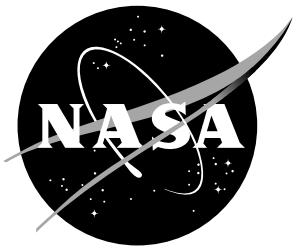
NASA/TM–20240003202



Toward a Psychoacoustic Annoyance Model for Urban Air Mobility Vehicle Noise

Matthew A. Boucher
Langley Research Center, Hampton, Virginia

Andrew W. Christian
Langley Research Center, Hampton, Virginia

Siddhartha Krishnamurthy
Langley Research Center, Hampton, Virginia

Tyler Tracy
Langley Research Center, Hampton, Virginia

Durand R. Begault
Ames Research Center, Moffett Field, California

Kevin Shepherd
Langley Research Center, Hampton, Virginia

Stephen A. Rizzi
Langley Research Center, Hampton, Virginia

June 2024

NASA STI Program Report Series

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA scientific and technical information (STI) program plays a key part in helping NASA maintain this important role.

The NASA STI program operates under the auspices of the Agency Chief Information Officer. It collects, organizes, provides for archiving, and disseminates NASA's STI. The NASA STI program provides access to the NTRS Registered and its public interface, the NASA Technical Reports Server, thus providing one of the largest collections of aeronautical and space science STI in the world. Results are published in both non-NASA channels and by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA Programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counterpart of peer-reviewed formal professional papers but has less stringent limitations on manuscript length and extent of graphic presentations.

- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.

- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.

- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.

- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.

- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services also include organizing and publishing research results, distributing specialized research announcements and feeds, providing information desk and personal search support, and enabling data exchange services.

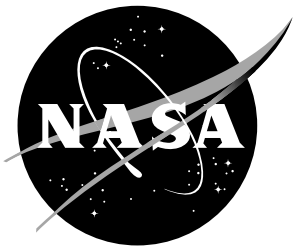
For more information about the NASA STI program, see the following:

- Access the NASA STI program home page at http://www.sti.nasa.gov

- Help desk contact information:

https://www.sti.nasa.gov/sti-contact-form/
and select the "General" help request type.

NASA/TM–20240003202



Toward a Psychoacoustic Annoyance Model for Urban Air Mobility Vehicle Noise

Matthew A. Boucher
Langley Research Center, Hampton, Virginia

Andrew W. Christian
Langley Research Center, Hampton, Virginia

Siddhartha Krishnamurthy
Langley Research Center, Hampton, Virginia

Tyler Tracy
Langley Research Center, Hampton, Virginia

Durand R. Begault
Ames Research Center, Moffett Field, California

Kevin Shepherd
Langley Research Center, Hampton, Virginia

Stephen A. Rizzi
Langley Research Center, Hampton, Virginia

National Aeronautics and
Space Administration

Langley Research Center
Hampton, Virginia 23681-2199

June 2024

Available from:

NASA STI Program / Mail Stop 050
NASA Langley Research Center
Hampton, VA 23681-2199

Abstract

A psychoacoustic test was performed to obtain annoyance responses to noise from a quadrotor Urban Air Mobility (UAM) vehicle to aid in the development of a model of annoyance to UAM vehicle noise. Previous analysis of that test concluded that a psychoacoustic annoyance (PA) model, including the effects of loudness, sharpness, fluctuation strength and roughness, correlated well with the collected annoyance responses. This motivated (1) the assessment of other PA models available in the literature and (2) the development of a new PA model that includes other sound quality effects (e.g., tonality). To build the PA model for UAM noise, annoyance ratings to individual sounds and annoyance comparisons between pairs of sounds are first combined into a latent annoyance scale that has a correlation coefficient of 0.98 with the raw responses and that is based on just-noticeable-differences (JNDs) in annoyance. This latent annoyance (JND) scale also gives insight into the interplay of various perceptual components, such as overall loudness, temporal effects and spectral effects. The proposed PA model for UAM noise is fit to the latent annoyance scale, includes the sound quality effects mentioned above with an added term for tonality, and offers an improvement over other methods for predicting annoyance to UAM vehicle noise.

1 Introduction

The goal of Urban Air Mobility (UAM) is to provide safe, efficient and accessible on-demand transportation services for passengers and cargo [1]. Early testing and development are underway in New York, São Paulo, New Zealand and Dubai, but vehicle-related challenges related to noise annoyance have already surfaced [2]. In order for UAM operations to mature and become practical in a wide range of locations, it is critical that the noise impact be minimized for communities in which UAM vehicles operate [3, 4]. To address some of these challenges related to human perception of UAM noise, gaps and recommendations were established by a working group on UAM noise [5]. Examples include further development of noise metrics and the need for predictive models of human response, including annoyance, which can be used in perception-influenced design of UAM vehicles [6].

Following these recommendations, annoyance responses to UAM noise by human test subjects were collected as part of the Test of UAM Sound Quality (TUSQ), which took place in 2022 [7]. Due to the scarcity of recorded UAM vehicle noise, that psychoacoustic test was based on noise predictions and auralizations of a NASA reference quadrotor UAM vehicle [8]. The test explored the relationship between objective measures of sound quality and subjective annoyance responses. In particular, sharpness, tonality, roughness and impulsiveness were found to be contributing factors in the annoyance response.

In the TUSQ dataset, a linear regression model showed that the tonality and

roughness metrics were similar in importance[1]. Additional analysis of the TUSQ data demonstrated that the PA model by Zwicker [9], a non-linear combination of sound quality metrics that includes roughness but not tonality, correlated more highly with annoyance than the linear regression model. These observations provided motivation to investigate the applicability of other published PA models to UAM noise and to incorporate tonality into a PA model for UAM noise. Zwicker's model includes the effects of loudness, sharpness, fluctuation strength and roughness. The models by More [10] and Di et al. [11] include an additional term for tonality, and the model by Torija et al. includes additional terms for both tonality and impulsiveness [12]. Tonality and impulsiveness are expected to be important for UAM noise because of the acoustical characteristics of rotors. A "chopping" or impulsive sound can occur when tip vortices from preceding blades interact with the following blades, creating blade-vortex interaction noise [13]. The perception of tonality may be caused by distinct harmonics of the blade passage frequency.

The main contributions of the current work are: (1) generating a latent annoyance scale built on TUSQ annoyance responses and just-noticeable-differences in annoyance, (2) assessment of available PA models from the literature to UAM noise and (3) a new PA model developed specifically for UAM noise. The hypothesis is that a model based on unique features of UAM vehicle noise will yield better annoyance predictions than a model developed from non-UAM noise sources. The latent annoyance scale is based on Thurstone's Law of Comparative Judgment [14], unifies both annoyance ratings and paired comparisons [15] and highly correlates with TUSQ annoyance responses. Because of this high correlation, assessment of PA models is evaluated in terms of the line of best fit between PA predictions and latent annoyance.

This work is organized as follows. Section 2 summarizes the methods and conclusions from the TUSQ [7]. Section 3 describes Thurstone's Law and the latent annoyance scale used to quantify annoyance responses collected from the TUSQ. Section 4 describes the forms of various psychoacoustic annoyance models available in the literature and evaluates their applicability to the TUSQ dataset. Finally, Section 5 presents the proposed PA model for UAM noise, which is based on Zwicker's model and includes a tonality term.

2 Psychoacoustic test

This section summarizes the TUSQ, which was administered in 2022 and documented in a publication at the SAE International Noise & Vibration Conference 2023 [7]. The main output from that work is a dataset that includes annoyance responses to UAM noise stimuli with a range of sound qualities. TUSQ test subjects provided annoyance ratings (from "not at all annoying" to "extremely annoying") of individual stimuli and annoyance comparisons between pairs of stimuli (i.e., paired comparisons). The TUSQ can be summarized in the following steps.

[1]The linear regression model mentioned here refers to a Least Absolute Shrinkage and Selection Operator (Lasso) regression model with sharpness, tonality, roughness and impulsiveness as predictors. [7].

2.1 UAM noise predictions

The acoustic stimuli were based on the NASA quadrotor reference vehicle, which is sized for six passengers and a 1200 lb. payload [8]. The blade geometry, number of blades and rotation rate were inputs to the Comprehensive Analytical Rotorcraft Model of Rotorcraft Aerodynamics and Dynamics (CAMRADII [16]) program. This resulted in a blade passage frequency of 20 Hz. For two operating conditions (90 knots with a 0° climb angle and 60 knots with a -5° climb angle), CAMRADII determined the corresponding blade loading, blade motion, inflow velocity and effective angle of attack. Source noise hemispheres were then calculated for each vehicle rotor using the Aircraft Noise Prediction Program 2 (ANOPP2 [17]). Within ANOPP2, Farassat's Formulation 1A was used for loading and thickness noise [18, 19], and self noise was calculated using an implementation of the Brooks-Pope-Marcolini model [20] cast in a rotating frame.

2.2 Auralizations and post-processing

Auralizations were generated from the source noise hemispheres using the NASA Auralization Framework (NAF) [21–23]. Both flight conditions were assumed to be stationary relative to a ground observer, because annoyance to time-varying sound quality was outside the scope of this psychoacoustic test. The auralizations assumed that the quadrotor was at an altitude of 1,000 ft with an emission angle of 60° elevation and 0° azimuth. A reference sound was used in TUSQ, which consisted of the self noise component of the level cruise auralization after removing its modulations. This reference sound was chosen due to its lack of dominant sound quality characteristics. Annoyance judgments of this sound at different levels would then be due to changes in loudness and not changes in other sound quality characteristics, such as roughness, tonality, etc. A total of 136 UAM stimuli (not including the reference sound) resulted from post-processing the auralizations, which involved: (1) mimicking changes in blade passage frequency, (2) adjusting the relative gain between loading and thickness noise and self noise components via a spectral weighting parameter, (3) adding a tone complex, (4) smoothing out the time signature of the loading and thickness noise via time averaging and (5) amplitude modulation. The post-processing was done to efficiently generate a range of sound quality instead of repeating the predictions described in Sec. 2.1 for a large number of BPFs and flight conditions. Further details can be found in Boucher et al. [7].

2.3 Experimental design

The psychoacoustic test was split into two parts, one that tested the annoyance to changes in loudness and another that tested annoyance to other aspects of sound quality. To test aspects of sound quality other than loudness, the post-processing steps described in Sec. 2.2 were applied to both flight conditions (level cruise and 5° descent) to create a full-factorial experimental design with 4 factors (sharpness, tonality, impulsiveness and fluctuation strength) and two qualitative levels for each factor (low and high). UAM stimuli were adjusted to a loudness of 6 sones by multiplying the acoustic pressure time history by a frequency independent amplitude fac-

tor. Roughness variations in the stimuli were not controlled, since a post-processing technique was not available to systematically adjust roughness independently of the other sound quality metrics. This resulted in 136 UAM stimuli of equal loudness that spanned a range of other sound quality values, as shown in Figure 1. To test annoyance to changes in loudness, the reference sound described earlier was presented at five different levels in 5 dB increments, resulting in loudnesses of 3.8, 5.6, 8.0, 11.3 and 15.7 sones.

Various methods are available to calculate sound quality metrics. The ones used in the TUSQ and the current work are given here. The loudness equalization process applied to the UAM stimuli followed ISO 532-1 [24] as implemented in the NAF Psychoacoustic Analysis Library. The resulting loudness was evaluated using the DIN 45631/A1 [25] standard in HEAD Acoustics ArtemiS Suite 13.6 [26]. Both methods are implementations of Zwicker's method for time-varying loudness. Sharpness, tonality, roughness, impulsiveness and fluctuation strength also used ArtemiS Suite 13.6. Sharpness used the DIN45692 (free field) standard [27]. Tonality, roughness, impulsiveness and fluctuation strength all used Sottek's Hearing Model [28, 29]. For tonality, the Hearing Model was preferred over Aures' method, because it considers many aspects of tonality (not just pure tones), including narrowband noise and impure tones [30]. The tonality and roughness calculation methods are included in the ECMA-418 standard for psychoacoustic metrics for noise emissions from Information Technology and Telecommunications equipment [31]. For a description of the perceptual effects of various sound quality metrics, see Section 4.

2.4 Collection of annoyance responses

The acoustic stimuli were reproduced in the Exterior Effects Room [32] at NASA Langley Research Center. All sounds were presented from a front, central location to 40 human test subjects (10 groups, 4 subjects at a time). Subjects listened attentively and gave their judgment of the sounds in terms of annoyance. As mentioned in Sec. 2.3, the psychoacoustic test was split into two parts. The first part tested annoyance ratings to UAM stimuli that had different sound qualities but equal loudness. As shown in Fig. 2, subjects were asked, "How annoying was the sound to you?" This rating scale spanned from just below "not at all" to just above "extremely", which corresponds to an 11-point scale from 1 to 11. Each subject gave their annoyance rating to all 136 UAM stimuli, including 8 sounds that had 4 replicates. This resulted in 6,400 rating responses in total. The second part of the test gauged subjects' annoyance via paired comparisons (see bottom screen prompt in Fig. 2). In each pair, one sound was a UAM stimulus at 6 sones, and the other sound was the reference sound at one of five levels, ranging from 3.8 to 15.7 sones (a 20 dB range). Twenty-six of the 136 UAM stimuli were chosen at random for each group of test subjects. This resulted in 119 out of the 136 UAM stimuli being included in the paired comparisons and a total of 1,040 (26×40) responses.

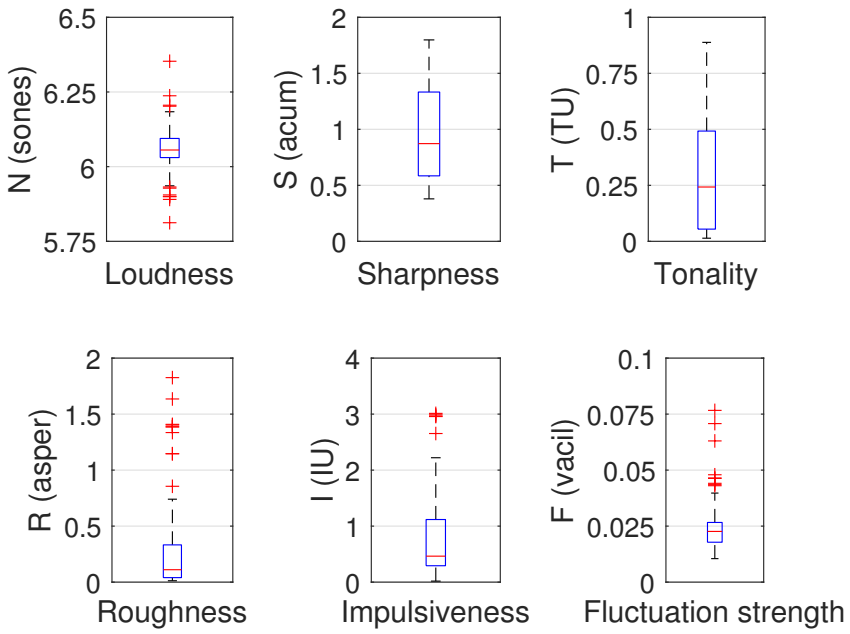


Figure 1. Box plots of sound quality of 136 UAM stimuli used in TUSQ. The reference sound, which only varied in loudness, was presented at five different levels, separated by 5 dB, which resulted in loudness of 3.8, 5.6, 8.0, 11.3 and 15.7 sones.

2.5 Observations

Previous work made the following conclusions from this psychoacoustic test, including those related to loudness, other sound qualities and psychoacoustic annoyance [7]:

- The point of subjective equality (i.e., equal annoyance point) between the UAM stimuli and the reference sound occurred when the reference sound was 9.3 sones and the UAM stimuli were an average of 6 sones (with all UAM stimuli within the range of 5.8 to 6.4 sones). This difference in loudness corresponds to approximately a 6.3 dB difference in sound pressure level. When the UAM stimuli and reference sound were compared at the same loudness, the UAM stimuli were judged more annoying 73% of the time. This indicates that differences in sound quality other than loudness can have a significant effect on annoyance.

- The annoyance rating responses to UAM stimuli showed that tonality, roughness and impulsiveness were positively correlated with annoyance. Sharpness was found to be negatively correlated with annoyance. The fact that higher sharpness did not result in higher annoyance is likely due to only 3 out of 136 UAM stimuli having sharpness above the 1.75 acum threshold to contribute to psychoacoustic annoyance [33]. However, the negative correlation is surpris-

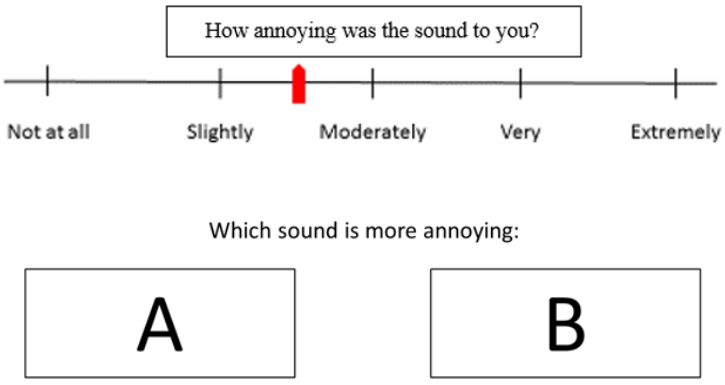


Figure 2. Computer tablet screen prompts were presented to test subjects to collect their annoyance responses. UAM stimuli of equal loudness were rated on an 11-point scale from "not at all" to "extremely" (top). In another part of the test, a reference sound at different levels was compared to UAM stimuli of equal loudness via paired comparisons (bottom).

ing and indicates that higher levels of broadband self-noise may mask some negative aspects of tonal, impulsive loading and thickness noise.

- Finally, the PA model by Zwicker [33–36] was found to have a higher correlation with annoyance than a linear regression model. This is one of the main motivating factors to assess the performance of other PA models for UAM noise (see Section 4) and to make potential improvements (see Section 5).

3 Latent annoyance scale

This section describes Thurstone's Law of Comparative Judgment [14] and how it relates to the subjective evaluation of annoyance during a paired comparison test. A method is then described to place each acoustic stimulus from the TUSQ on a linear psychological scale. This method combines annoyance rating responses with paired comparison data through a simple linear relationship. Thurstone's Law and the process of placing a sound stimulus on a psychological scale does not assume anything about the physical measurement of a sound. However, sound quality and A-weighted sound pressure level (SPL) spectra of example stimuli are examined to develop some intuition about what acoustical measures may be important for psychoacoustic annoyance. By applying Thurstone's Law, each stimulus from the TUSQ is given a latent annoyance value, which is used in Section 4 to assess the applicability of various PA models and in Section 5 to develop an improved model that includes tonality.

3.1 Thurstone's Law of Comparative Judgment

A fundamental concept of psychophysics, useful in the context of developing a PA model, is that the difference in annoyance between two stimuli in a paired comparison

is related to the proportion of comparisons for which one stimulus is judged more annoying than the other. When this proportion is further from 50%, the distance between the two sounds on the psychological (e.g., latent annoyance) scale is greater. The distance is called the discriminal difference, which is given by $q_i - q_j$ where q_i and q_j are the discriminal processes of sounds i and j, respectively. There is natural variation, or dispersion, in comparative judgments, meaning that a test subject, "gives different comparative judgments on successive occasions about the same pair of stimuli [14]." As a result, the discriminal process of the stimulus, as well as its location on the psychological scale, is actually a distribution. This dispersion is a measure of the reliability of the estimated location of a particular stimulus on the psychological scale. This behavior is depicted in Fig. 3.

These discriminal concepts as well as Thurstone's mathematical formulation of them are known as the Law of Comparative Judgment [14]. The concepts are applicable in a wide range of studies: when a human test subject provides a response about the quality of an image [37], in judging quality of handwriting [38] or the judgment of audio samples [39]. When applying the Law of Comparative Judgment to the TUSQ dataset, it is assumed that responses from multiple subjects can be combined such that the resulting distribution of responses is normal. It is also assumed that the discriminal dispersion, σ, is the same for all stimuli [2].

An important concept of Thurstone's Law of Comparative Judgment is that no assumption is made regarding physical measures of the stimuli. Although it is possible that physical measures contribute to the subject's decision, the method does not require it, which makes it applicable to comparative studies where physical measures of stimuli are not apparent or even absent. The method can, therefore, be applied to studies of opinions, affects, judgments of all types, even economics, morals and aesthetics [40]. Of course, for developing a PA model for UAM noise, the desire is that latent annoyance correlates with some measures of loudness, temporal effects and spectral effects. Since the latent annoyance scale described here is completely independent of any physical measure of sound, latent annoyance may be used to fit a PA model or another type of annoyance model with a different mathematical form or one that includes different physical measures of sound.

3.2 Latent annoyance scale for ratings and paired comparison data

Data collected during the TUSQ were: (1) annoyance ratings for UAM stimuli of equal loudness and (2) annoyance comparisons between UAM stimuli at equal loudness and a reference sound that varied in loudness. This was done so that changes in annoyance due to changes in sound qualities other than loudness could be evaluated independently from changes in loudness. However, to develop an annoyance model that accounts for all these effects, it is necessary to combine the rating and comparison responses into one common, latent annoyance scale.

When analyzing the latent scores on the psychological scale, modern methods use a probabilistic formulation for observing a given response based on the underlying discriminal processes, q, and discriminal dispersion, σ, which assumes a normal

[2] These assumptions, along with others, describe Case V of Thurstone's Law of Comparative Judgment. For a complete description of these assumptions, see Ref. [14].

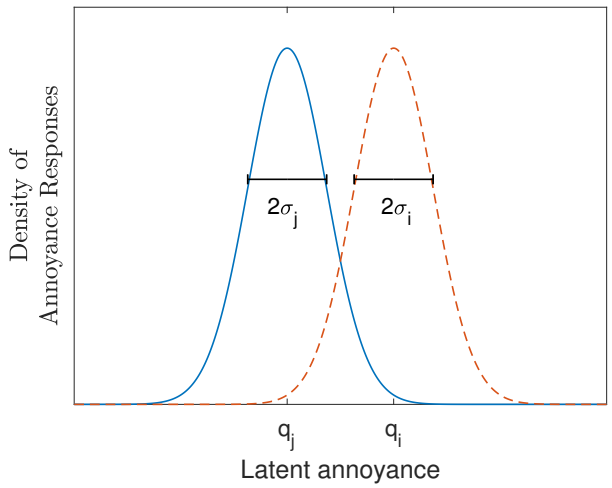


Figure 3. The discriminal processes for sounds i and j are assumed to be random variables with their means being their latent annoyance. The discriminal dispersions are assumed to be equal (i.e., $\sigma_i = \sigma_j$).

distribution for each latent annoyance, q_i, on the psychological scale. Although combining ratings with the latent annoyance from paired comparisons is not trivial, a simple linear relationship may be assumed between the ratings and latent annoyance, which is given by $m_i = a\, q_i + b$, where m_i is the annoyance rating. What follows is the probabilistic formulation used to obtain latent annoyance values for all TUSQ stimuli. The mathematical description and its development can be found in Perez-Ortiz et al. [15, 37], and the code is freely available [41].

The latent annoyance for all stimuli is found by obtaining the parameters q, σ, a, b and η that maximize the total probability of observing the annoyance response data contained in the paired comparison preference matrix $\mathbf{C}$ and the rating matrix $\mathbf{M}$. The maximization process is expressed as

$$\underset{q,a,b,\eta}{arg\,max}\; P(\mathbf{C}|q,\sigma) \cdot P(\mathbf{M}|q,\sigma,a,b,\eta) \tag{1}$$

where the arguments q, a, b and η are sought that maximize the total probability. The total probability is given by the scalar product of $P(\mathbf{C}|q,\sigma)$ and $P(\mathbf{M}|q,\sigma,a,b,\eta)$, which are the probabilities of observing the paired comparison and rating data, respectively. A prior in the Bayesian sense is also multiplied by this likelihood expression to enforce convexity and reduce the likelihood that any values of q are far from its mean [37]. The parameter η is introduced as a weighting parameter between rating and paired comparison response data [3].

The preference matrix $\mathbf{C}$ contains the observations collected during the paired comparison portion of the psychoacoustic test. The preference matrix is a square matrix with N rows and columns where $N = 141$ is the number of stimuli. Indexes

[3] The original publication of this method used c instead of η [37]. η is preferred here to distinguish it from the entries of the preference matrix c_{ij}.

1-136 are the UAM stimuli, and indexes 137-141 are the reference sound at different levels. In each entry c_{ij}, the i-th sound is compared to the j-th sound, and c_{ij} is the total number of times sound i was judged more annoying than sound j. None of the UAM stimuli were directly compared to each other, so $c_{ij} = 0$ for all entries in which $i < 137$ and $j < 137$. Each comparison contained one reference sound and one UAM stimulus, so c_{ij} may be non-zero for $i \geq 137$ or $j \geq 137$. The total number of comparisons made was 1,040, meaning $\underset{i,j}{\Sigma}\mathbf{C} = 1,040$.

The likelihood of observing pairwise comparisons is $P(\mathbf{C}|q,\sigma)$ and is given by

$$P(\mathbf{C}|\mathbf{q},\sigma) = \prod_{i,j} \binom{n_{ij}}{c_{ij}} \Phi(q_{ij},\sigma)^{c_{ij}} (1-\Phi(q_{ij},\sigma))^{n_{ij}-c_{ij}} \tag{2}$$

where Φ is the normal cumulative distribution function and n_{ij} is the total number of times sound i was compared to sound j. This assumes that for a given discriminal difference, q_{ij}, and discriminal dispersion, σ, the probability that sound i was judged more annoying than sound j is equal to: (1) $\Phi(q_{ij},\sigma)$ when the subject chooses sound i as more annoying and (2) $1-\Phi(q_{ij},\sigma)$ when the subject chooses sound j as more annoying.

The rating matrix $\mathbf{M}$ has a size of the number of stimuli ($N = 141$) by the number of subjects ($J = 40$). The entries, m_{ik}, are the ratings of the i-th sound by subject k and are in the range of 1-11. The matrix entries are non-zero for $i < 137$ and are populated with *nan* for stimuli 137-141, because the reference sound was not rated; it was only included in the paired comparison portion of the experiment. If a stimulus was presented more than once in the experiment, m_{ik} is the mean rating response of sound i for subject k.

The likelihood of observing rating scores is given by $P(\mathbf{M}|q,\sigma,a,b,\eta)$ and expressed as

$$P(\mathbf{M}|q,\sigma,a,b,\eta) = \prod_{i=1}^{N}\prod_{k=1}^{J} f(m_{ik}|q_i,\sigma,a,b,\eta) \tag{3}$$

where the Normal distribution, f, is given by

$$f(m_{ik}|q_i,a,b,\eta) = \frac{1}{\sqrt{2\pi a^2\eta^2\sigma^2}} e^{-\frac{(m_{ik}-(aq_i+b))^2}{2a^2\eta^2\sigma^2}} \quad . \tag{4}$$

The discriminal dispersion, σ, is the same as it is in Eq. (2). However, the weighting parameter, η, assumes that the effective dispersion on the rating scale (equal to $\eta\,\sigma$) may be different from the discriminal dispersion when only paired comparisons are considered. The latent annoyance vector, q, is found using a maximum likelihood estimator, which seeks to find q and parameters a, b and η that maximize the probability of explaining the collected data in the preference matrix $\mathbf{C}$ and the rating matrix $\mathbf{M}$.

The discriminal dispersion is set to a fixed value ($\sigma = 1.4826$) considering the just-noticeable-difference from Thurstone's analysis [14, 37]. Specifically, σ is specified so that when 75% of responses agree which sound is more annoying (i or j), then the discriminal difference, $q_i - q_j$, is equal to one just-noticeable-difference. Thurstone states that, "we may arbitrarily define the difference limen or the [JND]

as that stimulus difference which has a probability of 0.75 of being correctly discriminated [42]." Therefore, for 1 JND to correspond to a unit change in latent annoyance, it means that $\Phi(q_{ij}, \sigma) = 0.75$ when $q_{ij} = 1$. With these constraints, σ must be equal to 1.4826.

In a psychoacoustic experiment, the accuracy of ratings may be different from the accuracy of paired comparisons. In other words, the amount of information gained from a rating may be different from the amount gained from a paired comparison. The η parameter in the latent annoyance analysis is a way to measure this tradeoff. Eq. (4) shows that η is an extra parameter that modifies the discriminal dispersion for the rating responses. For $\eta > 1$, the dispersion for rating responses is greater than that for paired comparisons alone (i.e., $\eta\,\sigma > \sigma$). In this case, paired comparisons give a narrower distribution and, therefore, have less error than rating responses. The opposite is true for $\eta < 1$; rating responses give a narrower distribution and have less error than paired comparisons. Through the maximization process described in Eq. (1), the value of η is varied and found to be 1.284. This indicates that paired comparisons of the UAM stimuli resulted in less error in the annoyance judgments than ratings.

The maximization process expressed in Eq. (1) yields the latent annoyance vector, q, for all acoustic stimuli, including the UAM stimuli and the reference sounds. The corresponding q_i for the UAM stimuli are shown in Fig. 4 as red $\times$'s and are compared to the mean annoyance responses computed over all test subjects, shown in blue circles. The mean annoyance responses are on the 11-point scale described in Fig. 2. The abscissa is the stimulus number, sorted in increasing latent annoyance. As a result, mean annoyance is not monotonically increasing. The Pearson correlation coefficient between latent annoyance and mean annoyance responses is high, 0.98, for the 136 UAM stimuli. The black dots are the latent annoyance for the reference sound at five different loudness values, which do not have a mean annoyance response but can be placed on the latent annoyance scale among the UAM stimuli. The latent annoyance results vary from approximately -2 to 2, which spans 4 annoyance JNDs and corresponds roughly to the 20 dB range of the reference sound.

Because of the high correlation between mean annoyance responses and latent annoyance, as well as the previous linear assumption that $m_i = aq_i + b$, the latent annoyance in terms of JNDs can be converted to an approximate mean annoyance rating. The results from the Thurstone analysis yield the mean annoyance rating as $m = 1.44q + 5.06$. This relationship may not hold for stimuli whose mean annoyance is not normally distributed, such as near either extreme of the 11-point scale. To account for this, the analysis may be more robust if a skewed distribution for rating responses is used in Eq. (4). However, this is not expected to be an issue for the mean annoyance responses from the TUSQ dataset, since a Jacque-Bera test indicated that only 8 of 136 UAM stimuli had a non-normal distribution. Furthermore, these 8 stimuli covered a wide range of mean annoyance (i.e., 3.7, 4.1, 4.5, 4.5, 5.0, 5.8, 6.8, 7.8) and were not systematically near the boundaries of the scale.

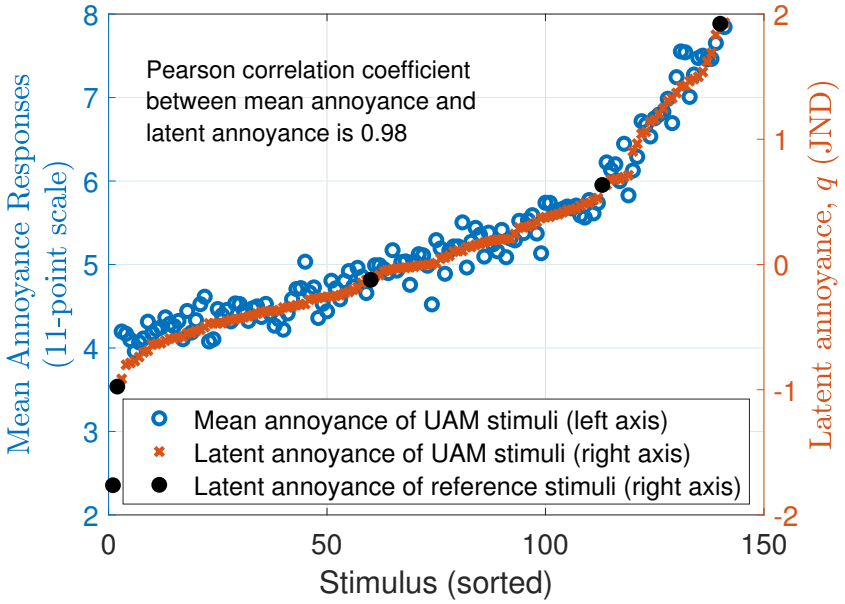


Figure 4. Correlation between mean annoyance responses and latent annoyance, q. Latent annoyance can be transformed to the 11-point rating scale by $1.44\,q + 5.06$. UAM stimuli are all approximately 6 sones yet span a range of 2.8 JNDs in annoyance. The reference sounds span ranges of 15.7 sones and 3.7 JNDs in annoyance.

3.3 Observations related to acoustical characteristics of stimuli

The latent annoyance values, q_i, shown in Fig. 4 are determined independently of the acoustical characteristics of the test stimuli. However, it can be useful to examine acoustical characteristics to understand what influenced the relative latent annoyance values. Indeed, it is assumed that subjects were sensitive to loudness, as well as temporal and spectral effects, since Zwicker's PA model was positively correlated with the mean annoyance responses [7]. Therefore, before evaluating psychoacoustic annoyance, some general observations about physical measures of sound, including loudness, other sound quality metrics and SPL spectra are made.

As shown in Fig. 4, the least annoying UAM sound (at 6 sones) is roughly equal in annoyance to the second quietest reference sound (at 5.6 sones). The most annoying UAM sound (also 6 sones) is about equally annoying as the loudest reference sound at 15.7 sones; this is about 3 JNDs more annoying than the least annoying UAM stimulus. Therefore, the range of annoyance to the UAM stimuli roughly corresponds to a change in loudness of the reference sound from 5.6 to 15.7 sones. This is approximately equivalent to a change of 15 dB. Two important observations can be made: (1) one annoyance JND corresponds to about a 5 dB change of the reference sound and (2) since all UAM stimuli were played at 6 sones, sound quality other than loudness can cause annoyance differences of about 3 JNDs or approximately a 15 dB change in the reference sound.

These observations are made more clear by looking at four example stimuli, as

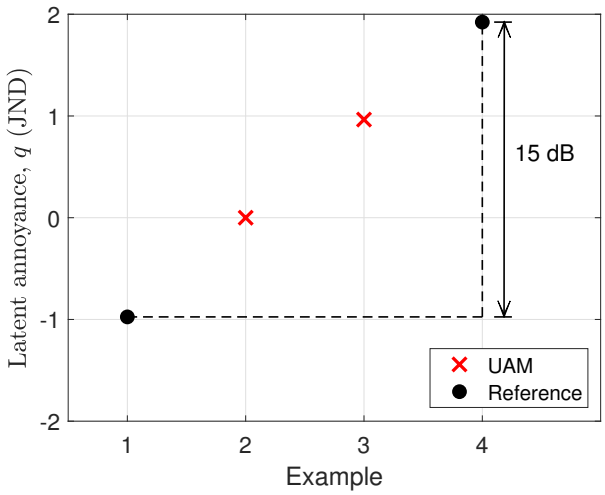


Figure 5. Latent annoyance, q, of example stimuli. A change of 15 dB of the reference sound corresponds to approximately 3 annoyance JNDs, approximately the same range of annoyance produced by UAM stimuli of equal loudness.

shown in Fig. 5. Examples 1 and 4 (black dots) are reference sounds separated by 15 dB. Examples 2 and 3 (red ×'s) are UAM stimuli. Example 2 is the level cruise flight condition at a BPF of 15 Hz with the self noise component suppressed (spectral weighting parameter is −10 dB) and a moving average applied to smooth out the loading and thickness noise. Example 3 is the 5° descent flight condition at a BPF of 80 Hz with an elevated self noise component (spectral weighting parameter is 10 dB) and an added synthesized motor tone complex. These stimuli were chosen as examples since their latent annoyance values are separated by approximately one annoyance JND, indicating that roughly 75% of the subjects agreed which sound was more annoying. For the following, example 2 is 0.98 JND more annoying than example 1, example 3 is 0.96 JND more annoying than example 2 and example 4 is 0.96 JND more annoying than example 3. The change in annoyance of 3 JNDs for a 15 dB change in the level of the reference sound is evident.

The sound quality of the examples is shown in Table 1. When comparing examples 1 and 2, example 2 is a little bit louder and is more tonal than example 1, which is shaped broadband. Blade passages as well as some fluctuations can be heard in example 2. For examples 2 and 3, which are both UAM sounds and equally loud, example 3 has a much higher roughness. Both sounds have roughly the same tonality, but example 3 has higher frequency tones. For examples 3 and 4, example 4 is much louder (difference of 9.6 sones). These observations based on sound quality indicate that increases in loudness, tonality, fluctuations and roughness can all contribute to higher annoyance.

The narrowband SPL spectra (A-weighted) of the example sounds are shown in Fig. 6. The low frequency tones in the spectrum of example 2 are very prominent over the shaped broadband from the spectrum of example 1 (top subplot). While both examples 2 and 3 exhibit prominent spectral peaks, example 3 has significantly

Table 1. Sound quality for example stimuli from the Test for Urban Air Mobility Sound Quality [7].

Example	*Loudness, N (sones)*	*Sharpness, S (acum)*	*Tonality, T (TU)*	*Roughness, R (asper)*	*Impulsiveness, I (IU)*	*Fluctuation strength, F (vacil)*
1	5.62	1.56	0.0362	0.0283	0.312	0.0107
2	6.03	0.783	0.110	0.0193	0.358	0.0176
3	6.06	1.15	0.125	0.856	0.471	0.0147
4	15.7	1.58	0.0539	0.0436	0.332	0.0165

more high frequency content than example 2 (middle subplot). Although spectral peaks below 1 kHz in example 3 are higher in level than the shaped broadband in example 4, the spectrum of example 4 is higher in level than the spectral peaks of example 3 at frequencies above 1 kHz (bottom subplot). Similar to the observations made based on sound quality, these observations based on spectral content indicate that narrowband peaks, higher frequency content, and higher SPL contribute to higher annoyance.

While latent annoyance is not based on acoustical characteristics of the sounds, such as those depicted in Table 1 and Fig. 6, these examples highlight specific instances when measurable differences in the sound lead to higher annoyance. This supports the use of sound quality metrics as a way to quantify this behavior, which may be combined in the form of a psychoacoustic annoyance model.

4 Psychoacoustic annoyance models

Several PA models are available in the literature [10–12, 36]. While Zwicker's PA model was based on synthesized narrowband and broadband noise, with and without modulations, the models by More, Di and Torija focused on particular types of noise sources. More focused on jet noise [10], Di on transformer noise [11] and Torija on drone/small UAS noise [12]. This section summarizes the forms of these models and compares their behavior using the annoyance data collected from the TUSQ.

There are two main assumptions in all four of these PA models: (1) loudness is the main predictor of annoyance and (2) spectral and temporal effects are of secondary importance as predictors of annoyance. An example metric that accounts for spectral effects on annoyance is the sharpness metric. A sound with higher sharpness has more higher frequency content or may be conversely characterized by a lack of low frequency content. Temporal effects, or how a sound changes over time, are usually taken into account by two different types of modulations: fluctuation strength and roughness. Fluctuation strength is perceived at low modulation frequencies, which peaks around a 4 Hz modulation frequency. At such low modulation frequencies, the human ear is able to track these slow modulations in amplitude. When

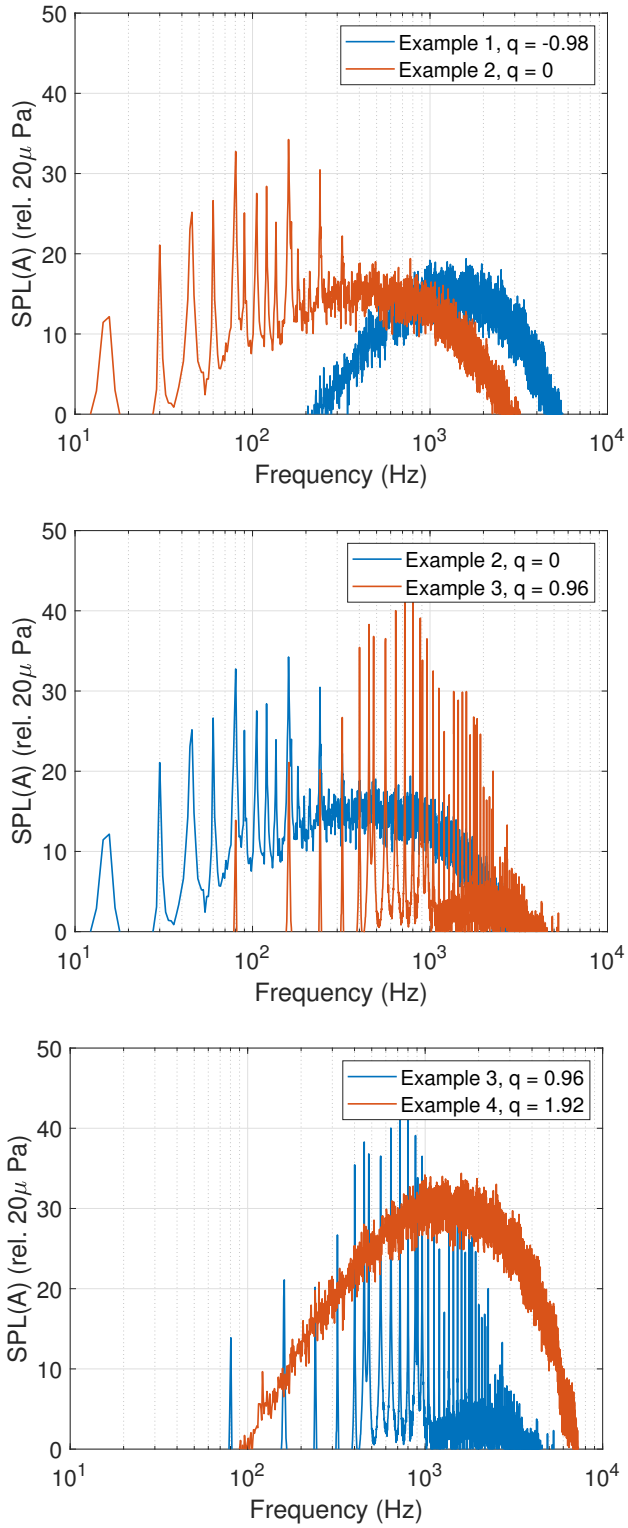


Figure 6. Spectra of pairs that differ by approximately 1 JND ($q_{ij} \approx 1$). In each pair, the spectrum in red corresponds to the more annoying sound.

the modulation frequency increases, the human ear is no longer able to track the time envelope of the sound. At these higher modulation frequencies, modulations are perceived as roughness, which has a maximum perceived effect around 70 Hz.

Changes in loudness alone give a mostly linear change in annoyance, as shown by the reference sound (black dots) in Fig. 7. Additionally, spectral or temporal effects also affect annoyance, as shown by the UAM stimuli in Fig. 7 (red ×'s), where annoyance responses span up to 3 annoyance JNDs for the same loudness level. This demonstrates that loudness alone does not explain the variation in annoyance to UAM noise. While all these PA models take loudness into account in the same way, they differ in how the other sound qualities are incorporated. These differences lead to horizontal adjustments and various degrees of scatter and alignment of all stimuli (UAM and reference). For an ideal model, all red ×'s and black dots would fall along the same line. How sound quality metrics are used in different PA models is described next. How well the PA models align with annoyance responses is assessed in Section 4.5.

4.1 Synthesized narrowband and broadband noise, with and without modulations (Zwicker)

Zwicker's PA model is based on psychoacoustic experiments involving narrowband and broadband noise and the comparison between modulated and unmodulated noise. Of the PA models considered here, it is based on more generic, synthetic types of noise. It is not based on a specific type of industrial or transportation noise source. It combines loudness, temporal and spectral effects in the following form [33–36]:

$$PA = N_5 \left(1 + \sqrt{w_S^2 + w_{FR}^2}\right) \tag{5}$$

where $w_S = (S - 1.75) \times 0.25 \log_{10}(N_5 + 10)$ for $S > 1.75$ acum (0 otherwise) and $w_{FR} = \frac{2.18}{N_5^{0.4}}(0.4F + 0.6R)$. N_5 is the loudness in sones that is exceeded 5% of the time over the duration of the stimulus, S is the sharpness in acum, F is fluctuation strength in vacil and R is roughness in asper.

4.2 Noise from aircraft (More)

The model by More includes the same terms for loudness, sharpness and fluctuation strength/roughness as Zwicker, but it also includes a tonality metric, which relates to the spectral effect. A pure tone is the obvious example of a signal with high tonality. This tonality effect is considered important for rotorcraft noise, since spectra of these sounds often include harmonic content related to blade passage frequencies. However, various degrees of tonality may be perceived for complex sounds. Considering the sound power in band-limited noise, if the frequency range of the passband is decreased enough, the sound no longer is perceived to be broadband, and the sound can be associated with a musical pitch or tone.

More's PA model is given by

$$PA = N_5 \left(1 + \sqrt{\gamma_0 + \gamma_1 w_s^2 + \gamma_2 w_{FR}^2 + \gamma_3 w_T^2}\right) \tag{6}$$

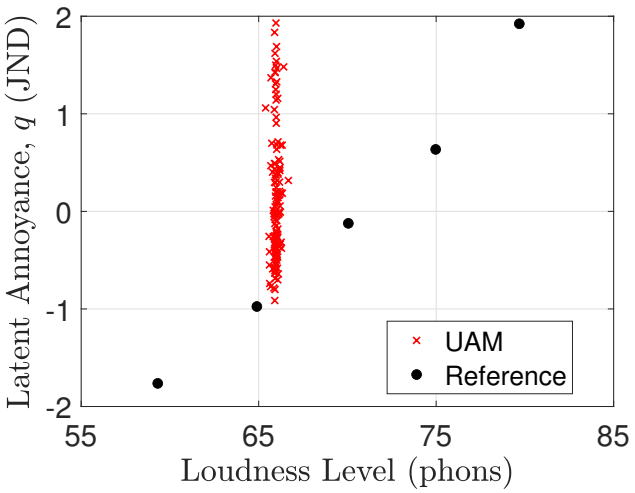


Figure 7. Latent annoyance versus loudness level for the TUSQ stimuli. For the reference sound, changes in loudness level alone induce a nearly linear change in annoyance. A wide range in annoyance is observed for UAM stimuli of equal loudness.

where w_S and w_{FR} are the same as in Zwicker's model and the tonality term, w_T, is given by

$$w_T^2 = \left[\left(1 - e^{-\gamma_4 N_5}\right)^2 \left(1 - e^{-\gamma_5 T}\right)^2\right] \quad . \tag{7}$$

Tonality, T, was calculated using Aures tonality. The γ coefficients were found by fitting the model to annoyance responses to jet noise ($\gamma_0 = -0.16$, $\gamma_1 = 11.48$, $\gamma_2 = 0.84$, $\gamma_3 = 1.25$, $\gamma_4 = 0.29$ and $\gamma_5 = 5.49$). For the TUSQ dataset, $w_S = 0$ for most stimuli, because $S < 1.75$ acum for 138 out of 141 stimuli. For stimuli where w_T is also low, this can lead to negative values under the radical and imaginary numbers for PA. More's model dealt with jet noise where w_S was most likely a contributing factor, so this did not pose a problem. To avoid this complication when applying More's model to the TUSQ dataset in this work, it is assumed that More's model reverts back to Zwicker's model when $\gamma_1 w_S^2 + \gamma_2 w_{FR}^2 + \gamma_3 w_T^2 < -\gamma_0$.

4.3 Noise from transformers (Di)

Di also developed a modified PA model by adding a tonality term to Zwicker's model. However, the form of the tonality term is different from More's. Di's model is given by [11]

$$PA = N_5 \left(1 + \sqrt{w_S^2 + w_{FR}^2 + w_T^2}\right) \tag{8}$$

where the tonality term, w_T, is given by

$$w_T = \beta \frac{T}{N_5^\alpha} \tag{9}$$

and T is calculated with Aures tonality. Di hypothesized the form of this term by fitting Zwicker's PA to annoyance responses with and without high tonality sounds.

By comparing these two fits, Di et al. found a negative correlation between w_T and loudness in phons, justifying the form of w_T. By searching a grid space of α and β, the maximum coefficient of determination, R^2, was found for $\alpha = 0.52$ and $\beta = 6.41$ [11].

4.4 Noise from drones and small Unmanned Aerial Systems (Torija)

The model by Torija [12] includes the sharpness and fluctuation strength/roughness terms from Zwicker, the tonality term from More and a new impulsiveness term, w_I. Impulsiveness is a temporal effect that quantifies sudden changes in amplitude that do not necessarily have a modulation frequency associated with them. Torija's model was developed for noise from aerial drones and UAS and is given by [12]

$$PA = N_5 \left(1 + \sqrt{\gamma_0 + \gamma_1 w_s^2 + \gamma_2 w_{FR}^2 + \gamma_3 w_T^2 + \gamma_4 w_I^2}\right) \tag{10}$$

The model parameters are: $\gamma_0 = 103.08$, $\gamma_1 = 339.49$, $\gamma_2 = 121.88$, $\gamma_3 = 77.20$ and $\gamma_4 = 29.29$. The impulsiveness term is

$$w_I = \frac{0.975I}{N^{-1.334}} \tag{11}$$

4.5 Assessment of models for UAM noise

Correlation between PA model outputs and subjects' annoyance responses from the TUSQ are now studied. The assessment of the models focuses on psychoacoustic annoyance level, L_{PA}, instead of PA itself. This is justified by Fechner's Law, which states that perceptual changes are related to the logarithm of the stimulus intensity [43]. If the stimulus intensity is assumed to be given by loudness in sones, then loudness level, given by

$$L_N = 40 + 10 \log_2 N \quad , \tag{12}$$

better represents the logarithm of stimulus intensity. Therefore, it is more logical to assume a linear relationship between perceived annoyance and L_N instead of N. Indeed, when sound qualities other than loudness are ignored, a mostly linear relationship between L_N and latent annoyance is observed (i.e., reference sound in Fig. 7). Likewise, when sound qualities are considered (i.e., PA instead of N alone), PA may be considered the stimulus intensity, and psychoacoustic annoyance level, L_{PA}, given by

$$L_{PA} = 40 + 10 \log_2(PA) \quad , \tag{13}$$

represents the logarithm of the stimulus intensity. Equation (12) is valid for $N \geq 1$ sone, below which $L_N = 40(N)^{0.35}$ [24]. Similarly, Eq. (13) should be used for $PA \geq 1$, below which $L_{PA} = 40(PA)^{0.35}$ may be more appropriate.

Psychoacoustic annoyance level, L_{PA}, for TUSQ stimuli is predicted using the four models given by Zwicker, More, Di and Torija reviewed previously. Since the latent annoyance, q, was found to be highly correlated with annoyance responses, PA levels are compared to q instead of directly to annoyance rating or paired comparison

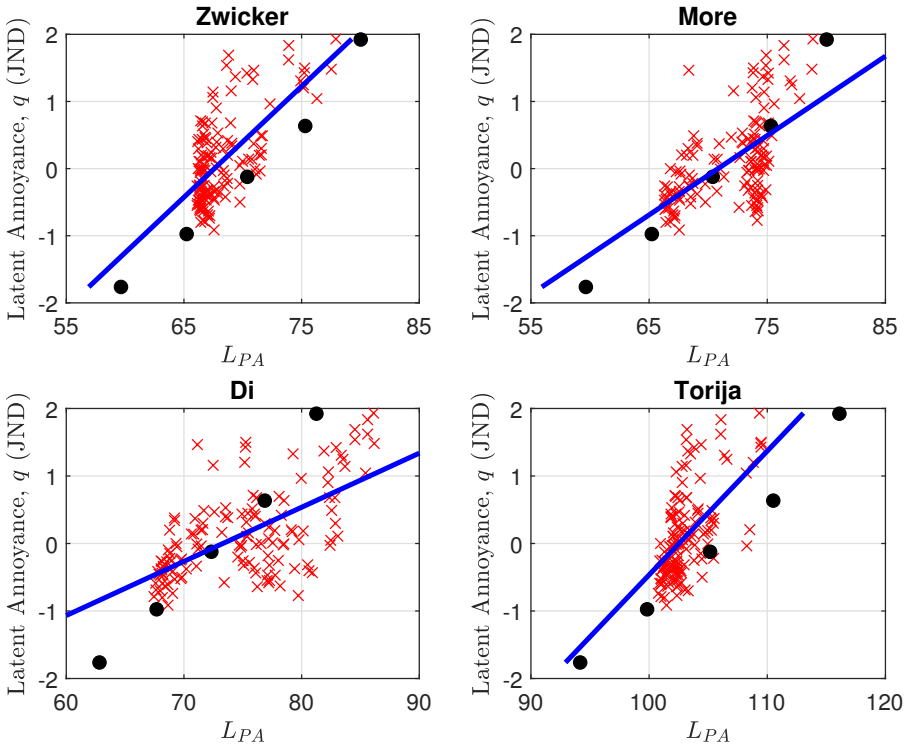


Figure 8. Correlation between psychoacoustic annoyance level, L_{PA} (see Eq. (13)), and latent annoyance, q, for the TUSQ dataset. Pearson correlation coefficients for Zwicker, More, Di and Torija are 0.70, 0.62, 0.63 and 0.65, respectively. Coefficients of determination, R^2, are 0.48, 0.37, 0.39 and 0.42, respectively. (Line of best fit: solid blue. Red $\times$: UAM stimuli. Black circle: reference stimuli.)

responses. The correlation of various PA levels with latent annoyance, q, is evaluated and plotted in Fig. 8. The models are shown to be moderately correlated with the annoyance responses, with Pearson correlation coefficients ranging from 0.62 to 0.70, as shown in Table 2. Zwicker's model has the highest correlation, but there are many stimuli varying from $-1 < q < 1$ that are clustered about a similar L_{PA}. With More's model, there is not a smooth transition from low to high L_{PA}, which is evidenced by two distinct clusters of stimuli. Di's model has less clustering than Zwicker or More but a large amount of scatter. For Torija's model, the data are more tightly grouped at low L_{PA} with gradually increasing spread of q as L_{PA} increases.

Pearson correlation coefficients between latent annoyance and PA or L_{PA} are summarized in Table 2. Looking across models, the correlation coefficient for PA is similar to the correlation coefficient for L_{PA}. While the published models do not transform PA to L_{PA}, as in Eq. (13), this transformation is commonly thought more important when stimuli loudness levels vary over ranges greater than 30 phon.

Table 2. Pearson correlation coefficients between either psychoacoustic annoyance, PA, or psychoacoustic annoyance level, L_{PA}, and latent annoyance, q.

	Zwicker	More	Di	Torija
PA	0.68	0.63	0.64	0.61
L_{PA}	0.70	0.62	0.63	0.65

5 A psychoacoustic annoyance model for urban air mobility noise

The proposed PA model for UAM vehicle noise is given by

$$PA = N_5 \left(1 + \sqrt{w_S^2 + w_{FR}^2 + w_T^2}\right) \tag{14}$$

where

$$w_T = \gamma_1 \frac{T}{N_5^{\gamma_2}} \quad . \tag{15}$$

The model uses w_S and w_{FR} from Zwicker's model and includes a tonality term whose coefficients are found using annoyance responses from the TUSQ dataset. The rationale for using this form of a PA model for UAM noise starts with the fact that Zwicker's model correlated better with latent annoyance than the other available models. Next, Zwicker's sharpness term is used without modification, because only 3 out of 141 stimuli in the test had sharpness above the threshold of 1.75 acum to contribute. Likewise, only low levels of fluctuation strength (maximum of only 0.08 vacil) were tested in the TUSQ. Therefore, there was insufficient variation in w_S and w_{FR} to warrant their modification for UAM noise. Finally, the form of Di's tonality term is used, because it is assumed that for higher loudness levels, higher tonality would be necessary to have the same change in psychoacoustic annoyance.

To find values for γ_1 and γ_2 for use in a PA model for UAM noise, the best linear fit (see Fig. 8) between PA level and latent annoyance is found where PA level is given by Eq. (13). Initially, the line of best fit was found by a non-linear optimization routine that sought to minimize the absolute difference between latent annoyance, q, and PA level when both PA level and latent annoyance were normalized to be within 0 and 1. It was assumed that minimizing these differences would find the parameters of the tonality term that correctly rank the UAM noises from least annoying to most annoying. The output of this optimization proved to be dependent on the initial guess of γ_1 and γ_2, pointing to the existence of many local minima. Instead of attempting more complicated global optimization schemes, a simpler approach based on inspection of R^2 over a 2D grid spanning ranges of γ_1 and γ_2 was followed.

Starting with large ranges for γ_1 and γ_2, it was found that the highest values for R^2 were for $\gamma_1 < 50$ and $\gamma_2 < 3$. From this, it was evident that there was a strong correlation between γ_1 and γ_2. This correlation is highlighted here by examining a smaller range of possible parameters ($0 < \gamma_1 < 10$ and $0 < \gamma_2 < 2$), as shown in Figure 9. The contour lines in that figure show computed R^2 values in steps of 0.05, as a function of γ_1 and γ_2. High correlation between γ_1 and γ_2 is evident since the

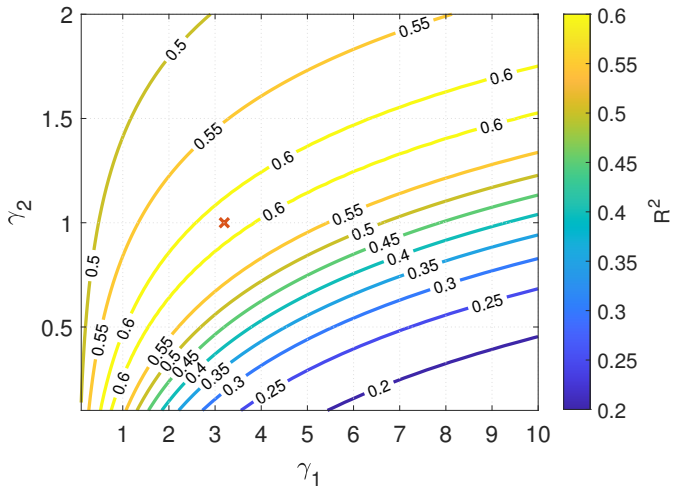


Figure 9. Coefficient of determination, R^2, for the line of best fit between psychoacoustic annoyance level and latent annoyance, q, as in Fig. 8.

contours form neatly ordered, well behaved contours of constant R^2, where many pairings of γ_1 and γ_2 would give the same R^2 value. The line of best fit can be realized from any combination of parameters along the maximum contour line that falls in between the two $R^2 = 0.6$ contours.

The correlation between the tonality term strength and the loudness exponent in the denominator was also observed in Di's work [11]. Although Di did report a maximum value, it appears somewhat arbitrary when the contour lines show many combinations that give an equally good result. Based on this, there is not sufficient evidence to support a unique solution based on two parameters. However, the dependence of the tonality term on loudness is observed for UAM noise just as it was for transformer noise. Based on these considerations, it is reasonable to assume that $\gamma_2 = 1$ and assume that γ_1 is found from the intersection of the maximum contour line in Fig. 9 and $\gamma_2 = 1$. Thus, the tonality term for psyhoacoustic annoyance level for UAM noise is

$$w_T = 3.2\frac{T}{N_5} \tag{16}$$

where tonality, T, is measured in Tonality Units (TU) using Sottek's Hearing Model and N_5 is the 5% exceedance of loudness in sones. The choices for $\gamma_1 = 3.2$ and $\gamma_2 = 1$ are marked in Fig. 9 by a red $\times$.

The line of best fit using the tonality term given above for UAM noise is shown in Fig. 10. When compared to Zwicker's model, the PA model for UAM noise shows a marked improvement in predicting the annoyance responses from TUSQ, which are represented by the latent annoyance, q. The Pearson correlation coefficient increases from 0.70 to 0.78, and the coefficient of determination, R^2, for the line of best fit increases from 0.48 to 0.606. For Zwicker's model, there were many stimuli around $L_{PA} = 66$ that spanned a range of $-1 < q < 1$. This undesirable behavior is absent in the PA model for UAM noise. Other undesirable characteristics, such as the

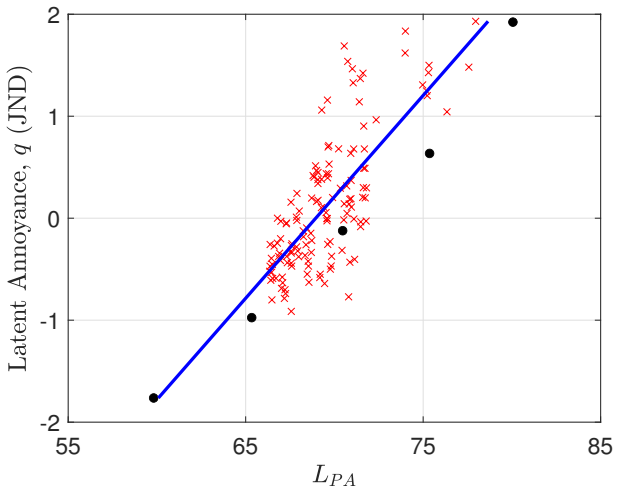


Figure 10. Correlation between psychoacoustic annoyance level, L_{PA} (see Eq. (13)), and latent annoyance, q, using a tonality term developed for UAM noise, where $w_T = 3.2\,T/N_5$. The Pearson correlation coefficient is 0.78, and $R^2 = 0.606$ for the line of best fit. (Line of best fit: solid blue. Red $\times$: UAM stimuli. Black circle: reference stimuli.)

clustering of predictions from More's model, the scatter from Di's model and the increasing spread from Torija's model, are also absent in the updated PA model for UAM noise. It is concluded, therefore, that the PA model for UAM noise with the tonality term described is an improvement upon existing models, both qualitatively and quantitatively.

6 Discussion

Thurstone wrote that investigators postulate discriminal processes that differ from the actual manner in which test subjects make their judgments [40]. In other words, there are limits to any psychophysical model, including the PA model described here and given in Eq. (14). It may be that other sound quality metrics are missing in the model or that included metrics do not capture the complete annoyance response. This model was developed using annoyance responses to steady-state sounds, so its applicability to transient sounds (e.g., flyover events) is unknown. A method to integrate loudness (or PA) level over time is needed in this case. Using 5% exceedance values of sound quality metrics may be a possible alternative, but further analysis or psychoacoustic testing is needed to answer this question. Additionally, the UAM stimuli that were used to develop the model were of equal loudness, so strong interaction effects between loudness and other sound quality metrics may not be properly modeled. Finally, TUSQ stimuli included auralizations of a quadrotor vehicle and modifications of those sounds to span a range of sound quality. Annoyance responses to recordings of flown UAM vehicles may lead to a better model.

With the limitations just described, future improvements or modifications to PA

models for UAM noise are likely. In that case, Thurstone analysis and the latent annoyance scale used in this work may still be used to fit any future annoyance model. While this intermediate step may not be necessary if only rating questions are used, model development is greatly simplified if rating and paired comparison responses can be unified into one common scale, as described in Sec. 3.2. Thurstone analysis also helps to distinguish between the internal psychological scale and the external scale that may be based on measurable quantities of sound, such as SPL, loudness or sound quality metrics. Latent annoyance is not based on any of these measures and merely places stimuli on a scale from least to most annoying and quantifies the difference in annoyance in terms of annoyance JNDs.

Another useful result from Thurstone analysis applied to rating and paired comparison data is that the relative accuracy of different pyschoacoustic tasks can be quantified. Specifically, paired comparison questions were found to be more accurate than rating questions. This may inform experimental designs of future psychoacoustic investigations on annoyance.

The results from Fig. 4 show that the range in annoyance to UAM sounds of equal loudness spans a 15 dB change in level of the reference sound. This is reminiscent of some of the earliest perceptual tests of jet noise in the 1950's, when the level of a new passenger jet (before being retrofitted with engine mufflers) was judged equally disturbing to a propeller-driven aircraft only after the jetliner's sound level was reduced by 15 dB [44]. The current work indicates that characteristics other than the loudness level of transportation noise continue to have large effects on annoyance.

Noise from UAM vehicles must be at a level that is acceptable to the communities in which the vehicles operate [4]. Otherwise, "UAM operations will not be allowed to occur at all times or at all locations when...detrimental impacts to people...are deemed to be too high [2]." Predicting noise annoyance, therefore, is critical to the evaluation of potential impacts of UAM missions and to inform vehicle designers and transportation service providers. While the model presented in this work is far from predicting community response to UAM noise, it does offer a step forward in predicting short-term annoyance to steady-state UAM noise.

7 Conclusions

The main result of this work is a psychoacoustic annoyance model for Urban Air Mobility noise, Eq. (14), that is based on Zwicker's model but includes an extra term for tonality. The tonality term, w_T, is $3.2\,T/N_5$ where T is tonality measured in TU units using an auditory model and N_5 is the loudness in sones exceeded 5% of the time. This proposed model shows improvements over Zwicker's and the three other existing models considered by showing a higher correlation between latent annoyance and psychoacoustic annoyance level, L_{PA}, given by Eq. (13).

The psychoacoustic test consisted of rating as well as paired comparison annoyance responses to UAM stimuli. The stimuli were based on noise predictions of a NASA reference quadrotor vehicle in two flight conditions, which were then auralized and modified to span a range of sound quality.

To compare models and arrive at the UAM psychoacoustic annoyance model, annoyance data collected during the psychoacoustic test were converted to a latent annoyance scale based on Thurstone's Law of Comparative Judgment. The model was then fit to latent annoyance. Latent annoyance is the underlying psychological scale for annoyance and assumes nothing about the physical measure of sound. It is, therefore, suitable to fit any model of annoyance that may be developed in the future. Results show that 1 just-noticeable-difference on the latent annoyance scale corresponds to about a 5 dB change in a shaped broadband sound and that the range of annoyance to UAM stimuli of equal loudness covered a range of 3 annoyance JNDs in the psychoacoustic test. Evaluating annoyance in terms of JNDs shows pairs of sounds where 75% of the subjects agree on which sound is more annoying than the other, highlighting aspects of sound quality and SPL spectra that contribute to human perception.

While the proposed model for UAM psychoacoustic annoyance may be limited to the ranges of sound quality tested, care has been taken to not overfit this model. Specifically, the sharpness and fluctuation strength/roughness terms from Zwicker, as well as their relative weights in the model, were not altered. This is in contrast to other psychoacoustic annoyance models that have been fit to certain types of noise. Nevertheless, improvements will likely be made to the proposed model based on future laboratory tests. In which case, the latent annoyance results shown here should provide a good basis to explore future UAM noise annoyance model refinements.

Acknowledgments

The authors would like to thank Brian Tuttle (Analytical Mechanics Associates, Inc.) for help preparing the auralizations and executing the test, Aric Aumann (Analytical Services and Materials, Inc.) for help executing the test and Erin Thomas (Analytical Mechanics Associates, Inc.) for recruiting test subjects and performing hearing screenings. Menachem Rafaelof (National Institute of Aerospace) is acknowledged for many fruitful discussions. The authors also gratefully acknowledge the support of Ran Cabell, Noah Schiller, Susan Gorton, Benny Lunsford. This research was conducted in support of the NASA Revolutionary Vertical Lift Technology Project. Acknowledgment is given for technical reviews from Aaron Vaughn and Jacob Klos (NASA Langley Research Center).

References

1. Patterson, M.; Antcliff, K.; and Kohlman, L.: A proposed approach to studying Urban Air Mobility missions including an initial exploration of mission requirements. *74th Annual Forum & Technology Display of the American Helicopter Society*, 2018.

2. Thipphavong, D. P.; Apaza, R.; Barmore, B.; Battiste, V.; Burian, B.; Dao, Q.; Feary, M.; Go, S.; Goodrich, K. H.; Homola, J.; Idris, H. R.; Kopardekar, P. H.; Lachter, J. B.; Neogi, N. A.; Ng, H. K.; Oseguera-Lohr, R. M.; Patterson, M. D.;

and Verma, S. A.: Urban Air Mobility Airspace Integration Concepts and Considerations. *2018 Aviation Technology, Integration, and Operations Conference*, American Institute of Aeronautics and Astronautics, jun 2018.

3. Goodrich, K. H.; and Theodore, C. R.: Description of the NASA Urban Air Mobility Maturity Level (UML) Scale. *AIAA Scitech 2021 Forum*, American Institute of Aeronautics and Astronautics, Jan. 2021.

4. B.P. Hill et al.: UAM Vision Concept of Operations (ConOps) UAM Maturity Level (UML) 4. , NASA, 2020. URL https://ntrs.nasa.gov/citations/20205011091.

5. Rizzi, S.; Huff, D.; Boyd, Jr., D.; Bent, P.; Henderson, B.; Pascioni, K.; Sargent, D.; Josephson, D.; Marsan, M.; He, H.; and Snider, R.: Urban Air Mobility Noise: Current Practice, Gaps, and Recommendations. NASA/TP-2020-5007433, 2020.

6. Rizzi, S.: Toward reduced aircraft community noise impact via a perception-influenced design approach. *Inter.Noise*, Hamburg, 2016.

7. Boucher, M.; Rafaelof, M.; Begault, D.; Christian, A.; Krishnamurthy, S.; and Rizzi, S.: A Psychoacoustic Test for Urban Air Mobility Vehicle Sound Quality. *SAE Technical Paper Series*, NVC, SAE International, May 2023.

8. Silva, C.; Johnson, W. R.; Solis, E.; Patterson, M. D.; and Antcliff, K. R.: VTOL Urban Air Mobility Concept Vehicles for Technology Development. *2018 Aviation Technology, Integration, and Operations Conference*, American Institute of Aeronautics and Astronautics, jun 2018.

9. Zwicker, E.; and Fastl, H.: *Psychoacoustics: Facts and Models*. Springer, 2nd ed., 2013.

10. More, S.: Aircraft noise characteristics and metrics. Ph.D. Thesis, Purdue University, West Lafayette, Indiana, Dec. 2010.

11. Di, G.-Q.; Chen, X.-W.; Song, K.; Zhou, B.; and Pei, C.-M.: Improvement of Zwicker's psychoacoustic annoyance model aiming at tonal noises. *Applied Acoustics*, vol. 105, apr 2016, pp. 164–170.

12. Torija, A. J.; Li, Z.; and Chaitanya, P.: Psychoacoustic modelling of rotor noise. *The Journal of the Acoustical Society of America*, vol. 151, no. 3, mar 2022, pp. 1804–1815.

13. Hubbard, H. H.: Aeroacoustics of Flight Vehicles: Theory and Practice. Volume 1: Noise Sources. 1991.

14. Thurstone, L. L.: A law of comparative judgment. *Psychological Review*, vol. 34, no. 4, July 1927, pp. 273–286.

15. Perez-Ortiz, M.; and Mantiuk, R. K.: A practical guide and software for analysing pairwise comparison experiments. 2017.

16. Johnson, W.: Comprehensive Analytical Rotorcraft Model of Rotorcraft Aerodynamics and Dynamics, Version 4.10. Johnson Aeronautics, 2017. Volumes I-9.

17. Lopes, L. V.; and Burley, C. L.: Design of the Next Generation Aircraft Noise Prediction Program: ANOPP2. *17th AIAA/CEAS Aeroacoustics Conference*, 2013.

18. Farassat, F.; and Succi, G.: A review of propeller discrete frequency noise prediction technology with emphasis on two current methods for time domain calculations. *J. of Sound and Vib.*, vol. 71, no. 3, 1980, pp. 399–419.

19. Farassat, F.; and Succi, G.: The prediction of helicopter rotor discrete frequency noise. *Vertica*, vol. 7, 1983, pp. 309–320.

20. Brooks, T.; Pope, T.; and Marcolini, M.: Airfoil self-noise and prediction. *NASA/RP-1218*, 1989.

21. Aumann, A.; Tuttle, B.; Chapin, W.; and Rizzi, S.: The NASA auralization framework and plugin architecture. *inter.noise*, San Francisco, Aug. 2015.

22. Krishnamurthy, S.; Rizzi, S. A.; Cheng, R.; Boyd, D. D.; and Christian, A. W.: Prediction-Based Auralization of a Multirotor Urban Air Mobility Vehicle. American Institute of Aeronautics and Astronautics, jan 2021.

23. Krishnamurthy, S.; Aumann, A. R.; and Rizzi, S. A.: A Synthesis Plugin for Auralization of Rotor Self Noise. *AIAA AVIATION 2021 FORUM*, American Institute of Aeronautics and Astronautics, July 2021.

24. ISO: ISO 532-1:2017 Acoustics–Methods for calculating loudness–Part 1: Zwicker method. 2017.

25. DIN 45631/A1. Calculation of loudness level and loudness from the sound spectrum - Zwicker method - Amendment 1: Calculation of the loudness of time-variant sounds. 2009.

26. Acoustics, H.: ArtemiS Suite 13.6.

27. DIN 45692:2009, Measurement technique for the simulation of the auditory sensation of sharpness. German Institute for Standardization.

28. Sottek, R.: Modelle zur Signalverarbeitung im menschlichen Gehör (in German). Ph.d. thesis, RWTH Aachen, 1993.

29. Sottek, R.; and Genuit, K.: Models of signal processing in human hearing. *AEU - International Journal of Electronics and Communications*, vol. 59, no. 3, jun 2005, pp. 157–165.

30. Bray, W.: A new psychoacoustic method for reliable measurement of tonalities according to perception. *Inter-Noise 2018*, Chicago, Aug. 2018.

31. ECMA-418: Psychoacoustic metrics for ITT equipment - Part 2: Models based on human perception, 2nd edition. Dec. 2022.

32. Faller II, K.; Rizzi, S.; and Aumann, A.: Acoustics performance of a real-time three-dimensional sound reproduction system. Technical Memorandum NASA TM-2013-218004, National Aeronautics and Space Administration, 2013.

33. Widmann, U.: A psychoacoustic annoyance concept for application in sound quality. *Noise-Con 97*, 1997, pp. 491–496.

34. Fastl, H.; and Zwicker, E.: *Psychoacoustics*. Springer Berlin Heidelberg, third ed., 2007. URL https://www.ebook.de/de/product/25374312/hugo_fastl_eberhard_zwicker_psychoacoustics.html.

35. Fastl, H.: From psychoacoustics to sound quality engineering. *Proc. of the Institute of Acoustics*, vol. 24, 2002.

36. Widmann, U.: Subjektive Beurteilung der Lautheit und der Psychoakustischen Lästigkeit von PKW-Geräuschen. *Proc. German Annual Conference on Acoustics, DAGA 95*, Oldenburg, 1995, pp. 875–878.

37. Perez-Ortiz, M.; Mikhailiuk, A.; Zerman, E.; Hulusic, V.; Valenzise, G.; and Mantiuk, R. K.: From Pairwise Comparisons and Rating to a Unified Quality Scale. *IEEE Transactions on Image Processing*, vol. 29, 2020, pp. 1139–1151.

38. Thorndike, E.: The measurement of the quality of handwriting. *Teachers College Record*, vol. 11, 1910, pp. 86–151.

39. Bech, S.; and Zacharov, N.: *Perceptual Audio Evaluation–Theory, Method and Application*. Wiley, apr 2006.

40. Thurstone, L.: The measurement of value. *Pscychological Review*, vol. 61, 1954, pp. 47–58.

41. Mikhailiuk, A.; and Mantiuk, R.: pwcmp_rating_unified. 2019. URL https://github.com/gfxdisp/pwcmp_rating_unified.

42. Thurstone, L.: A mental unit of measurement. *Psychological Review*, vol. 34, 1927, pp. 415–423.

43. Gescheider, G.: *Psychophysics: The Fundamentals*. Lawrence Erlbaum Associates, Mahwah, New Jersey, third ed., 1997.

44. Beranek, L. L.: *Riding the Waves*. The MIT Press, 2008.