\,\mathrm{dB}$. Since this is more than $2\,\mathrm{dB}$ outside the ranges for the EAP confidence intervals found in Table 5 for pair 3 and is over $4\,\mathrm{dB}$ from the EAP, it is concluded that the EC130 would also be more annoying than the AS350 even if they were compared at their recorded absolute levels.

Pair 4: steady vs. unsteady AS350 recordings

The results regarding the recorded steady and unsteady flights of the AS350 are now discussed.

When only considering the case when the steady and unsteady AS350 flyovers are played at the same SEL, a binomial test shows that the percentage of responses choosing the unsteady flyover to be more annoying than the steady flyover was not significant. From Table 4, the null hypothesis (that both sounds

are equally annoying) cannot be rejected (p-value>0.05 and confidence interval on percentage overlaps 50%). Since both helicopter flyovers were played at the same SEL and there was close to an even probability of annoyance responses, it remains a possibility that the steady and unsteady flyovers are equally annoying.

A more complete picture is revealed by examining the annoyance responses at different relative levels between the two sounds, as shown in Fig. 7. The EAP for pair 4 found through the probit model is shown in Table 5, along with the confidence interval using Eq. (4). Since $\Delta L_{AE} = 0$ is outside the confidence interval of the EAP, the null hypothesis should be rejected for pair 4. From this statistically significant result, it is concluded that, when played at the same SEL, the unsteady AS350 maneuver is perceived to be more annoying than the steady AS350 maneuver.

The results from the binomial test and the probit model do not agree for pair 4. This can be understood by inspecting the responses shown in Fig. 7. The subjects' responses show that the probability that sound B is more annoying than sound A is 0.56 at $\Delta L_{AE} = 0$. Looking at only this data point is inconclusive, but when the other data points are considered, the probit fit predicts a higher $Pr(B \succ A) = 0.60$ at $\Delta L_{AE} = 0$. Raising of the curve lowers the predicted EAP as well as its confidence interval so that the confidence interval does not include 0. Since more data goes into the probit fit analysis, this is the preferred method. Therefore, the unsteady flight is perceived to be *slightly* more annoying than the steady flight.

Again, 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, leading to a symmetric confidence interval. The MCMC simulations are performed to check these assumptions, the results of which are shown in Fig. 9c and Table 5.

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. This supports the conclusion that the unsteady AS350 flight is perceived to be slightly more annoying than the steady one when presented at the same SEL. Further analysis of sound quality metrics for these two sounds confirmed an increase in impulsiveness for the unsteady flight and an even larger increase of fluctuation strength. It suggests that SEL 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 response to the unsteady sound could be that some subjects reported (in a post-test questionnaire) that unsteadiness in the sound was interpreted as an unsteadiness in the flight or control of the vehicle, which signaled an elevated safety risk for some listeners. This notion is supported by many other studies in which fear was determined to be a moderating factor on an individual's annoyance to noise (Ref. 30).

Another reason for the disparity is the tradeoff between the maximum level and the length of a flyover. In typical SEL computations on "haystack" events, where the time history is first a (mainly) increasing function up to the maximum of the noise and then a decreasing function as the aircraft flies away, there will be a natural tradeoff between the maximum level that is achieved and the length of time over which the integration takes place. This is evidenced quite clearly in Fig. 3d for the steady/unsteady pair. In the case of pair 4, the shorter sound with the higher maximum level is found to be more annoying.

Pairs 1 and 2: auralizations of noise-optimized blades

The results regarding the auralizations with the noise-optimized blades are now discussed.

When only considering the case when the baseline and SEL-optimized flyovers are played at the same SEL, a binomial test shows that the percentage of responses choosing the optimized flyover to be more annoying than the baseline flyover was not significant. From Table 4, the null hypothesis (that both sounds are equally annoying) cannot be rejected (p-value$>$0.05 and confidence interval on percentage overlaps 50%). Since there was close to an even probability of annoyance responses, it remains a possibility that the baseline and SEL-optmized flyovers are equally annoying when played at the same SEL.

The null hypothesis also cannot be rejected for pair 2, meaning that the EPNL-optimized flyover is not significantly more annoying than the SEL-optimized flyover when played at the same SEL.

For the SEL-optimized flyover, a more complete picture is revealed by examining the annoyance responses at different relative levels with respect to the baseline flyover auralization, as shown in Fig. 9b. The EAP for pair 1 found through a probit model is shown in Table 5. The confidence interval using Eq. (4) is also shown. For the baseline vs. SEL-optimized flyover pair, the confidence interval contains

0, indicating that the null hypothesis cannot be rejected. From this result, it is concluded that the flyover with the SEL-optimized rotor design is not perceived to be significantly more annoying than the baseline flyover when both flyovers are played at the same SEL. The result of the probit model for pair 1 in Table 5 agrees with the binomial result in Table 4, confirming the conclusions that the null hypothesis cannot be rejected for pair 1.

The results from the MCMC simulation give further insight into this conclusion. As shown in Fig. 9b and Table 5, 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 $2.67\,\mathrm{dB}$ 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 SEL-optimized flyover is not significantly more annoying than the baseline flyover when presented at the same SEL.

Nevertheless, the estimate of the EAP for the SEL-optimized flyover compared to the baseline flyover is $-2.98\,\mathrm{dB}$, indicating that perception gravitates toward being less annoying for the baseline flyover than for the SEL-optimized flyover when both flyovers are played at the same SEL. Although not conclusive, one possible cause for this is that there is an aspect of the sound in the optimized flyover that is more annoying than the baseline flyover. This could be due to the fact that only the main rotor was optimized (the tail rotor geometry was left unchanged). Reducing the SEL of the main rotor while leaving the tail rotor mostly constant, then normalizing in terms of SEL means that the tail rotor noise is more prominent compared to the main rotor in the SEL-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 annoyance response even though the main rotor noise was reduced. Further analysis showed that the flyover stimuli with the optimized main rotors had an increase in tonality when compared to the baseline flyover. The results suggest that annoyance might be further reduced if both main and tail rotor geometries are optimized simultaneously.

Another possibility for why the SEL-optimized flyover is perceived to be slightly more annoying when played at the same level as the baseline, as discussed earlier, is that there is a tradeoff between a shorter sound with higher maximum level and a longer sound. The duration of the $10\,\mathrm{dB}$-down-time interval for the SEL-optimized flyover is 50% longer than that of the baseline flyover (see Fig. 3a). Through written

responses, several subjects reported that sounds that appeared to loiter (i.e., longer sounds) were found to be more annoying, in general. On the other hand, the shorter sound with the higher maximum level in pair 4 was also found to be more annoying. In the case of pair 1, the longer sound is indicated (though not significantly) to be more annoying. Therefore, there seems to be no *consistent* bias to this tradeoff when looking across all of the pairs of sounds in this test. The length of a sound can act as a penalty or reward, as loitering and startling have opposite effects. This is supported by both the annoyance responses as well as through self-reported impressions collected from the post-test questionnaire.

The case in which the flyovers are played at their original levels (i.e., before set to the same SEL) also deserves discussion. Before being adjusted to the same SEL, the original two sounds for Pair 1 had a difference in SEL of -4.22 dB. Both confidence intervals for the EAP recovered for this pair overlap -4.22 dB. This indicates that if the psychoacoustic test were done on the pair of sounds played at their original SEL levels, then the result of the test would have been that the baseline and optimized flyovers could not be distinguished in terms of annoyance even though the optimization removed more than 4 dB of noise. It is possible that having more test subjects would have decreased the width of the confidence interval on the EAP, further distinguishing the EAP from the ΔL_{AE} of the original auralizations. On the other hand, reducing the tail rotor noise also could have made the difference in annoyance more pronounced. As shown in Table 5, the MCMC lower bound of the CI is -5.33 dB. If optimizing the tail rotor noise led to a total reduction (main and tail rotor combined) of more than 5.33 dB, then the SEL-optimized flyover would be perceived to be less annoying than the baseline flyover when played at their absolute levels. Since only optimizing the main rotor reduced SEL by 4.22 dB, a total noise reduction of 5.33 dB is considered achievable if the tail rotor was also optimized. This assumption is supported in that previous auralizations of only the main rotor reduced SEL and EPNL an average of 14 dB (Ref. 12).

This also suggests that perceptual differences in annoyance would likely be comparable to the designed reduction in SEL if both the main rotor and tail rotor were optimized. Since there was not a significant difference between the SEL- and EPNL-optimized flyovers, either metric may be considered suitable as a design criteria and that there are no obvious 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 SEL or EPNL were not perceived to be significantly different in terms of annoyance.

Three-point perceptually adjusted SEL (RQ 5)

This section discusses the results of the TPPAS approach and how they can be used to evaluate the efficacy of SEL as a metric for rating annoyance to helicopter noise. As explained in the Analysis Techniques section (and in more detail in the Appendix), TPPAS generates perceptually-adjusted A-weighted SPL 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 A-weighted SPL and analyzed with MCMC simulations. This gave a range of possible adjustments to the A-weighted SPL time history and a distribution of possible perceptually adjusted SEL values.

An example of annoyance comparisons of short sounds centered around t_{max} for the steady/unsteady AS350 flyover (pair 4) are shown in Fig. 10. It shows the probability that the unsteady flyover was perceived to be more annoying than the steady flyover as a function of its A-weighted SPL (relative to the steady flyover). A probit model was fit to the responses and the Equal Annoyance Point was found. The MCMC simulation gave the confidence interval on the EAP, and the MCMC resamples give the likely perceptual adjustments to be made at t_{max}. The results show that the unsteady flyover should be $2\,\mathrm{dB}$ higher than the steady flyover to be equally annoying when subjects only compare short stimuli centered around t_{max}. Since $\Delta L_A = 0$ lies outside the confidence interval, this perceptual difference is considered to be significant; the A-weighted SPL alone does not capture the perceptual difference.

Short stimuli centered around t_1 and t_2 were also compared for the steady/unsteady flight pair, and comparisons at all three control points of the A-weighted SPL time history for the baseline/SEL-optimized flyover were made. For each comparison, an MCMC simulation was run that gave the possible perceptual adjustments to the A-weighted SPL time history and a confidence interval on the EAP (see the Appendix

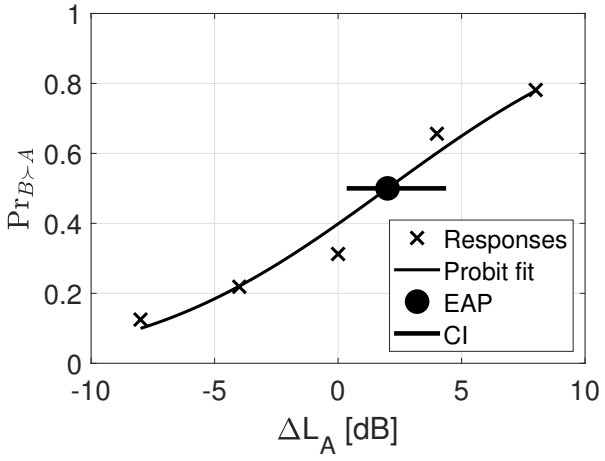


Fig. 10: Probability that the unsteady AS350 short stimulus is more annoying than the steady AS350 short stimulus ($\Pr_{B\succ A}$). Both stimuli were $1\,\mathrm{s}$ in duration, centered around t_{max}. This comparison was repeated at t_1 and t_2 to determine perceptually adjusted A-weighted SPL time histories.

Table 6: Equal Annoyance point (EAP) with confidence intervals of the A-weighted SPL (dB) for the short stimuli comparisons at three control points, found from Monte Carlo simulations. Adjustments are to make sound B equally annoying to sound A for pairs 1 (baseline/SEL-optimzed) and 4 (steady/unsteady).

	Pair 1		Pair 4	
Control point	EAP (ΔSPL)	CI	EAP (ΔSPL)	CI
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]

for more details). The resulting EAPs and confidence intervals from these MCMC simulations are summarized in Table 6. For pair 4, there was a significant perceptual difference between the steady/unsteady stimuli at t_{max}. However, for t_1 and t_2 for pair 4 as well as all three points for pair 1, the annoyance responses were not significantly different from what the A-weighted SPL indicated.

By comparing Tables 5 and 6, the confidence intervals for full flyovers in terms of SEL are larger than the confidence intervals for shorter stimuli in terms of A-weighted SPL. For the baseline/SEL-optimized

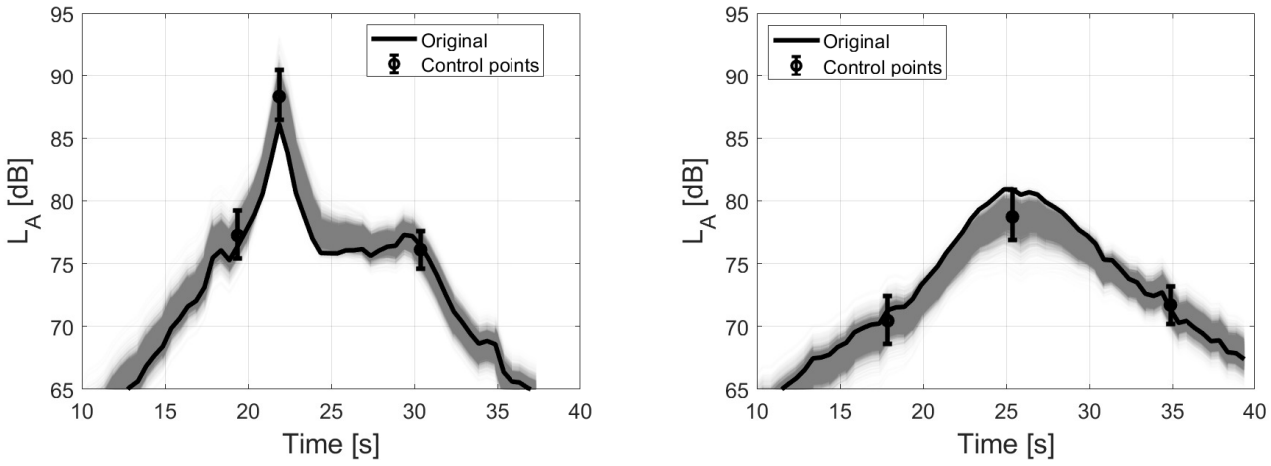
flyover pair, the confidence intervals are approximately $3.0\,\mathrm{dB}$ wider than the confidence intervals for the shorter stimuli, suggesting SEL does not fully capture the range of responses provided by test subjects. This is perhaps because inter-subject differences are larger when considering longer test stimuli. Comparing longer stimuli is likely a more difficult perceptual task, resulting in more variation in perception when comparing stimuli longer than $1\,\mathrm{s}$. It highlights one of the main drawbacks of a time-integrated metric like SEL, which is that human test subjects not only respond differently to each short portion of a flyover but their response to the entire flyover may be caused by different portions of the flyover, leading to a larger variation in the annoyance response.

The MCMC simulations for the short stimuli comparisons gave 100,000 resamples, which are considered bootstrapping samples (see Appendix) and used to adjust the A-weighted SPL time history. The adjustments for the unsteady vs. steady AS350 recorded flights (pair 4) are shown in Figure 11. Each adjusted A-weighted SPL time history is shown as a light gray curve; many overlapping curves yield a shaded region of possible adjustments. The mean A-weighted SPL adjustment and confidence interval for the control points comparisons are shown in Table 6.

In Figure 11, the stimuli from t_1 and t_2 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 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 A-weighted SPL time history of the unsteady flight to higher values. Reciprocally, most of the A-weighted SPL time history of the steady flight is adjusted lower in order to match the responses of the unsteady flight.

Similar to pair 1, the distributions for pair 4 using the TPPAS approach are narrower than the distribution on the EAP found through MCMC simulations for the full flyover comparison, as shown in Fig. 12. Again, this indicates that it is a simpler task for subjects to give responses to shorter sounds.

The distributions for pair 4 shown in Fig. 12 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 5), which means 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,



(a) Adjustment of the unsteady AS350 flight **(b) Adjustment of the steady AS350 flight**

Fig. 11: Perceptually-adjusted A-weighted SPL time histories for pair 4 using the TPPAS approach. (Actual playback levels were lower. See Section Test Design and Execution.) "Original" indicates the A-weighted SPL time history used in the full flyover comparison. Control points are at the maximum (t_{max}) and $10\,\mathrm{dB}$ down points (t_1 and t_2). Thin curves are the adjusted A-weighted SPL time histories, which are interpolated from the distributions found at the control points.

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 represent an interesting result that gives further insight into the efficacy of SEL. In particular, it is evident that neither Eq. (4) nor the TPPAS approach takes into account the fact that t_{max}-t_1 is only around $2.5\,\mathrm{s}$ for the unsteady flight, an example of the onset rate being a factor of annoyance (Ref. 31). The fast rise in A-weighted SPL 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 $10\,\mathrm{dB}$ difference in $2.5\,\mathrm{s}$ does not appear in the calculation for SEL. The analyses from probit fits and MCMC simulations should be considered more robust, since they include subjects' responses to the entire flyover, including the fast rise in A-weighted SPL.

The TPPAS distribution of the EAP when adjusting the A-weighted SPL time history for the steady

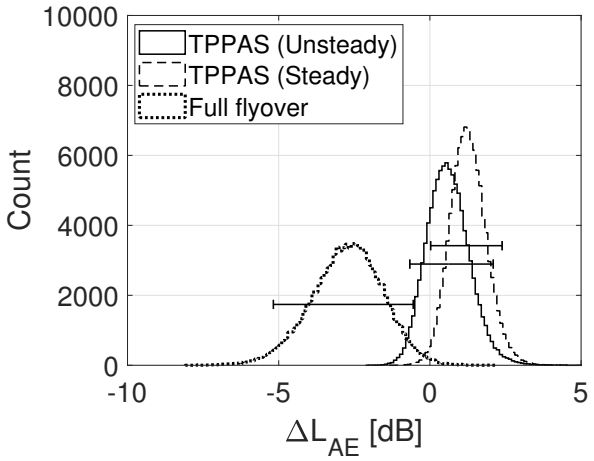


Fig. 12: Distributions of bootstrapped SEL values resulting from the TPPAS method for the unsteady/steady AS350 flights (pair 4), compared to the distribution of EAP for full flyovers.

flyover is slightly different from the distribution when adjusting the unsteady flyover, as already mentioned. It indicates that there are important perceptual differences at instances other than t_1, t_{max} and t_2, which may limit the TPPAS approach to sounds where the noise characteristics do not change very much throughout the flyover. The same limitation applies to SEL using Eq. (1) alone, since perceptual differences at specific time instances are also absent from this calculation. Furthermore, it indicates that sound quality cues, such as spectral, temporal or spatial noise characteristics not taken into account by the integrated A-weighted SPL time history, are important for the annoyance perception of helicopter noise flyovers.

Discussion

The results found from this psychoacoustic test have shown that there are statistically significant perceptual aspects of helicopter noise that are not taken into account by SEL, which may include spectral, temporal or spatial noise characteristics. While the main goal of this work is not to identify a better metric or to modify SEL, further work is needed to better predict the annoyance response to helicopter noise.

In particular, sound quality metrics may play an important role. For example, while the baseline flyover had higher impulsiveness than the SEL-optimized flyover in pair 1, the SEL-optimized flyover

had higher tonality in the range of $200 - 300\,\mathrm{Hz}$ and higher roughness. In pair 3, the EC130 low noise approach had higher tonality than the AS350 low noise approach around $1\,\mathrm{kHz}$, which may have induced its higher annoyance response. In pair 4, the unsteady AS350 had a quick rise to its maximum loudness (evident from the A-weighted SPL time history), which was also characterized by high impulsiveness. This suggests that sound quality metrics may play an important role in predicting the annoyance response to helicopter noise. All these sound quality metrics can be combined into a psychoacoustic annoyance model, which should be the subject of further research. A complicating factor related to that is what metric over the flyover to use: a 5% exceedance value, maximum value or some other integrated measure. The above sound quality analysis was done in HEAD Acoustics ArtemiS Suite v11.7, and the Hearing Model was used for tonality and impulsiveness.

The stimuli used in this test include three auralizations and four recordings from a flight test. Psychoacoustic testing that covers a larger range of perceptual attributes may be needed to better understand the aspects that are perceptually relevant to helicopter noise and that are not taken into account by an A-weighted integrated metric like SEL.

The results from pair 4 indicate that noise characteristics that occur at a particular point or throughout a flyover may lead to higher annoyance. This poses a challenge when designing psychoacoustic studies. While using recordings in perceptual studies adds natural variations in noise characteristics, it is difficult, costly and often impractical to control these characteristics during flight testing, putting some constraints on the experimental design of psychoacoustic tests. Auralizations based on noise predictions offer precise control of aerodynamic loads but are computationally expensive when specifying time varying loads throughout a flyover. This highlights the need for more efficient noise prediction tools and the ability to auralize those predictions, so that variation of noise characteristics during a flyover, including transitions in flight, can be carefully controlled and studied.

Conclusions

The main results of a psychoacoustic test related to the efficacy of SEL in the rating of annoyance to helicopter noise are presented. Recorded flights of an AS350 and EC130 as well as auralizations based

on optimized rotor designs for an AS350 were judged by human subjects in the Exterior Effects Room at the NASA Langley Research Center. The responses collected demonstrate significant qualitative aspects of helicopter noise not captured by SEL that are on the order of 3-4 dB. Further studies, most likely with greater resolution, are required to determine which metrics can account for this difference.

When a recorded low noise approach was played at the same SEL, the EC130 was judged to be more annoying than the AS350. This result was found by comparing annoyance responses at different relative levels, and it also holds for the absolute levels as recorded in the flight test. This suggests that SEL may not be a sufficient indicator of annoyance when comparing helicopters with differing tail rotor technologies. In this case, temporal, spectral or spatial components of the sound that are not captured by SEL may be important for annoyance evaluations.

When comparing different recorded maneuvers for the AS350, the flyover having unsteady sound characteristics (shorter duration, high impulsiveness) was perceived to be more annoying than one having mostly steady sound characteristics (longer duration, low impulsiveness). Again, Sound Exposure Level did not fully capture annoyance responses, this time comparing different maneuvers for the same helicopter. Besides the variations in sound quality, the unsteady flyover also had a steep rise in A-weighted SPL at the beginning of the sound stimulus.

In addition to comparing recordings of helicopter maneuvers, the psychoacoustic test also used auralized sounds based on different rotor designs in order to evaluate SEL as an indicator of annoyance. To within the resolution of the test, a flyover with a rotor design optimized in terms of SEL was not judged to be significantly different than the baseline flyover when played at the same SEL, which means that SEL can be an effective metric to minimize in the design phase of low noise rotors. It was found that optimizing rotor designs in terms of SEL or EPNL does not give significant differences in the perception of annoyance, indicating that reducing either metric will lead to a similar reduction in annoyance. However, these results should be taken with caution, because tail rotor noise was more prominent in the SEL-optimized flyover, which could have increased the annoyance responses. Therefore, it is suggested that both main and tail rotors should be optimized so that: (1) unwanted changes in sound quality are not introduced and (2) designed reductions in SEL are well-separated from the Equal Annoyance Point. The optimized rotor resulted in an auralized flyover within 10 dB of its maximum that was around 50%

longer than the baseline auralization, which also could have contributed to the slightly higher annoyance responses. Further study that systematically varies the duration of a flyover event would be necessary to determine if this represents a deficiency in using SEL as an indicator of annoyance to helicopter noise.

An alternative to presenting subjects with complete flyovers was tested in which subjects compared short sounds at the beginning, maximum A-weighted SPL and end of a flyover. The two methods gave similar results for the auralizations but differed for the unsteady and steady AS350 stimuli recorded from a flight test. This reinforces the idea that perceptually significant features of a sound that change over time, an aspect not taken into account by SEL or the TPPAS approach, can be important for annoyance judgments.

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

Appendix

This section is meant to give more insight into the Markov Chain Monte Carlo (MCMC) simulations and three-point perceptually adjusted SEL (TPPAS) technique used in this work. In particular, an example of how an MCMC simulation is initiated and how it converges is shown, and the bootstrapping technique used in the TPPAS approach is described in greater detail.

Markov Chain Monte Carlo simulations

The MCMC method iteratively varies the parameters in Eq. (5) and finds distributions of the most likely values that fit the paired comparison annoyance response data. The simulations start with an estimate of the EAP, μ, and the standard deviation, σ, of the distribution, Φ. For each response, if sound B is judged more annoying than sound A, the likelihood of that response is given by Φ in Eq. (5), assuming the current estimates of μ and σ. Otherwise, if sound A is judged more annoying, the likelihood is given by $1 - \Phi$. Since ΔL_{AE} is the SEL of sound B minus the SEL of sound A, it is more likely that sound B is judged more annoying for higher levels of sound B. The total likelihood for a given combination of μ and σ is the product of the likelihood of all the individual paired comparison annoyance responses (i.e., over all test subjects, all presentation levels, both A-B and B-A orderings, for a given signal pair).

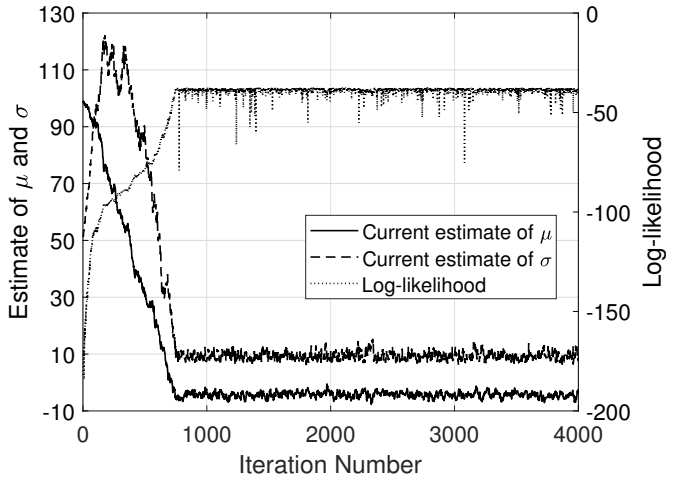
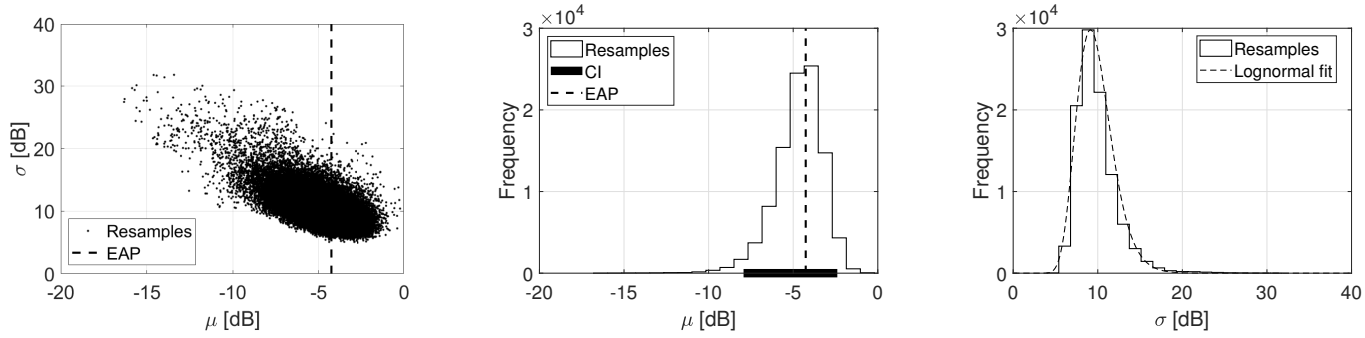


Fig. 13: Example of the first 4,000 iterations in a Markov Chain Monte Carlo simulation for determining the confidence interval of the Equal Annoyance Point (μ) for pair 3. The log-likelihood measures how well the current estimates of μ and σ fit the annoyance response data.

When moving to the next iteration of the chain, new values for μ and σ are randomly generated, and the total likelihood is again calculated. The change to the new values is accepted if the total likelihood is greater at the new step. If the total likelihood is not greater at the new step, the new μ and σ values are only accepted if the ratio of the new to old total likelihoods is greater than a random number generated between 0 and 1. Otherwise, the old μ and σ values are retained.

An example of the first 4,000 iterations of this process is shown in Fig. 13. The initial guess is at $\mu = 100$ and $\sigma = 50$. Since these values do not explain the annoyance response data well, this results in a low value for the log of the total likelihood. During the first few hundred iterations, combinations of μ and σ that are more likely are accepted until the estimates for the current iteration stay close to the highest likelihood values. Since unlikely values for μ and σ have a non-zero chance of being accepted, the log-likelihood occasionally and temporarily dips down.

An initial "burn-in" phase of 1,000 steps was used in this work. Fig. 13 shows that after 1,000 steps, the algorithm does not wander far from the best estimate of the EAP. After that, 100,000 steps of the algorithm produced a random walk around the most likely combinations of μ and σ (i.e., when Φ most closely matches the response data). The burn-in steps are not used in the analysis; all of the last 100,000



(a) The resamples from MCMC simulation and EAP.

(b) Histogram of resamples for μ and confidence interval of EAP.

(c) Histogram of resamples for σ and lognormal fit.

Fig. 14: An MCMC simulation gives the most likely values of μ that fits the response data for pair 3 (a). The histogram (i.e., marginalization) for μ gives an estimate of the confidence interval on EAP (b). A lognormal fit on the marginalization of results on σ is used as a prior for MCMC simulations for pairs 1 and 4 (c).

steps are the resamples used to calculate the confidence interval on the EAP, which are shown in Fig. 14a. The resamples tend to accumulate around the EAP, which is shown more clearly in the histogram for μ in Fig. 14b. 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 Fig. 14b as the thick, solid line.

The histogram of the likely values of σ for pair 3 is shown in Fig. 14c, which closely follows a log-normal 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 (in a Bayesian context) in the MCMC simulations for pairs 1 and 4. This prior is multiplied by the total likelihood at each iteration of the MCMC simulation and leads to accurate predictions of the CI for the EAP.

The acceptance rate is the total steps accepted divided by the total steps and should be close to 0.5 for data being fit by functions of two parameters (Ref. 32). All MCMC simulations in this work resulted in acceptance rates between 0.45 and 0.55, indicating a good convergence. Ten successive MCMC simula-

tions for the pair 3 full flyovers resulted in a standard deviation of $0.15\,\mathrm{dB}$ when calculating the width of the confidence interval, indicating good repeatability of the MCMC approach.

The text that was the primary source for the development of this approach is Kruschke (Ref. 33). The initial usage of this approach from this research group can be found in Ref. 29, where more details are given.

Three-point perceptually adjusted SEL (TPPAS)

The three-point perceptually adjusted SEL (TPPAS) approach is an alternative way to calculate SEL, after making perceptual adjustments to the A-weighted SPL time history at the beginning, end and maximum (t_1, t_2 and t_{max}, respectively) of a complete flyover. This section steps through the TPPAS approach using the auralized flyovers from pair 1. Discussion of the results and their importance are left for the main body when discussing the TPPAS results for pair 4.

The TPPAS approach consists of the following steps:

1) Collecting subjects' annoyance responses to short sounds ($1\,\mathrm{s}$ in duration) centered around t_1, t_{max} and t_2 at different relative levels of L_A

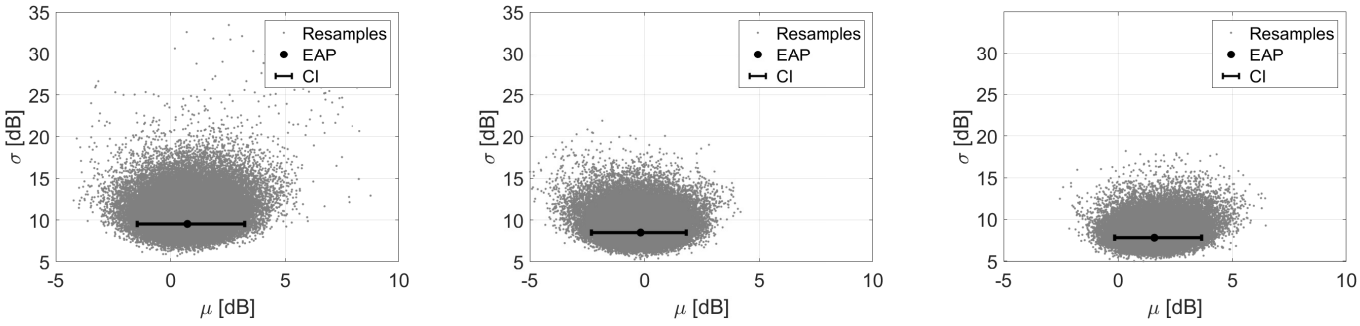
2) Do an MCMC simulation for each set of responses for t_1, t_{max} and t_2

3) Adjust the A-weighted SPL using the MCMC resamples

4) Calculate the perceptually adjusted SEL using the adjusted A-weighted SPL time histories

Finding TPPAS for a given pair of full flyovers starts with comparing short segments of the flyovers when the A-weighted SPL is at its maximum (t_{max}) and when it is $10\,\mathrm{dB}$ lower than the maximum (t_1 and t_2). Let t_{1A} be the first time during the flyover of sound A that the A-weighted SPL is within $10\,\mathrm{dB}$ of its maximum, and let t_{1B} be the first time during the flyover of sound B that the A-weighted SPL is within $10\,\mathrm{dB}$ of its maximum. Subjects' annoyance of sound A at t_{1A} compared to sound B at t_{1B} are collected at relative level differences between t_{1A} and t_{1B}. Then an MCMC simulation gives a distribution of 100,000 resamples of the most likely adjustment, ΔL_A, needed to make the two sounds at t_1 equally annoying.

The resamples at t_1, t_{max} and t_2 for the paired comparisons for pair 1 are shown in Fig. 15. Each blue



(a) Resamples for t_1 **(b) Resamples for t_{max}** **(c) Resamples for t_2**

Fig. 15: Resamples from MCMC simulations for paired comparisons of the short stimuli at t_1 (a), t_{max} (b) and t_2 (c) for pair 1. The original A-weighted SPL is at $\mu = 0$. Resamples are used as the bootstrapped samples when adjusting the A-weighted SPL time history.

dot represents one step in the MCMC simulation. The horizontal error bars show the confidence interval on the EAP for the short stimuli comparisons. All three overlap with $0\,\mathrm{dB}$, indicating that there were not strong perceptual differences in annoyance for the comparisons at these samples taken from the full flyover. However, the lower bound of the EAP confidence interval at t_2 is very close to zero, indicating that the SEL-optimized and baseline stimuli at t_2 were equally annoying when the SEL-optimized stimulus was $1.8\,\mathrm{dB}$ higher than the baseline. At this instance, the SEL-optimized stimulus was slightly less annoying than the baseline when played at the same L_A.

Since the MCMC resamples are a random sample of the distribution of likely adjustments, they are considered bootstrapping samples and used to adjust the A-weighted SPL time history of the full flyover. An example of this is shown in Fig. 16 for one resample for pair 1. For $t \leq t_1$, ΔL_A for t_1 is added to the original A-weighted SPL. For $t_1 < t \leq t_{max}$, a linear interpolation between ΔL_A at t_1 and t_{max} is added to the original A-weighted SPL. For $t_{max} < t \leq t_2$, a linear interpolation between ΔL_A at t_{max} and t_2 is added to the original A-weighted SPL. For $t > t_2$, ΔL_A at t_2 is added to the original A-weighted SPL. In Fig. 16, the adjustments are made to the A-weighted SPL of the SEL-optimized flyover. The confidence intervals on the resamples from Fig. 15 are shown again in Fig. 16 as vertical error bars at t_1, t_{max} and t_2. The blue dots are one MCMC resample at each control point. The dotted line is the adjusted A-weighted

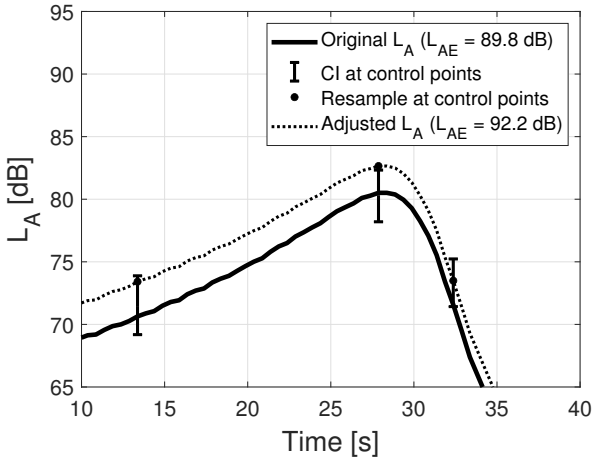
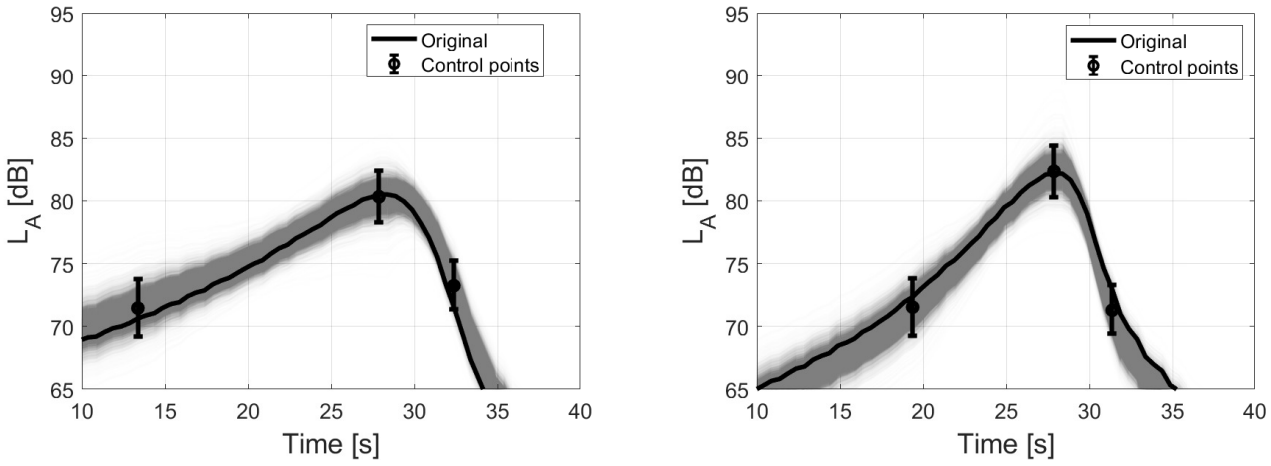


Fig. 16: Example of the TPPAS method for one resample at each control point t_1, t_{max} and t_2. The A-weighted SPL time history is interpolated between the control points, resulting in an adjusted SEL of $92.2\,\mathrm{dB}$.

SPL time history based on the three resamples, which gives an adjusted SEL that is $1.4\,\mathrm{dB}$ higher than the original.

Repeating the adjustments for all 100,000 resamples of the MCMC simulations give a range of possible adjusted A-weighted SPL time histories. Each adjusted A-weighted SPL time history is shown in Fig. 17a as a light gray curve; many overlapping curves yield a shaded region of possible adjustments. 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 A-weighted SPL. As a result, the adjusted A-weighted SPL time histories significantly overlap with the originals. Since the resamples give the relative adjustments between sound A and B, the opposite adjustments can be made to the A-weighted SPL time history of the baseline flyover as well, which is shown in Fig. 17b. Each thin curve gives a possible EAP for the full flyover, resulting in the distributions shown in Fig. 18.

Although all three confidence intervals in Fig. 18 overlap with 0, there are some important differences. First, 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 sim-



(a) Adjustment of the SEL-optimized flyover **(b) Adjustment of the baseline flyover**

Fig. 17: Perceptually-adjusted A-weighted SPL time histories for pair 1 using the TPPAS approach. (Actual playback levels were lower. See Section Test Design and Execution.) "Original" indicates the A-weighted SPL time history used in the full flyover comparison. Control points are at the maximum (t_{max}) and $10\,\mathrm{dB}$ down points (t_1 and t_2). Thin curves are the adjusted A-weighted SPL time histories, which are interpolated from the distributions found at the control points.

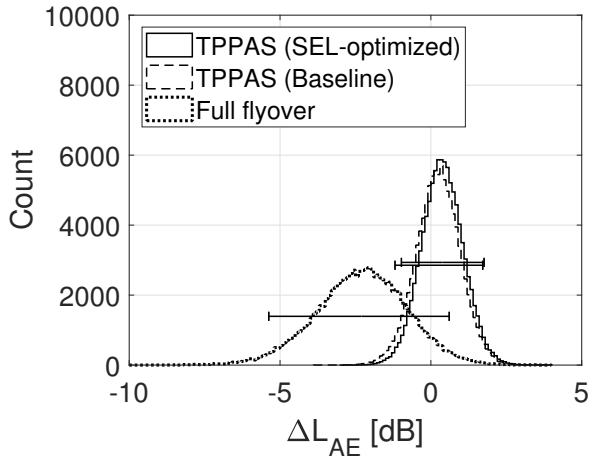


Fig. 18: Distributions of bootstrapped SEL values resulting from the TPPAS method for the SEL-optimized flyover case (pair 1), compared to the distribution of EAP for full flyovers.

pler tasks produces a more well-controlled response from subjects. On the other hand, responding to an entire flyover, subjects may respond to different parts of the flyover and give more varied responses. A

second difference is that the distribution of the full flyover tends to be negative while the TPPAS results are centered closer to 0, indicating slightly different EAPs between the two methods.

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]U.S. Federal Government. Code of Federal Regulations, Title 14, Part 36. Washington, DC: Federal Register, 2017.

[2]International Civil Aviation Organization. Annex 16 to the convention on international civil aviation environmental protection volume i: Aircraft noise. 7th Ed., Technical report, ICAO, 2014.

[3]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, 2017.

[4]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.

[5]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.

[6]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.

[7]S. Krishnamurthy, M. Boucher, A. Christian, and S.A. Rizzi. Rotorcraft Sound Quality Metric Test 1: Stimuli Generation and Supplemental Analyses. NASA Technical Memorandum 20205008997, 2021.

[8]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*, 153(2):867–876, 2023.

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

[10]M. Kleiner, B.-I. Dalenbäch, and P. Svensson. Auralization - an overview. *J. Audio Eng. Soc.*, 41(11):861–875, 1993.

[11]S.A. Rizzi, A.K. Sahai. Auralization of air vehicle noise for community noise assessment. *CEAS Aeronautical Journal*, 10(1):313–334, 2019.

[12]S. Krishnamurthy, S.A. Rizzi, D.D. Boyd, and A.R. Aumann. Auralization of rotorcraft periodic flyover noise from design predictions. Paper 74-2018-0166, Proceedings of the 74th Annual Forum of the American Helicopter Society, Phoenix, AZ, 2018.

[13]H. Fastl and E. Zwicker. *Psychoacoustics: Facts and Models*. 3rd Ed., Springer, 2007.

[14]E. Greenwood. Helicopter flight procedures for community noise reduction. Paper 73-2017-0016 , Proceedings of the 73rd Annual Forum of the American Helicopter Society, Fort Worth, TX, 2017.

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

[16]C.M. Heath, J.S. Gray. OpenMDAO: Framework for Flexible Multidisciplinary Design, Analysis and Optimization Methods. Paper AIAA 2012-1673, Proceedings of the 53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference, AIAA-2012-1673, Honolulu, HI, 2012, 1-13.

[17]H. H. Hubbard (Ed.). Aeroacoustics of Flight Vehicles: Theory and Practice Reference Publication RP-1258, NASA, August 1991.

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

[19]S. Krishnamurthy, B.C. Tuttle, and S.A. Rizzi. Auralization of unsteady rotor noise using a solution to the Ffowcs Williams-Hawkings equation. Paper F-0075-2019-14440, Proceedings of the 75th Annual Forum of the Vertical Flight Society, Philadelphia, 2019.

[20]S. Krishnamurthy, B.C. Tuttle, and S.A. Rizzi. A synthesis plug-in for steady and unsteady loading and thickness noise auralization. Paper AIAA 2020-2597, Proceedings of AIAA Aviation, Virtual, 2020.

[21]A.R. Aumann, B.C. Tuttle, W.L. Chapin, and S.A. Rizzi. The NASA auralization framework and plugin architecture. Paper IN15-298, Proceedings of InterNoise15, San Francisco, CA, 2015.

[22]NASA Langley Research Center. Aircraft Flyover Simulation. Sructural Acoustics Branch. http://stabservdata.larc.nasa.gov/flyover/VFS76/VFS-Forum76-2020-SoundFiles.zip, 2022.

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

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

[25]W.A. Simpson The method of constant stimuli is efficient. *Perception & Psychophysics*, 44(5): 433-436, 1988.

[26]L.L. Beranek. Balanced noise-criterion (NCB) curves. *The Journal of the Acoustical Society of America*, 86(2):650–664, 1989.

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

[28]A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007.

[29]S.A. Rizzi and A. Christian. A psychoacoustic evaluation of noise signatures from advanced civil transport aircraft. Paper AIAA 2016-2907, Proceedings of the 22nd AIAA/CEAS Aeroacoustics Conference, 2016–2907, Lyon, France, 2016.

[30]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.

[31]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.

[32]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.

[33]J.K. Kruschke. *Doing Bayesian Data Analysis*. Elsevier LTD, Oxford, 2015.

List of Figures

List of Tables