1. INTRODUCTION

Although every helicopter in operation has to go through a noise certification process, annoyance due to helicopters still persists within various communities. This implies that certification metrics do not capture the full range of human response and that predicted reactions could be supplemented with other information, which could be acoustic or nonacoustic in nature. The rotorcraft sound quality metric (RoQM) psychoacoustic experiment was designed to determine the relative importance of sound quality metrics (SQMs), such as sharpness, tonality, loudness, fluctuation strength and impulsiveness, on human annoyance to rotorcraft sounds.[1] Starting from a baseline helicopter recording, SQMs were varied synthetically and presented to subjects who responded with an annoyance rating. The RoQM test took place in 2017 at the NASA Langley Research Center in the Exterior Effects Room.[2] A total of 105 sounds were played to 40 subjects. The relationship between helicopter noise sound quality and annoyance is modeled using multilevel regression in this work, which takes into account the variability of responses across subjects. Previous analyses did not consider such a grouping of the data.[1]

A. HELICOPTER SOUND QUALITY

Researchers have long asked if helicopter noise can be adequately rated and whether certification metrics correlate well enough with public reactions.[3] For example, Schomer indicated that day-night level (DNL) correlated well with annoyance to aircraft noise except for helicopters.[4] Similarly, some evidence of the deficiency of sound exposure level (SEL) was found, showing that helicopter noise requires a $4\,\mathrm{dB}$ penalty when compared to white noise of equal SEL.[5] Although the difficulty is often ascribed to "blade slap" or the impulsive nature of helicopter noise,[3] the literature is not in agreement that impulsiveness is the dominant driver of annoyance.[6] Recent work suggests that a more complete palette of SQMs, such as loudness, sharpness, tonality and impulsiveness, may be more appropriate in predicting human annoyance to helicopter noise.[7]

Leverton describes the difficulty of applying a given metric to helicopter noise, as it is influenced by impulsive noise at the blade passage frequency of the main rotor and tonal noise up to 10 harmonics of the tail rotor.[3] This description alone is a hint that the SQMs of impulsiveness, tonality and sharpness may be investigated. Due to this complex nature of helicopter noise, these and possibly other SQMs, which are based on how the human auditory system reacts to acoustic stimuli, may be good predictors of annoyance to helicopter noise. A simple annoyance model based on SQMs, as well as its subsequent modifications, has already been applied to a variety of sounds.[8] Furthermore, recent work has highlighted that a subjective listening test that varies SQMs independently could lead to a deeper understanding of annoyance to helicopter noise.[7]

This idea has been applied in the area of domestic product sound quality for quite some time. An important point is that sound quality is specific to each product. For example, a definition of sound quality has been proposed as: "sound quality is the perceptual reaction to the sound of a product that reflects the listener's reaction to the acceptabilty of that sound for that product...."[9] As it has been said that "a good refrigerator does not sound like a good lawnmower," (*ibid*) it could also be said that a good fixed-wing aircraft does not sound like a good helicopter. This implies that the perception of sound quality that corresponds to a physical scale of sound may be different for different types of aircraft and that certification metrics alone may not encompass the complete human response. This is the same motivation for reducing community impact due to aircraft noise through perception-influenced design.[10] Since SQMs are a physical scale of sound that corresponds to human responses, it is the aim of this work to model the relationship between these SQMs and annoyance responses to helicopter noise.

B. MULTILEVEL ANALYSIS OF TRANSPORTATION NOISE

Multilevel analysis, which is explained in more detail in Section 2.D, is a linear regression technique that assumes a random distribution of regression parameters based on assigned groupings of the response data.[11] There are many reasons to use multilevel analysis instead of conventional linear regression (i.e., without grouping) when analyzing the relationship between helicopter sound quality and annoyance. First, it is a good compromise between analyzing unpooled and aggregated data. Secondly, a variable intercept and/or slope can be represented explicitly in the model, which makes grouping variables simple to implement. Also, data that are nested can be modeled effectively by including another level in the analysis.[11,12]

Further motivation for using multilevel analysis in this work is that it has already been successfully applied to a wide range of dose-response studies related to transportation noise. The following is meant to be a representative, although not exhaustive, review of the use of multilevel analysis in transportation noise annoyance studies.

Groothuis-Oudshoorn and Miedema applied multilevel analysis to studies on annoyance to noise from different modes of transportation.[13] The relationship between the predictor DENL (day-evening-night level) and annoyance was analyzed, using data from 53 studies of aircraft, road traffic and railway noise. The grouping was based on different studies, which was implemented as a variable intercept. They were able to determine that the percentage of people highly annoyed was highest for aircraft noise and lowest for railway noise. The explained proportion of variance of DENL was only around 16%, which indicates that other (possibly nonacoustic) factors also contributed to the differences in annoyance.

Trollé applied multilevel analysis in a study in which recordings of tramway noise were played back to subjects in the laboratory.[14] Acoustical predictors were the A-weighted equivalent sound pressure level and one that quantified the tonal energy within critical bands. A similar procedure was applied to aircraft flyover noise.[15] The individual subject was the grouping variable in both studies. Not only was loudness a significant predictor in the analysis, as expected, but also the temporal derivative of loudness. This implies that sudden changes in the noise are also annoying to individuals. Along with the analysis, verbal reports from the test subjects indicated that spectral content, temporal variation and perceived sound intensity were all contributing factors to annoyance. In both studies, noise sensitivity was used as a predictor at the individual subject level. Although noise sensitivity had a small effect on the annoyance ratings, it was still suggested that future studies include this nonacoustic factor in the statistical model.[14,15]

Mulitlevel analysis was applied to a comparative study between wind turbine noise and road traffic noise.[16] Source type, A-weighted sound pressure level and the presence of amplitude modulation were used as predictors, and the responses were grouped by individual subject. It was found that for sounds of equal A-weighted sound pressure level, wind turbine noise was more annoying than road traffic noise. Furthermore, it was found that the amplitude modulation of the wind turbine noise was an important attribute.

Wilson et al. used multilevel analysis to model annoyance to airport noise measured in DNL.[17] They used 43 community airport noise surveys compiled by Fidell et al.[18] The data were grouped by survey. Since the fitted logistic curve was in terms of the 50% point of highly annoyed individuals, the Community Tolerance Level was explicit in the formulation of the multilevel model. The "within community" variation was larger than the "between community" variation, but the latter was significant and quantifiable. It was shown that multilevel analysis is well-suited to account for individual differences as well as differences between communities.

Rathsam applied multilevel analysis to community noise social surverys of sonic booms and military blast noise.[19] Noise exposure (C-weighted DNL) was used as a predictor. Multilevel models fit the data better than ordinary regression models, because the variability among participants was taken into account.

Taghipour et al. used multilevel analysis (mixed effects model) to model short-term annoyance to civil helicopter and propeller-driven aircraft noise.[20] Their set of stimuli consisted of four helicopters, five propeller-driven aircraft types, takeoffs and landings and original or processed recordings. The predictors

were the source type, type of maneuver and the sound exposure level. Responses were grouped by subject. They mainly conclude that the annoyance due to civil helicopter and propeller-driven aircraft noise is determined by the sound exposure level, and the source type is not a determining factor. Only small differences were found based on the type of procedure (takeoff or landing).

Overall, multilevel analysis has proven to be a robust type of analysis when it comes to modeling the dose-response relationship between transportation noise and annoyance. In particular, the differences across subjects, surveys, mode of transportation and vehicle operations have all been successfully analyzed. Different factors, including loudness, various weightings of sound pressure level and noise parameters related to tonal components and temporal variation, have been investigated as possible indicators of annoyance to transportation noise. Several studies have indicated the importance of nonacoustic factors, especially noise sensitivity.

Several reasons for using multilevel analysis in this work were stated at the beginning of this section. However, the most important reason is that after taking into account the variations between subjects, the relationship between helicopter noise sound quality and annoyance is more clear.

C. OUTLINE OF PAPER

The rest of the paper is organized as follows: Section 2 describes the SQMs used in this work and how they were calculated, along with a description of the noise stimuli used in the psychoacoustic test. It also details how the annoyance responses were collected and presents the multilevel models used to relate the SQMs of the noise stimuli with the annoyance responses. Section 3 presents the results of the analysis and is followed by the main conclusions of this work.

2. METHODS

In this section, the methods used in "The Rotorcraft Sound Quality Metric" (RoQM, 2017) psychoacoustic test are described. Sound quality metrics are introduced, along with how they were calculated. Then, the process of creating helicopter noise stimuli is explained, which consists of creating a baseline sound from recordings, followed by synthesizing changes in the SQMs. Next, the psychoacoustic test is described, which was performed in the NASA Langley Exterior Effects Room. Finally, multilevel analysis is detailed, which was used to find the relationship between annoyance and changes in SQMs of helicopter noise.

A. SOUND QUALITY METRICS IN PLACE OF AUDITORY SYSTEM

Sound quality metrics transform a physical measure of sound into units of measure that correspond to a linear response of the human auditory system (e.g., a doubling of the psychoacoustic quantity corresponds to a doubling of the perceived level).[21] There are standards for some calculations, such as loudness and sharpness, while others (e.g., impulsiveness) lack an international standard. Many options for the calculation procedures are included in the ArtemiS Suite software,[22] which was used in this work for calculating the SQM values of the noise stimuli. In the following, the process for calculating loudness, sharpness, fluctuation strength, roughness, tonality and impulsiveness is briefly explained.

Loudness, measured in sone, is the sensation that most closely corresponds to sound intensity and indicates the perception of sound level. While various loudness standards exist, the one used here is DIN 45631/A1, which is a standardization of the Zwicker loudness model and includes a modification for time varying signals. Sharpness, measured in acum, starts from the loudness DIN 45631/A1 standard and then calculates the spectral balance of the sound. Sharper sounds have more higher frequency content. The DIN 45692 sharpness standard is used in this work, which does not take into account the absolute loudness level.

For fluctuation strength, roughness and impulsiveness, the hearing model given by Sottek[23] (implemented in ArtemiS) is used. The hearing model substitutes for the signal processing chain present in the human auditory system. It starts with filters for the middle and outer ear, and then a filterbank of overlapping bandpass filters is applied that takes into account the frequency selectivity of the inner ear. The threshold in quiet is taken into account with a specific adjustment for low frequencies. Subsequent steps mimic the limitations of human hearing's ability to track fast temporal changes within a critical band.

Sounds that vary slowly, in which the envelope changes at less than $20\,\mathrm{Hz}$, give the perception of fluctuation strength. Measured in vacil, it has a maximum for a modulation frequency of $4\,\mathrm{Hz}$. For modulation frequencies above $20\,\mathrm{Hz}$ and less than $300\,\mathrm{Hz}$, human hearing is not able to respond in real-time to the fluctuations, and this perception is given by roughness. Measured in asper, roughness has a maximum for a modulation frequency of $70\,\mathrm{Hz}$. Impulsiveness is the perception caused by short, sudden changes in sound intensity and is measured in IU.

If the sound spectrum is strong within a very narrow band, the sound is perceived as tonal. Tonality, measured in TU, was computed in ArtemiS based on the Aures/Terhardt calculation.[24]

B. HELICOPTER NOISE STIMULI

The acoustic stimuli used for the psychoacoustic test consisted of synthesized (rather than recorded) helicopter sounds, which was done for two reasons. First, in a simple flyover recording, the sound quality of the helicopter changes. If subjects were to report their annoyance response, the specific point in the flyover recording corresponding to the response would be unclear. Secondly, it is difficult to find any pair (let alone a series) of recorded sounds that differ only in a specific SQM, so the influence of an individual metric would be a challenge to isolate. In contrast to recordings, simulated helicopter sounds that have a mostly steady sound quality over a few seconds can be easily created, and it is possible to generate a series of test sounds in which one SQM changes while the others remain roughly constant. For these reasons, the test stimuli played to the test subjects consisted of simulated helicopter noise.

The simulated stimuli for the listening test were created after examining a set of 172 flyover recordings of 6 helicopters of different make and model. The minimum and maximum value of the SQMs from the recordings were calculated, and the synthesized sounds were required to stay within this range. This requirement helped to keep the synthesized sounds helicopter-like.[1]

The baseline synthesized sound started from a flyover recording of an AS350 helicopter. The average main and tail rotor noise was separated using methods given by Greenwood and Schmitz.[25] Starting with this baseline sound, a series of perturbation methods were developed to vary an individual SQM while keeping others roughly constant. The perturbation methods consisted of changes to the amplitude, phase or frequency of either the main or tail rotor blade passage frequencies. For example, changing the blade passage frequency of the tail rotor primarily changed tonality while keeping other metrics mostly constant.[1]

Perturbation methods that changed one metric while keeping the others roughly constant were found for fluctuation strength, sharpness, tonality and impulsiveness. Changes in roughness could not be separated from changes in impulsiveness. Since loudness is believed to be a dominant indicator of annoyance,[7,26] all test sounds were adjusted to give approximately the same loudness. More details on the perturbation methods used to generate the test stimuli are given by Krishnamurthy.[1]

In total, 105 unique sounds, each $5\,\mathrm{s}$ long, were used in the psychoacoustic test. Each SQM value is calculated after an initial $0.5\,\mathrm{s}$ transient period, except for fluctuation strength, which is calculated after $3\,\mathrm{s}$. Then, the SQM value that is exceeded 5% of the time is used as a predictor for the multilevel analysis. The following short-hand is used when referring to the SQMs:

- N: loudness exceeded 5% of the time
- S: sharpness exceeded 5% of the time

- T: tonality exceeded 5% of the time
- F: fluctuation strength exceeded 5% of the time
- I: impulsiveness exceeded 5% of the time
- R: roughness exceeded 5% of the time

.

C. COLLECTION OF ANNOYANCE RESPONSES

The psychoacoustic test was conducted in the Exterior Effects Room at the NASA Langley Research Center (LaRC). The test and its protocol were approved by the Institutional Review Board at NASA LaRC. The test had paid subjects from the general population around Hampton, VA (USA). The sounds were played to 40 subjects in total, in groups of four. Each group listened to the 105 sounds spread out over 4 sessions, giving the subjects any needed breaks to limit fatigue. Subjects reported their annoyance responses (via tablets) on a scale from "Not at all" up to "Extremely", with intermediate, equal intervals marked at "Slightly", "Moderately" and "Very". The subjects were free to respond in between these markings as well. The responses were converted after the test to a number scale from 0 to 10 (decimal values allowed) where the descriptors corresponded to odd numbers.

Recordings of the sound stimuli were made at the four test seat locations with no subjects present. The microphones were placed approximately at the between-the-ears location of an average listener. The SQMs were evaluated for each seat individually so that each SQM value corresponded as closely as possible to what was heard by each subject.

D. MULTILEVEL ANALYSIS

Multilevel analysis can be thought of as a combination of no pooling and complete pooling, which is shown in Figure 1. With no pooling, a regression analysis would be done for one subject's responses at a time, leading to subject-specific regression parameters that are not influenced by how other subjects responded. With complete pooling, annoyance responses of all subjects are aggregated, leading to regression parameters that apply to the entire population, not individuals. Furthermore, complete pooling assumes that annoyance responses from the same subject are independent. However, it is often accepted that a subject's annoyance responses are more similar to his/her own than they are to another subject's responses. In fact, "...observations from within the same group are generally more similar to each other than the observations from different groups."[12] Therefore, the assumption of independence of the annoyance responses may not be valid, which is why multilevel models are needed. This line of reasoning indicates that the grouping variable should be the subject.[a] Therefore, multilevel analysis, i.e., partial pooling, is achieved by assuming a random distribution of regression parameters across subjects. The most important advantages are that the lack of statistical independence is taken into account and that aggregate as well as subject-specific regression parameters are found.

A linear regression model using one SQM as an explanatory variable (i.e., predictor) X_1 has the form

$$Y_i = \gamma_0 + \gamma_1 X_{1i} + e_i \tag{1}$$

in which Y_i is an annoyance response to the i-th sound, X_{1i} is the 5% exceedance SQM value of the i-th sound,[b] γ_0 is an intercept, γ_1 is a slope of the metric X_1 and e_i is the residual error. If responses to all 105 sounds from all 40 subjects are used in the above analysis, it is equivalent to the complete pooling case as described above. There is one intercept and one slope that fits all the data.

[a] In this work, "grouping" always refers to a grouping variable in a multilevel analysis context. It does not refer to the no pooling case where a linear regression would be done on one subject's responses at a time.

[b] The SQM values are centered about the mean of 105 sounds, which is common practice in multilevel analysis.[12]

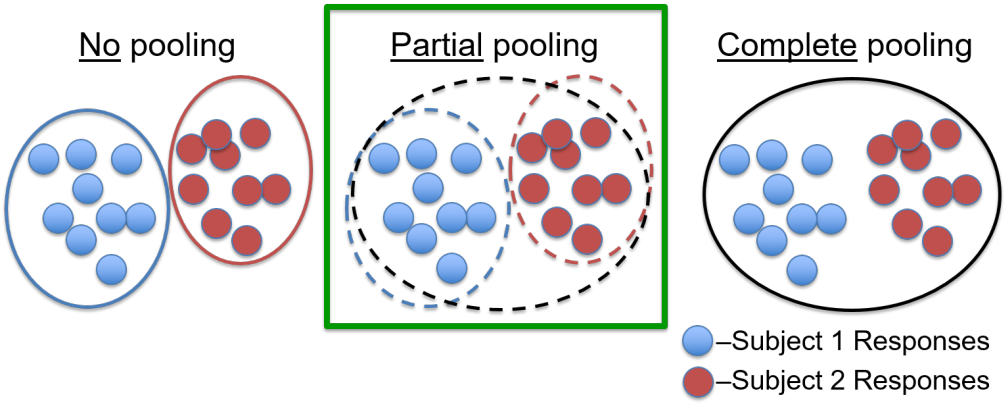

Figure 1: Multilevel analysis can be thought of as partial pooling of the response data, a combination of no pooling and complete pooling. Annoyance responses are grouped by subject, because independence of each annoyance response can not be assumed.

If the responses are grouped based on the subject, a multilevel model is introduced. For example, the intercept may vary across subjects. This adds a second level to Eq. (1), resulting in

$$Y_{ij} = \beta_{0j} + \gamma_{10} X_{1i} + e_{ij} \tag{2}$$

in which Y_{ij} is an annoyance response to the i-th sound by the j-th subject. The variable intercept β_{0j} is the sum of two terms, $\beta_{0j} = \gamma_{00} + u_{0j}$, where γ_{00} is an overall mean intercept for all subjects and u_{0j} is a subject-specific offset. The subject-specific intercept offsets are assumed to be a part of a normal distribution such that $u_{0j} \sim N(0, \sigma^2_{u_0})$, with $\sigma^2_{u_0}$ being the variance of the distribution. The slope of metric X_{1i} is γ_{10} in which the second subscript 0 indicates it does not vary across subjects. Although Eq. (2) is written in terms of the j-th subject, it is connected to other subjects by the overall mean intercept, γ_{00}, and is the mechanism in which partial pooling is applied.

Similarly, the slope may also vary across subjects ($\beta_{1j} = \gamma_{10} + u_{1j}$), yielding a subject-specific slope offset u_{1j}. Again, a normal distribution is assumed such that $u_{1j} \sim N(0, \sigma^2_{u_1})$, with $\sigma^2_{u_1}$ being the variance of the distribution.

Since the intercept and slope can have different assumptions, several different models are studied in this work, which are summarized in Table 1. $M0$ is a standard linear regression model equivalent to Eq. (1). $M1$ is a multilevel model with a variable intercept and no sound quality metric predictors. $M2$ is a multilevel model with a variable intercept and the same slope for all subjects. Lastly, $M3$ is a multilevel model in which both the intercept and slope are variable.[c]

Other models analyzed in this work contain several SQMs as predictors. For a multilevel model with K sound quality metrics as predictors,

$$Y_{ij} = \beta_{0j} + \sum_{k=1}^{K} \beta_{kj} X_{ki} + e_{ij} \quad . \tag{3}$$

in which X_{ki} is the k-th SQM for sound i. Although each slope in Eq. (3) is shown to be variable, it is possible to assume some SQMs have fixed slopes while others are variable. If σ^2_{uk} was small, for example, it might be advantageous to assume a fixed slope and eliminate this parameter.

To refer to models and describe which predictors are included in the model, some short-hand is adopted. For example, a model with a variable intercept, fixed slope and loudness as a predictor is given by $M2$.N.

[c]The reason for not including a model with fixed intercept and variable slope is discussed in Section 3.A.

Table 1: Summary of simple models used to analyze ROQM annoyance response data. Only one explanatory variable (SQM) is used at a time. $M0$ ***is equivalent to standard linear regression, which represents complete pooling of the data. For variable regression parameters, the annoyance responses are grouped by subject, and two subscripts indicate a multilevel model.***

Model	Intercept		Slope	
$M0$	fixed	γ_0	fixed	γ_1
$M1$	variable	$\gamma_{00} + u_{0j}$	0	0
$M2$	variable	$\gamma_{00} + u_{0j}$	fixed	γ_{10}
$M3$	variable	$\gamma_{00} + u_{0j}$	variable	$\gamma_{10} + u_{1j}$

Similarly, $M3$.STF is a variable intercept and variable slope model with three predictors (sharpness, tonality and fluctuation strength).

Two information criteria are reported below that measure the goodness-of-fit of each model. They are Akaike's Information Criterion (AIC)[27] and Schwartz's Bayesian Information Criterion (BIC).[28] Higher numbers indicate less preferable models. Both criteria increase with increasing deviance and include a penalty term based on the number of parameters; simpler models are preferred. The BIC depends on the sample size, so there is also a penalty for a higher number of observations. In addition to the two information criteria, the adjusted R^2 coefficient of determination is also given for each regression analysis.

3. RESULTS

A. SIMPLE MODELS

The mean annoyance for each subject is shown in Figure 2. These values are found using the simplest multilevel model, one with only a variable intercept based on subject ($M1$); no SQMs were used as predictors. This model gives the mean annoyance for each subject as $\gamma_{00} + u_{0j}$, in which γ_{00} is the overall mean and u_{0j} is the estimate of the best linear unbiased predictor (BLUP). This represents the partial pooling result. The mean annoyance for each subject with no pooling is slightly different, but the results differ by less than 0.04 for all subjects. The differences are typical of partial pooling where the largest $|u_{0j}|$ gives the largest difference from the no pooling result, an effect known as shrinkage.[11]

Some details of $M1$ are given in Table 2. The fixed part shows the parameters common to all subjects, which for this model, is simply γ_{00}. The random part includes the variance of the stimulus level error, σ_e^2, and the variance of the variable intercept, $\sigma_{u_0}^2$. The mean annoyance values for individual subjects vary between 2.7 and 9.6, which is part of a normal distribution with variance $\sigma_{u_0}^2$. This is different from the 95% confidence interval for γ_{00}, which is $(5.76; 6.75)$.

The large range in intercepts with no SQMs included in the model points to nonacoustic factors as indicators of annoyance. For example, noise sensitivity has been shown in other work to have an important effect in the model.[14,15] The influence of subject-specific factors may at first seem to be a barrier to building a model based on acoustic factors. However, the use of a subject-specific intercept term weakens this barrier and allows the effects due to changes in sound quality to be better modeled.

Previous analysis of the RoQM data indicated that sharpness, tonality and fluctuation strength were important indicators of annoyance to helicopter noise.[1] This was found by including loudness, sharpness, tonality, fluctuation strength and impulsiveness as indicators, measuring the R^2 value, subtracting one metric at a time, and then calculating the difference in R^2 with that metric removed. A large reduction in R^2 for a particular metric indicated its importance in the model. A similar one-off approach is shown in Figure 3.

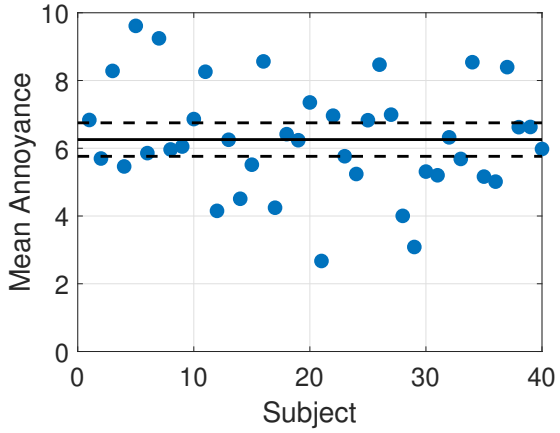

Figure 2: Mean annoyance for each subject ($\gamma_{00} + u_{0j}$) found using a variable intercept model ($M1$). Solid line: mean of all subjects (γ_{00}). Dashed lines: 95% confidence intervals.

Table 2: Results from the variable intercept, no metric model ($M1$). Values in parentheses are the 95% confidence intervals.

Model	$M1$
Fixed part	
γ_{00} (Intercept)	6.26 (5.76; 6.75)
Random part	
Stimulus level σ_e^2	2.87 (2.75; 2.99)
Individual level $\sigma_{u_0}^2$	2.52 (1.62; 3.92)
Explained variance	
Stimulus level R^2	0.47
AIC	16532
BIC	16551

The reduction of R^2 is shown for three different models (as defined in Table 1):

- $M0$: both the intercept and slope are fixed (i.e., the same for all subjects),
- $M2$: variable (i.e., subject-specific) intercept with fixed slope and
- $M3$: variable intercept and variable slope.

For example, the largest difference is shown for $M3$ with sharpness removed, which is given by the difference in R^2 between $M3$.NSTFI and $M3$.NTFI. It is shown that sharpness, tonality and fluctuation strength show the largest reductions in R^2 for $M0$, $M2$ and $M3$, which agrees with previous analysis.[1] The ranking of importance is also consistent for the three model types. Low reductions in R^2 for loudness demonstrate that loudness was effectively controlled for; presenting sounds of similar loudness caused subjects to listen to other differences in the sounds. Low reductions in R^2 for impulsiveness may be due to the fact that sounds with high enough impulsiveness were not tested. Roughness was not included in this analysis for two reasons: first, perturbation methods could not be found that changed roughness while keeping other metrics roughly constant and second, roughness was highly dependent on impulsiveness.

While loudness appears to have been effectively controlled for based on Figure 3, further evidence of this is shown by the resulting regression slopes, which give the change in annoyance for a given change in sound quality, i.e.,

$$Slope = \frac{change\ in\ annoyance}{change\ in\ sound\ quality} \quad . \tag{4}$$

For the k-th sound quality metric in a model, the mean slope for all subjects is γ_{k0}, and the estimate of the slope for the j-th subject is $\gamma_{k0} + u_{kj}$. The regression slopes for loudness are shown in Figure 4, which are all less than $0.2/\text{sone}$. This is a direct consequence of making all test stimuli similar in loudness levels. All 105 sounds presented to the subjects were between 14 and 24 sones (49 and 63 dBA). If loudness were to vary freely, it is expected to be the dominant indicator of annoyance,[7,26] which would result in higher reduction of R^2 in Figure 3 as well as higher regression slopes in Figure 4.

Having controlled for loudness, the change in annoyance due to other SQMs can be better understood. The regression slopes for sharpness, tonality and fluctuation strength are found to be statistically significant indicators of annoyance ($p < 0.05$). These results are shown in Figure 5. Similar to Figures 3 and 4, models

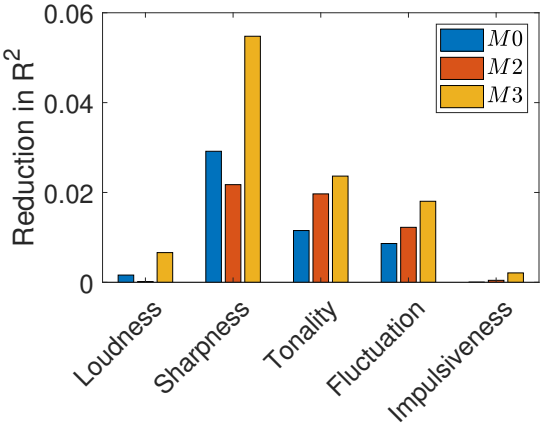

Figure 3: Reduction in R^2 by removing one metric, starting with loudness, sharpness, tonality, fluctuation strength (fluctuation) and impulsiveness. Larger reductions indicate a metric's importance.

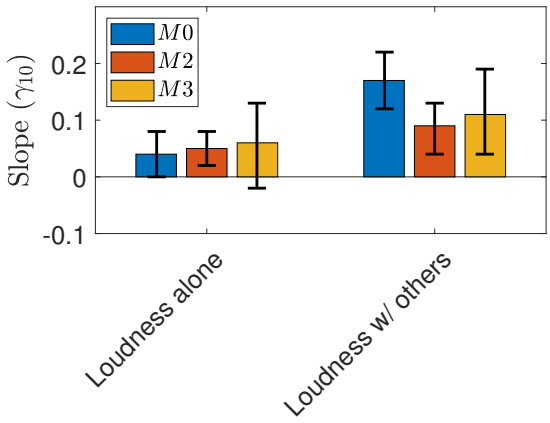

Figure 4: Regression slopes with confidence intervals for loudness when it is the only predictor in the model and when it is included with sharpness, tonality, fluctuation strength and impulsiveness in the same model.

with different combinations of fixed/variable intercept and slope are shown. The slopes for models with different number of SQMs are also shown, which was done to test for dependencies among the metrics.

For Figure 5(a), the slope outputs from models $M0$.S, $M0$.T, $M0$.F, $M2$.S, $M2$.T, $M2$.F, $M3$.S, $M3$.T and $M3$.F are shown. For these models, sharpness, tonality and fluctuation strength are considered *individually* for each multilevel analysis, $M0$, $M2$ and $M3$.

For Figure 5(b), the slope outputs from models $M0$.STF, $M2$.STF and $M3$.STF are shown. In these models, sharpness, tonality and fluctuation strength are considered *simultaneously* as predictors in each multilevel analysis, $M0$, $M2$ and $M3$.

For Figure 5(c), the slope outputs from models $M0$.NSTFI, $M2$.NSTFI and $M3$.NSTFI are shown. For these models, loudness, sharpness, tonality, fluctuation strength and impulsiveness are considered *simultaneously* as predictors in each multilevel analysis, $M0$, $M2$ and $M3$.

For the slopes shown in Figure 5, the averages are $1.4/\text{acum}$ for sharpness, $2.0/\text{TU}$ for tonality and $1.8/\text{vacil}$ for fluctuation strength, values that are relatively consistent across all models. This consistency is evidence for treating each metric independently when calculating the change in annoyance for a given change in sound quality.

The statistics of the multilevel models including sharpness, tonality and fluctuation strength together are shown in Table 3. These correspond to the regression slopes found in Figure 5(b). With higher R^2 and lower AIC and BIC, these models are an improvement over $M1$ shown in Table 2. Table 3 shows that variable slopes is a more accurate model than the fixed slope model. Although the mean slopes γ_{10}, γ_{20} and γ_{30} do not change much when the variable slopes are included, the 95% confidence interval does slightly increase. Subject-specific slopes can be found from $M3$.STF, but if overall mean slopes are needed, $M2$.STF gives almost the same result.

Although the variable slopes in $M3$.STF are significant, it was also investigated if the variable slope was more or less important for some metrics than it was for others. Keeping a variable slope for sharpness and fluctuation strength but making it fixed for tonality yields a model denoted as $M3.\text{S}\overline{\text{T}}\text{F}$. The reduction in R^2 from $M3$.STF to $M3.\text{S}\overline{\text{T}}\text{F}$ was only 0.01, while the reduction by fixing the slope of sharpness ($M3.\overline{\text{S}}\text{TF}$) or fluctuation strength ($M3.\text{ST}\overline{\text{F}}$) was 0.03. Therefore, it was found that the variable slope of sharpness and fluctuation strength was more important than that for tonality.

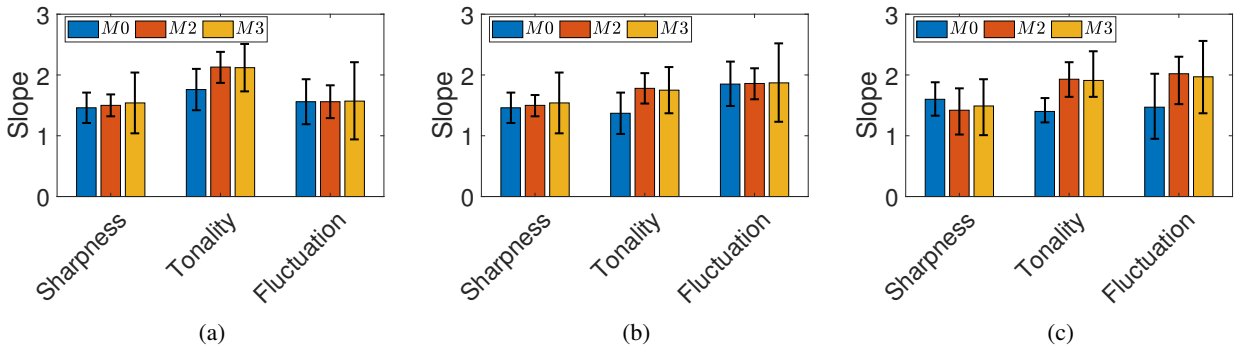

Figure 5: Regression slopes for sharpness, tonality and fluctuation strength (fluctuation) when (a) analyzed individually in each model (b) analyzed simultaneously in each model and (c) analyzed simultaneously with loudness and impulsiveness in each model.

Table 3: Results from the variable intercept and fixed slope model and the variable intercept and variable slope model, both with sharpness, tonality and fluctuation strength as predictors ($M2.\mathrm{STF}$ ***and*** $M3.\mathrm{STF}$***, respectively). Values in parentheses are the 95% confidence intervals.***

Model	$M2.\mathrm{STF}$	$M3.\mathrm{STF}$
Fixed part		
γ_{00} (Intercept)	6.26 (5.76; 6.76)	6.26 (5.76; 6.77)
γ_{10} (Sharpness)	1.50 (1.32; 1.67)	1.54 (1.04; 2.04)
γ_{20} (Tonality)	1.78 (1.53; 2.03)	1.75 (1.37; 2.13)
γ_{30} (FluctuationStrength)	1.86 (1.60; 2.11)	1.87 (1.23; 2.52)
Random part		
Stimulus level σ_e^2	2.45 (2.35; 2.56)	2.12 (2.03; 2.21)
Individual level $\sigma_{u_0}^2$	2.58 (1.66; 4.01)	2.60 (1.67; 4.04)
Individual level $\sigma_{u_1}^2$	-	2.33 (1.42; 3.81)
Individual level $\sigma_{u_2}^2$	-	0.95 (0.47; 1.90)
Individual level $\sigma_{u_3}^2$	-	3.73 (2.24; 6.21)
Explained variance		
Stimulus level R^2	0.55	0.61
AIC	15883	15496
BIC	15921	15553

Given that Figure 2 shows that a variable intercept is important and that Table 3 shows variable and fixed slope models give similar slope values, it raises the question of whether a variable intercept or variable slope is more important. Table 3 shows that removing the variable slopes for sharpness, tonality and fluctuation strength reduces R^2 by only 0.06. If the variable intercept is removed from $M3$.STF, the reduction in R^2 is 0.46 (not shown), proving that a variable intercept is much more important than the variable slopes.[d] This important result means that although subjects use different parts of the annoyance scale, they do not necessarily use different ranges of the scale. Said in another way, although the absolute annoyance is different among subjects, their changes in annoyance are similar for fixed changes in sound quality.

B. INCLUSION OF 2ND-ORDER TERMS

As already mentioned, a perturbation method that varies roughness while keeping other metrics relatively constant could not be found, and roughness was found to be dependent on impulsiveness. It is reasonable to assume, therefore, that some interactions or dependencies may exist among the metrics.

It was decided to test the statistical significance of 2nd-order or cross-terms. This adds a quadratic term for each metric along with the combinations among the metrics. This is depicted in Table 4 where the combination of each row and column gives a 2nd-order term (here, only the lower triangular matrix is considered to avoid duplicate terms). The statistical model first considered consists of a variable intercept, variable slopes for 1st-order terms (N, S, T, F, I and R) and fixed slopes for all 2nd-order terms.

$$Y_{ij} = \beta_{0j} + \sum_{k=1}^{K} \beta_{kj} X_{ki} + \sum_{k=K+1}^{K+P} \beta_{k0} X_{ki} + e_{ij} \tag{5}$$

in which 1st-order terms correspond to $1 \leq k \leq 6$ and 2nd-order terms correspond to $k > 6$ ($K = 6$ and $P = 21$). From this model, the following 2nd-order terms were found to be statistically significant ($p < 0.05$): N^2, $T \times S$, $F \times N$, $F \times S$, $I \times S$, $I \times T$ and $R \times I$, indicated by checkmarks in Table 4. The other terms were found to be statistically insignificant.

After finding the statistically significant 2nd-order terms, their relative importance was investigated. This was done in a new variable intercept model that contained all 1st-order terms (variable slopes) as well as the statistically significant 2nd-order terms (fixed slopes). Then, the reduction in R^2 was found through the process described earlier by removing one term at a time. The reduction in R^2 for each term is shown in Table 5. The importance of each term are subjectively put into three groups in order of highest to lowest importance (bold, underligned, plain). The most important factors, with large reduction in R^2, are fluctuation strength, sharpness and tonality, which is a different order of importance than found with only 1st-order terms. Factors of secondary importance include impulsiveness, roughness and the impulsiveness/tonality cross-term. Since the reduction for tonality is closer to that for impulsiveness than it is for sharpness, tonality could have been assigned to the second group. However, it is kept with the first group for two reasons: (1) a full range of tonality values was tested (in contrast to impulsiveness) and (2) perturbation methods were found to vary tonality while keeping other metrics relatively constant (in contrast to roughness). Although one model cannot confirm one metric as being dominant, slightly different rankings with the inclusion of 2nd-order terms does indicate that interaction effects could be important and may be considered in future analysis or new studies.

Finally, it was found that including 2nd-order terms increases the accuracy of the multilevel model. However, this is on the border of statistical versus practical significance. While increased accuracy is statistically significant, it is more practical to think about regression slopes only for 1st-order SQMs. For example, it is difficult to interpret the meaning of the change in annoyance for a unit change in the product of impulsiveness and tonality.

[d] This is also the reason a model with fixed intercept and variable slope is not included in Table 1.

Table 4: Lower triangular matrix of 2nd-order terms considered in a statistical model (variable intercept, variable slopes for 1st-order terms and fixed slopes for 2nd-order terms). Checkmarks indicate second-order terms found to be statistically significant.

	N	S	T	F	I	R
N	✓					
S						
T		✓				
F	✓	✓				
I		✓	✓			
R					✓	

Table 5: Reduction in R^2 when removing a 1st- or 2nd-order term. Base model included 1st-order terms and statistically significant 2nd-order terms. Bold entries are most important, and underligned entries are somewhat important.

Removing metric	Reduction in R^2
F	**0.055**
S	**0.034**
T	**0.015**
I	<u>0.012</u>
R	<u>0.010</u>
I×T	<u>0.010</u>
N×N	0.006
I×R	0.005
S×I	0.003
N×F	0.002
S×T	0.001
S×F	0.001

4. CONCLUSION

A psychoacoustic listening test was carried out that consisted of 105 synthesized helicopter-like sounds. Sound quality metrics were varied in a systematic way in order to capture most of the variation found from field recordings. The noise stimuli were normalized in terms of loudness. The annoyance responses from 40 subjects were collected. Multilevel analyses were performed that consisted of a noise stimulus lower level and a second level where the responses were grouped by subject number. It was found that although a subject-specific intercept plays a large role, SQMs may be used as predictors of annoyance to helicopter noise. Sharpness, tonality and fluctuation strength were found to be important predictors and showed consistent slopes for a variety of models tested. For unit changes of a sound quality metric, a variable slope had a larger effect for sharpness and fluctuation strength than it did for tonality. Interactions among the metrics may be important, as indicated by increased accuracy when including cross-terms. However, the utility of such higher-order terms cannot be confirmed and may be unique to each data set.

ACKNOWLEDGMENTS

The authors would like to thank Regina Johns and Erin Thomas for the recruitment of test subjects and performing the audiometric testing. Also, Jonathan Rathsam and Jasme Lee have contributed important insight into multilevel analysis.

REFERENCES

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

Illinois, August 2018.

[2] K.J. Faller II, S.A. Rizzi, and A.R. Aumann. Acoustics performance of a real-time three-dimensional sound reproduction system. Technical Report NASA TM-2013-218004, National Aeronautics and Space Administration, 2013.

[3] J.W. Leverton. Helicopter noise: can it be adequately rated? *J. Sound and Vib.*, 43(2):3, 1975.

[4] P.D. Schomer. A survey of community attitudes towards noise near a general aviation airport. *J. Acoust. Soc. Am.*, 74(6):1773–1781, 1983.

[5] P.D. Schomer and R.D. Neathammer. The role of helicopter noise-induced vibration and rattle in human response. *J. Acoust. Soc. Am.*, 81(4):966–976, 1987.

[6] Vincent Mestre, Sanford Fidell, Richard D. Horonjeff, Paul Schomer, Aaron Hastings, Barbara G. Tabachnick, and Fredric A. Schmitz. *Assessing Community Annoyance of Helicopter Noise*. Transportation Research Board, nov 2017.

[7] A.L. McMullen. Assessment of noise metrics for application to rotorcraft. Master's thesis, Purdue University, 2014.

[8] H. Fastl and E. Zwicker. *Psycho-Acoustics: Facts and Models*. Springer-Verlag, third edition, 2007.

[9] Richard H. Lyon. *Designing for product sound quality*. Marcel Dekker, Inc., 2000.

[10] S.A. Rizzi. Toward reduced aircraft community noise impact via a perception-influenced design approach. In *Inter.Noise*, Hamburg, 2016.

[11] A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, Cambridge, UK, 2007.

[12] J.J. Hox. *Multilevel Analysis: Techniques and Applications*. Routledge, 2nd edition, 2010.

[13] C.G.M. Groothuis-Oudshoorn and H.M.E. Miedema. Multilevel grouped regression for analyzing self-reported health in relation to environmental factors: the model and its application. *Biometrical Journal*, 2006.

[14] A. Trollé, C. Marquis-Favre, and A. Klein. Short-term annoyance due to tramway noise: determination of an acoustical indicator of annoyance via multilevel regression analysis. *Acta Acustic united with Acustica*, 100:34–45, 2014.

[15] L.-A. Gille, C. Marquis-Favre, and C. Weber. Aircraft noise annoyance modeling: consideration of noise sensitivity and of different annoying acoustical characteristics. *Applied Acoustics*, 115:139–149, 2017.

[16] Beat Schäffer, Sabine J. Schlittmeier, Reto Pieren, Kurt Heutschi, Mark Brink, Ralf Graf, and Jürgen Hellbrück. Short-term annoyance reactions to stationary and time-varying wind turbine and road traffic noise: A laboratory study. *J. Acoust. Soc. Am.*, 139(5):2949–2963, may 2016.

[17] D.K. Wilson, C.L. Pettit, N.M. Wayant, E.T. Nykaza, and C.M. Armstrong. Multilevel modeling and regression of community annoyance to transportation noise. *J. Acoust. Soc. Am.*, 142:2905–2918, 2017.

[18] Sanford Fidell, Vincent Mestre, Paul Schomer, Bernard Berry, Truls Gjestland, Michel Vallet, and Timothy Reid. A first-principles model for estimating the prevalence of annoyance with aircraft noise exposure. *J. Acoust. Soc. Am.*, 130(2):791–806, 2011.

[19] J. Rathsam. Multilevel modeling of recent community noise annoyance surveys. In *Proc. Meet. Acoust. (POMA)*, volume 33, 2019.

[20] A. Taghipour, R. Pieren, and B. Schäffer. Short-term annoyance reactions to civil helicopter and propeller-driven aircraft noise: a laboratory experiment. *J. Acoust. Soc. Am.*, 145:956–967, 2019.

[21] . Loudness and sharpness calculation. Technical Report Application Note - 02/18, HEAD Acoustics, 2018.

[22] HEAD Acoustics. Artemis Suite 12.05. Herzogenrath, Germany.

[23] R. Sottek. *Modelle zur Signalverarbeitung im menschlichen Gehör (in German)*. phdthesis, RWTH Aachen, 1993.

[24] W. Aures. Berechnungsverfahren für den sensorischen wohlklang beliebiger schallsignale. *Acta Acustica united with Acustica*, 59(2):130–141, 1985.

[25] E. Greenwood and F.H. Schmitz. Separation of main and tail rotor noise from ground-based acoustic measurements. *AIAA Journal of Aircraft*, 51(2):464–472, 2014.

[26] S.R. More. *Aircraft noise characteristics and metrics*. PhD thesis, Purdue University, West Lafayette, Indiana, December 2010.

[27] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, Tsahkadsor, Armenia, USSR, September 1973.

[28] G. E. Schwartz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.