JASA ARTICLE

Bayesian statistical models for community annoyance survey data

Jasme Lee,[1] Jonathan Rathsam,[2,a)] and Alyson Wilson[3]
[1]National Institute of Aerospace, 100 Exploration Way, Hampton, Virginia 23666, USA
[2]Structural Acoustics Branch, National Aeronautics and Space Administration Langley Research Center, MS 463, Hampton, Virginia 23681, USA
[3]Department of Statistics, North Carolina State University, 2311 Stinson Drive, Campus Box 8203, Raleigh, North Carolina 27695, USA

ABSTRACT:
This paper demonstrates the use of two Bayesian statistical models to analyze single-event sonic boom exposure and human annoyance data from community response surveys. Each model is fit to data from a NASA pilot study. Unlike many community noise surveys, this study used a panel sample to collect multiple observations per participant instead of a single observation. Thus, a multilevel (also known as hierarchical or mixed-effects) model is used to account for the within-subject correlation in the panel sample data. This paper describes a multilevel logistic regression model and a multilevel ordinal regression model. The paper also proposes a method for calculating a summary dose-response curve from the multilevel models that represents the population. The two models' summary dose-response curves are visually similar. However, their estimates differ when calculating the noise dose at a fixed percent highly annoyed. https://doi.org/10.1121/10.0001021

(Received 3 October 2019; revised 12 March 2020; accepted 16 March 2020; published online 13 April 2020)
[Editor: Sanford Fidell] Pages: 2222–2234

I. INTRODUCTION

Community annoyance due to loud and startling sonic booms is the major contributing factor to the existing ban on commercial supersonic flights over land. Since before the ban, NASA and partners have researched how to make these sonic booms quieter, specifically how to strategically design the aircraft to achieve a shaped sonic boom (Maglieri *et al.*, 2014). Lockheed Martin, under contract with NASA, is currently building an experimental aircraft designed to demonstrate this quiet supersonic technology, the X-59 Quiet SuperSonic Technology (QueSST). NASA will use this aircraft to conduct social surveys in order to understand how community members perceive the sounds of quiet supersonic flight. The resulting survey data will support efforts to develop international standards for replacing the current ban on commercial supersonic flight over land with a noise-based limit. For example, a dose-response curve established from the data can be used to predict the degree of annoyance for a population at a particular noise level within the tested noise dose range.

This paper describes how to calculate a summary dose-response curve from two different Bayesian statistical models. Each summary dose-response curve characterizes the population on average. Both models are fit to panel sample data from NASA's pilot community annoyance survey, the Quiet Supersonic Flights 2018 (QSF18) test, where each participant was asked to respond to an annoyance survey multiple times. The two models are the multilevel logistic regression model and the multilevel ordinal regression model. The goal is to demonstrate fitting the two multilevel models to the QSF18 data, and we do not attempt to down-select to either model.

The analysis method is selected based on the potential uses of the models. The primary goal is to estimate a population representative summary dose-response curve. In addition, it is desired to use the models to inform experimental design and planning of future surveys, such as estimating the minimum number of participants required and investigating a sufficient range of noise doses for the sonic boom events.[1] For these two reasons, we choose to use a statistical model over a curve-fitting approach to analyze the data because a statistical model considers the data-generating process with randomness and can be used for statistical inference and prediction.

Because each participant in the study can respond multiple times, a multilevel model is used to model the correlation among the multiple responses from the same participant via individual-level parameters (Fitzmaurice *et al.*, 2012; Hu *et al.*, 1998). The X-59 community surveys will likely also use panel samples because Fidell and Horonjeff (2019) found that telephone surveys without callbacks after each event resulted in a very low survey completion rate that made a cross-sectional sampling approach impractical for sonic boom community response surveys. By contrast, many community noise surveys collect cross-sectional data, which are often assumed to be independent because each participant in the study responds only once. These two types of data, single and multiple responses from each participant, require different statistical modeling techniques. A multilevel model is one

a)Electronic mail: jonathan.rathsam@nasa.gov

https://doi.org/10.1121/10.0001021

appropriate method for analyzing panel sample data,[2] whereas a non-multilevel model is appropriate for cross-sectional data. Using maximum likelihood estimation to find parameter values can be a complicated optimization problem for a model with many parameters, like those proposed in this paper. If the analyst has an efficient multidimensional optimizer, it may be possible to fit these models using maximum likelihood estimation. We instead use a Bayesian inference framework. Bayesian approaches supplement the likelihood with a probability distribution (called the prior distribution) that captures knowledge about the unknown parameters in the model. Our prior distributions will typically be noninformative, as discussed in Sec. II D. The Bayesian specification allows us to use Markov chain Monte Carlo (MCMC) algorithms to perform inference and quantify uncertainty about the model parameters. MCMC frames the inference problem as a sampling problem (how to draw a random sample from the posterior distribution of the unknown parameters) instead of an optimization problem (which values of the parameters maximize the likelihood), and can simplify the computation of parameter estimates and uncertainty intervals for high-dimensional models like those we employ.

A. Literature review

The community noise literature consists of a mix of statistical models and curve-fitting methods for calculating a dose-response curve. One of the first examples is by Schultz (1978), which proposed a single dose-response curve derived from a meta-analysis of multiple transportation noise surveys to describe human annoyance. Schultz (1978) also proposed to use percent highly annoyed to describe the community response, which was later adopted by the United States Environmental Protection Agency (EPA) (U.S. Environmental Protection Agency, 1982) as the impact criterion of noise on communities. The Schultz Curve is an average of multiple third-degree polynomial curve fits, each fit to data from an individual survey from a community noise survey database. One major disadvantage is that the Schultz Curve does not constrain the probability of high annoyance between 0 to 1. The Federal Interagency Committee on Noise (FICON) (Federal Interagency Committee on Noise, 1992) recommends the logistic regression model instead because this statistical model fits similarly to the Schultz Curve while bounding the probability. Another curve-fitting method is proposed by Fidell *et al.* (2011) and Fidell *et al.* (1988), with the specific function: $p = \exp(-A/m)$, where A is the parameter of interest, and m is a transformation of the noise dose based on Stevens' Power Law (Stevens, 1975). This curve bounds the probability and is parsimonious because it only has one free parameter. Instead of fitting to data from each survey separately, Miedema and Vos (1998) proposed two methods to pool together data from multiple surveys. The first is to pool together data from surveys with the same type of transportation noise source (aircraft, railway, or road noise) and fit a quadratic regression model to each of the three datasets. The second is to use a multilevel model to partially pool together all the survey data, and use the type of transportation noise source as the grouping level. The quadratic regression model, with or without the multilevel model structure, suffers the same drawback as the Schultz curve of not bounding the probability of high annoyance. Groothuis-Oudshoorn and Miedema (2006) suggest a multilevel interval regression model, which also models the transportation noise source as the grouping level. It models ordinal responses, such as a response from an ordered 1–5 response scale. Other examples of multilevel modeling of cross-sectional data, where each participant in the study only responds once, are Wilson *et al.* (2017) and Miller *et al.* (2014). While a multilevel model can be used to pool together cross-sectional data from different surveys or communities, it can also be used to model panel sample data as demonstrated with sleep disturbance data and laboratory noise annoyance data by Schäffer *et al.* (2017), Trollé *et al.* (2014), and Gille *et al.* (2016). This paper fits multilevel models to panel sample data from a community annoyance survey, adding an additional application to those cited above, by refining the analysis methods introduced in Rathsam *et al.* (2018a).

II. METHODS

The data necessary for this dose-response analysis are the collected survey responses and the corresponding noise doses. First, the data collection method is briefly described. Then the two statistical models and a general outline for assessing the model fits are described.

A. Field test overview

The Quiet Supersonic Flights 2018 test was conducted in Galveston, Texas on November 5–15, 2018. The quiet sonic booms were produced by an F-18 research aircraft performing a supersonic dive maneuver (Haering *et al.*, 2006) over the Gulf of Mexico, 10 to 20 nautical miles off the coast of Galveston. There were a total of 22 flights and 52 quiet sonic boom events distributed across 9 test days. The data analyzed in this paper are responses from the single-event survey that participants were asked to complete promptly after every event in order to characterize their perceptions of individual occurrences of the sonic booms.

B. Estimated noise dose

The dose-response analysis requires matching each survey response with an estimated noise dose. Noise doses are estimated via a combination of measurement and prediction. Eleven noise monitors were set up across the community survey area to measure the quiet sonic booms. Multiple noise metrics were calculated from the pressure waveforms measured at each monitor. In this paper, quiet sonic boom exposure is quantified in terms of perceived level (PL) (Shepherd and Sullivan, 1991; Stevens, 1972) because PL is one of several noise metrics shown to correlate well with human annoyance in the lab (Rathsam *et al.*, 2018b), and the acoustic requirements for X-59 were written in terms of

PL. Due to the sometimes low signal to noise ratio between the quiet sonic booms and the background noise, measurements were only retained for analysis if the PL of the quiet sonic boom was 5 dB or greater than the PL of the background noise measured immediately preceding the quiet sonic boom. As described by Shepherd and Sullivan (1991), a 650 ms analysis window was used. The acoustic levels calculated from the noise monitor data were interpolated to the participant locations with the aid of sonic boom exposure predictions from PCBoom (Page *et al.*, 2010) given the aircraft trajectory and measured meteorological data. A similar interpolation method can be found in the final report for a previous pilot study (Page *et al.*, 2014).

C. Survey data

This dose-response analysis focuses on the single-event survey responses related to annoyance. About half of the survey participants were randomly assigned to receive reminders, either via text message or email, after the sonic boom events and occasional false reminders. Note that the data analyzed in this paper do not include the false reminders.[3] Each participant was first asked whether he/she heard the event. If the participant reported hearing the event, he/she was then asked to rate his/her annoyance level to the sonic boom. If the participant reported not hearing the event, the survey software skipped the annoyance rating question. The annoyance survey question is phrased: "How much did the sonic boom bother, disturb, or annoy you?" The response scale is a five-point ordinal scale with 1 to 5 corresponding to the following descriptions: *not at all annoyed, slightly annoyed, moderately annoyed, very annoyed*, and *extremely annoyed*. Recall that the convention is to use "percent highly annoyed" as the community response. The recommended cutoff for "high annoyance" on a five-point scale is 4 or above (Fields *et al.*, 2001). All responses for which the participant reported "did not hear the event" are grouped with the "not at all annoyed" responses.

D. Statistical models

Two Bayesian statistical models are fit to the data: the multilevel logistic regression and the multilevel ordinal regression models. These are the two best models from a downselection on seven candidate models fit to a different NASA pilot study dataset (Lee *et al.*, 2019). For both models, the noise dose is assumed to be known precisely without measurement error, and order effects or sequence of the boom events are not considered. The model parameters are estimated using Markov chain Monte Carlo (MCMC) sampling with the software Just Another Gibbs Sampler (JAGS) Version 4.3.0 (Plummer, 2003).

The Bayesian approach considers the observed data to be fixed, and models the parameters to be random given the data. A prior probability distribution is used to describe the uncertainty about the model parameters before observing the data. The prior probability distribution is then updated to the posterior probability distribution using Bayes theorem after observing the data, which are summarized using a likelihood function. The posterior distribution describes the uncertainty about the model parameters after observing the data and is used for statistical inference about the model parameters. For example, the posterior distribution is used to calculate a 95% credible interval,[4] which gives a range of values for which the probability that the unknown model parameter falls in the range is 0.95 (Kruschke, 2014). Bayesian modeling and inference have also found applications in other acoustics research areas (Xiang and Fackler, 2015).

In general, the prior distribution, the likelihood function and the posterior distribution are all multidimensional. Thus, calculation of the summary statistics of the posterior distribution, such as the expected value, require high-dimensional integration, which is often intractable. Markov chain Monte Carlo is a class of algorithms for drawing a random sample from the posterior distribution that can be used to compute integrals of interest using Monte Carlo integration. A realization of a MCMC random sample is called a posterior draw or sample.

1. Model 1: Multilevel logistic regression

The model structure for a multilevel model is hierarchical or nested: there are multiple responses from the same participant, and the participants are all sampled from the same community. Both multilevel models proposed in this paper are random intercept models[5,6]: each individual is modeled to have his/her own intercept β_{0i} instead of the same β_0, while sharing the same slope, β_1. The random intercepts account for the baseline differences among the participants, whereas the shared slope indicates that the effect of noise dose on annoyance is assumed to be equivalent for all participants. The random intercepts model the correlation among responses from the same participant, and are assumed to come from a common distribution: $\beta_{0i} \sim N(\beta_0, \sigma^2)$.

The multilevel logistic regression model requires dichotomizing the ordinal responses to binary responses: 0 for annoyance ratings of 1 to 3 or "not highly annoyed" responses, and 1 for annoyance ratings of 4 to 5 or "highly annoyed" responses. Let H be the binary response; p be the probability of high annoyance; $i \in 1, \ldots, S$ be the set of participant indices; $j \in 1, \ldots, n_i$ be the set of observation indices for participant i, where n_i indicates the total number of responses from subject i. Equation (1) is the multilevel logistic regression model. The first three lines describe the standard multilevel logistic regression model, and the last three describe the noninformative prior distributions assigned to the model parameters β_0, β_1, and σ^2. Noninformative prior distributions are used because little is known about the model parameters *a priori*, and so the prior distributions should contribute little information compared to the data. In other words, the prior distributions should be flat relative to the likelihood. Without substantial engineering evidence, the prior distributions should not restrict the

https://doi.org/10.1121/10.0001021

model parameter values, but instead allow for a large range of possible values.

$$
\begin{aligned}
&H_{ij}|p_{ij} \sim Bernoulli(p_{ij})\\
&p_{ij}|\beta_{0i},\beta_1 = logit^{-1}(\beta_{0i}+\beta_1 PL_{ij})\\
&\beta_{0i}|\beta_0,\sigma^2 \sim N(\beta_0,\sigma^2)\\
&\beta_0 \sim N(0,100)\\
&\beta_1 \sim N(0,100)\\
&\sigma^2 \sim InverseGamma(0.01,0.01).
\end{aligned}
\tag{1}
$$

The notation $H|p$ denotes the binary response given the probability of high annoyance, $H \sim Bernoulli(p)$ denotes H is distributed as a Bernoulli random variable with parameter p, and $\beta \sim N(0,100)$ denotes β is distributed as a normal random variable with mean of 0 and variance of 100.[7] Note that the second line of Eq. (1) can be rewritten as $\log[p_{ij}/(1-p_{ij})] = \beta_{0i}+\beta_1 PL_{ij}$, and log indicates natural logarithm rather than logarithm of base 10.

An example of a Bernoulli (0/1) random variable is the outcome of a random coin flip. These random coin flips are modeled with the probability of landing heads dependent on the PL and the participant via β_{0i}. With the individual-level parameters, each participant has his/her own dose-response curve.

2. Model 2: Multilevel ordinal regression

In contrast to the previous model, the response variable for the ordinal regression model is the ordered five-point annoyance rating. Suppose that the ordinal responses are mapped to a continuous latent variable with range $(-\infty,\infty)$. Each ordinal level is mapped to an interval on the latent variable scale, analogous to how a letter grade can be mapped to a range of numerical scores. The interval thresholds are unknown and estimated from the data as if a grade curve was applied to the numerical scores (i.e., a curved letter grade A corresponds to 85%–100%, B to 75%–84%, etc.). Let Y be the ordinal response, and Y^* be the latent variable. Mathematically, the relationship between Y^* and Y is

$$Y_{ij} = k \text{ if } \gamma_{k-1} < Y^*_{ij} \le \gamma_k \text{ for } k = 1,\ldots,5,$$

where k is the ordinal response, γ_{k-1} corresponds to the lower threshold, and γ_k corresponds to the upper threshold on the latent variable scale.[8] The first and last thresholds are $\gamma_0 = -\infty$ and $\gamma_5 = \infty$. Note that the thresholds or gamma parameters are strictly increasing: $\gamma_0 < \gamma_1 < \cdots < \gamma_5$.

The model specifies that the latent variable is linearly related to the noise dose: $Y^*_{ij} \sim N(\beta_{0i}+\beta_1 PL_{ij},1)$. The variance of Y^* is fixed to 1, and γ_1 is fixed to 0 in order to make the model identifiable (Long, 1997). Figure 1 shows the relationship among the ordinal variable (Y), the latent variable (Y^*) and the covariate (PL). In this illustration, the line on the right representing the expected value of Y^*, $E(Y^*)$, has a nonzero slope. Therefore, the expected value of Y^* changes as PL changes, causing the normal distribution on the left to shift. The normal distribution is essential for deriving the probabilities of each ordinal level, denoted by $\boldsymbol{\pi}$ in the model given in Eq. (2). For example, a rating of 3 corresponds to the latent variable, Y^*, falling between γ_2 and γ_3, so

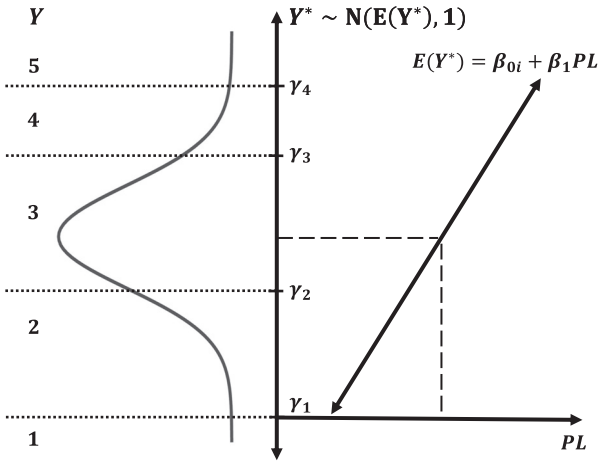

FIG. 1. Relationship among the ordinal variable (Y), latent variable (Y^*), and covariate (PL). This is a modified version of FIG. 23.6 from Kruschke (2014). Each ordinal value is mapped to an interval of Y^* values. The latent variable is modeled to have a linear relationship with PL. When the PL changes, the expected value of Y^* or the mean of the normal distribution on the left also changes. The probabilities of each ordinal level are calculated using the normal distribution on the left.

$$
\begin{aligned}
P(Y_{ij}=3) &= P(\gamma_2 < Y^*_{ij} \le \gamma_3)\\
&= \Phi(\gamma_3 - E(Y^*_{ij})) - \Phi(\gamma_2 - E(Y^*_{ij})).
\end{aligned}
$$

Note that $E(Y^*_{ij}) = \beta_{0i}+\beta_1 PL_{ij}$, and Φ is the standard normal cumulative distribution function.

The multilevel ordinal regression model is given in Eq. (2). An example of a multinomial random variable is the outcome of tossing a five-sided die. In this context, these random tosses, where outcomes of 1 to 5 have a clear ordering, are modeled and the probabilities of landing on each side depend on PL and the individual. The probabilities of landing on each of the five sides is given in $\boldsymbol{\pi}$, so $\boldsymbol{\pi}$ has five elements that sum to 1. The parameters are again assigned noninformative prior distributions [last four lines of Eq. (2)].

$$
\begin{aligned}
&Y_{ij}|\pi_{ij} \sim Multinomial(1,\boldsymbol{\pi_{ij}})\\
&\boldsymbol{\pi_{ij}}|\beta_{0i},\beta_1,\gamma_2,\ldots,\gamma_4 = \begin{bmatrix}
\Phi(0-E(Y^*_{ij}))\\
\Phi(\gamma_2-E(Y^*_{ij}))-\Phi(0-E(Y^*_{ij}))\\
\Phi(\gamma_3-E(Y^*_{ij}))-\Phi(\gamma_2-E(Y^*_{ij}))\\
\Phi(\gamma_4-E(Y^*_{ij}))-\Phi(\gamma_3-E(Y^*_{ij}))\\
1-\Phi(\gamma_4-E(Y^*_{ij}))
\end{bmatrix}\\
&\beta_{0i}|\beta_0,\sigma_0^2 \sim N(\beta_0,\sigma_0^2)\\
&\beta_0 \sim N(0,100)\\
&\beta_1 \sim N(0,100)\\
&\gamma_k \sim N(0,10) \text{ for } k=2,3,4\\
&\sigma_0^2 \sim InverseGamma(0.01,0.01).
\end{aligned}
\tag{2}
$$

3. Model assessment: Posterior predictive checking

After fitting each model, posterior predictive checking is used to assess the fit of each by checking whether the data replicated using each model are similar to the observed data. Model checking is an important step in the model fitting process because it helps determine whether the selected model class is appropriate for the data. For example, a linear regression model is not an appropriate model class if the data exhibit a highly nonlinear pattern.

The types of posterior predictive checking introduced in this paper are discrepancy statistics, which are one-number summaries that capture or summarize key features of the data. Gelman *et al.* (2000) emphasize that the choice of discrepancy statistics depends on the problem.[9] For example, the total proportion of responses that are highly annoyed is a suggested discrepancy statistic for this data.

The procedures for posterior predictive checking for each model follow.

(1) Select a set of discrepancy statistics of interest. An example is the total proportion of highly annoyed responses in the data.
(2) Calculate the discrepancy statistics using the observed data. For example, calculate the total proportion of highly annoyed responses in the observed data.
(3) Generate responses using parameter values at each posterior sample, while fixing the vector of noise doses. For example, for the multilevel logistic regression model given in Eq. (1), generate binary responses using the values of β_{0i} and β_1 at every posterior sample and the vector of observed PL values.
(4) Calculate the discrepancy statistics using each set of generated responses.
(5) Compare the discrepancy statistics from the observed data to the histogram of discrepancy statistics from the generated data.

If the observed discrepancy statistic falls outside the middle 95% probability region of the histogram, this indicates lack of fit for the particular data feature and that the data replicated from the model do not match the observed data.

III. RESULTS

A. Data summary

There are a total of 4998 single-event survey responses from 371 participants. Of those, 2194 (43.9%) indicate the event was heard. Figure 2 shows the distribution of the collected ordinal responses, with "not heard" responses grouped with the rating of 1 or "not at all annoyed." About 1% of the observed data is categorized as a "highly annoyed" response (either a 4 or a 5 rating). The data analyzed are included as supplementary material.[10]

Figure 3 shows the distribution of estimated noise doses assigned to each survey response, ranging from 56 to 90 dB PL. Most of the responses correspond to noise doses of about 65 to 85 dB. Note that because not every participant has the same estimated noise dose for the same event, there is a different number of possible responses at each dose.

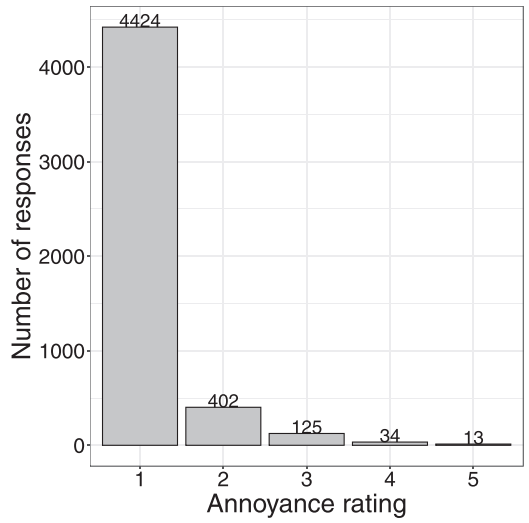

FIG. 2. Distribution of observed ordinal annoyance responses, with "not heard" responses grouped with rating of 1.

Figure 4 shows the number of responses from each participant in the data, split by the reminder groups that the participants were assigned to. Recall that the participants were randomly assigned to either receive reminders or no reminders, and the maximum number of events was 52. About half the participants contribute fewer than 10 responses. As expected, the participants who did not receive reminders responded fewer times than the participants with reminders. Note that some participants may have responded more times, but the response(s) may have been discarded due to either dose estimation problems or the data cleaning process[11] (e.g., failure to estimate a noise dose due to no location available, or an incomplete survey response).

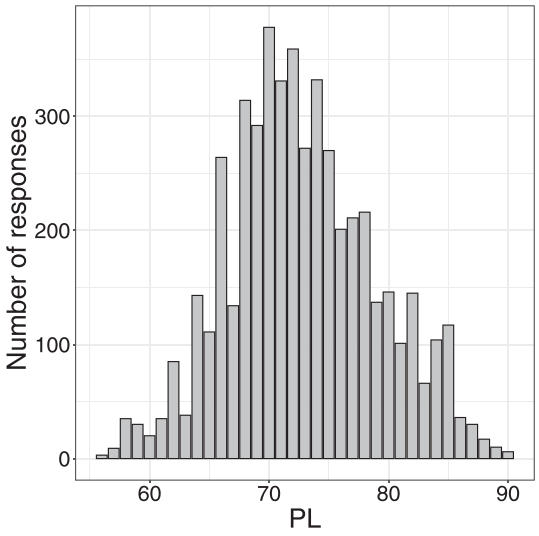

FIG. 3. Distribution of observed noise doses in PL (dB).

https://doi.org/10.1121/10.0001021

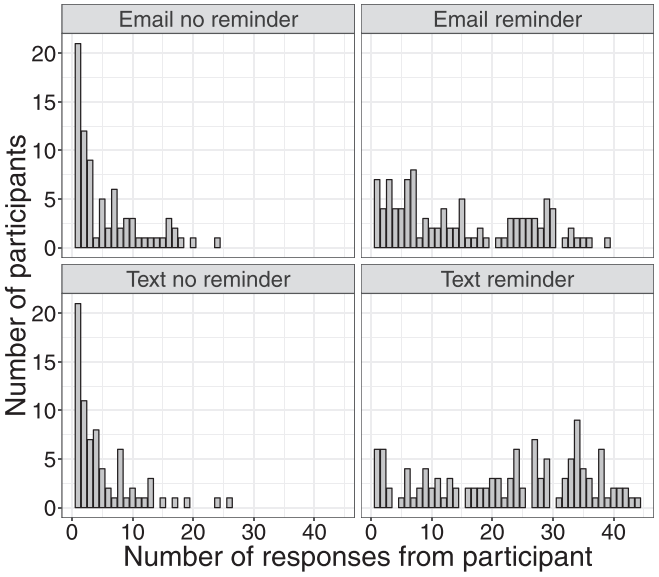

FIG. 4. Distribution of the number of responses from each participant split by the participants' reminder grouping.

B. Parameter estimation

Both models are fit using MCMC sampling. Convergence diagnostic plots, e.g., traceplots and Gelman-Rubin plots (Gelman and Rubin, 1992), are used to check that the chains of MCMC samples have converged to samples from the posterior distribution.[12] For both models, the noise dose is centered by subtracting the mean PL from every PL value in order to improve the efficiency of the MCMC sampling, and to reduce the autocorrelation of the MCMC chains (Kruschke, 2014).

1. Model 1: Multilevel logistic regression

The multilevel logistic regression model parameters are estimated using 400 000 posterior samples after discarding 4000 burn-in samples. The initial draws of an MCMC chain are often discarded as burn-in samples when the chain is moving from an initial value to a high-probability part of the posterior distribution. For this model, we found that 400 000 posterior samples are sufficient to reach convergence according to the traceplots and Gelman-Rubin statistics. Table I shows the posterior summaries of the β_0 and β_1 parameters, which include the posterior mean and 95% credible interval constructed using the 0.025 and 0.975 quantiles.[13] The summary statistics for the 371 β_{0i} parameters are not shown.

TABLE I. Summary statistics of the multilevel logistic regression model parameters β_0 and β_1.

	Mean	SD	0.025 quant.	0.25 quant.	Median	0.75 quant.	0.975 quant.
β_0	−19.0	2.4	−24.1	−20.6	−19.0	−17.3	−14.5
β_1	0.153	0.029	0.098	0.133	0.152	0.172	0.211

Posterior predictive checking is used to assess the model fit following the procedures outlined in Sec. II D 3. Five discrepancy statistics are calculated: the deviance, the proportion of highly annoyed responses, the 0.1 quantile of the PL distribution where highly annoyed responses occur, the mean number of highly annoyed responses per participant, and the number of participants who respond highly annoyed at least once. Deviance is an overall goodness-of-fit statistic; the 0.1 PL quantile is a one-number summary of the PL distribution corresponding to highly annoyed responses; and the last two discrepancy statistics are related to annoyance at the individual level. None of these checks indicate lack of fit.

For a multilevel model, there are two types of dose-response curves: individual and summary. An individual dose-response curve is representative of one participant, and a summary dose-response curve is representative of the population. The summary dose-response curve proposed in this paper is calculated by taking an average of the sampled participants' individual dose-response curves. Therefore, in order to calculate the summary dose-response curve, first the individual dose-response curves are calculated. An individual dose-response curve is calculated at each posterior sample, providing a distribution of individual dose-response curves for each participant. In this case, there are 400 000 individual dose-response curves calculated for each of the 371 participants. To calculate one participant's individual dose-response curve at one posterior sample, first, the observed PL range of 56 to 90 dB is evenly divided into 1000 values. Then, at each of the 1000 PL values, p_{ij} as defined in Eq. (1) is calculated based on the individual's β_{0i}. The curve connecting these 1000 points is the individual dose-response curve for the particular posterior sample.

To calculate a population representative summary dose-response curve, the individual dose-response curves are averaged at each posterior sample, resulting in a distribution of summary dose-response curves. The estimate of the summary dose-response curve is then the curve connecting the pointwise means of the 400 000 averaged curves. The 95% credible intervals are the pointwise 0.025 and 0.975 quantiles.

2. Model 2: Multilevel ordinal regression

The multilevel ordinal regression model parameters are estimated using 50 000 posterior samples after discarding 4000 burn-in samples. Table II shows the posterior summaries of the β_0, β_1 and the three γ parameters. Recall that γ_1 is fixed to 0, so it is not listed.

The same five discrepancy statistics described in Sec. III B 1 are used to assess the model fit of the multilevel ordinal regression model. None of these checks indicate lack of fit.

The first step in calculating the summary dose-response curve for the multilevel ordinal regression model is also to calculate the individual dose-response curves. For the multilevel ordinal regression model, the probabilities for each ordinal level are estimated. So, the probability of high

TABLE II. Summary statistics of the multilevel ordinal regression model parameters β_0, β_1, and γ_k for $k = 2, 3, 4$.

	Mean	SD	0.025 quant.	0.25 quant.	Median	0.75 quant.	0.975 quant.
β_0	−6.80	0.40	−7.61	−7.07	−6.80	−6.53	−6.01
β_1	0.0678	0.0051	0.0578	0.0644	0.0678	0.0712	0.0778
γ_2	0.973	0.045	0.886	0.942	0.972	1.003	1.065
γ_3	1.75	0.08	1.61	1.70	1.75	1.80	1.91
γ_4	2.36	0.12	2.14	2.28	2.36	2.44	2.61

annoyance for individual i can be calculated by adding the estimated probabilities of a 4 or a 5 rating,

$$\begin{aligned} P(Y_{ij} \geq 4) &= P(Y_{ij} = 4) + P(Y_{ij} = 5) \\ &= \left[\Phi(\gamma_4 - E(Y^*_{ij})) - \Phi(\gamma_3 - E(Y^*_{ij}))\right] \\ &\quad + \left[1 - \Phi(\gamma_4 - E(Y^*_{ij}))\right], \\ p_{ij} &= 1 - \Phi(\gamma_3 - E(Y^*_{ij})). \end{aligned} \tag{3}$$

Note that to estimate a different degree of annoyance, such as percent moderately or more annoyed, only the corresponding γ parameter in the last line of Eq. (3) changes. For example, for a dose-response curve estimating the probability of moderate to high annoyance (a response of 3, 4, or 5), $p_{ij} = 1 - \Phi(\gamma_2 - E(Y^*_{ij}))$. The summary dose-response curve is calculated using the method outlined in Sec. III B 1, with p_{ij} as defined in Eq. (3). Figure 5 compares the three summary dose-response curves for three different degrees of annoyance calculated from the multilevel ordinal regression model. The annoyance degree "slightly or more annoyed" corresponds to any rating greater than or equal to 2, "moderately or more annoyed" corresponds to any rating greater than or equal to 3, and "highly annoyed" corresponds to any rating greater than or equal to 4. Note that the sample size at each PL is not displayed.

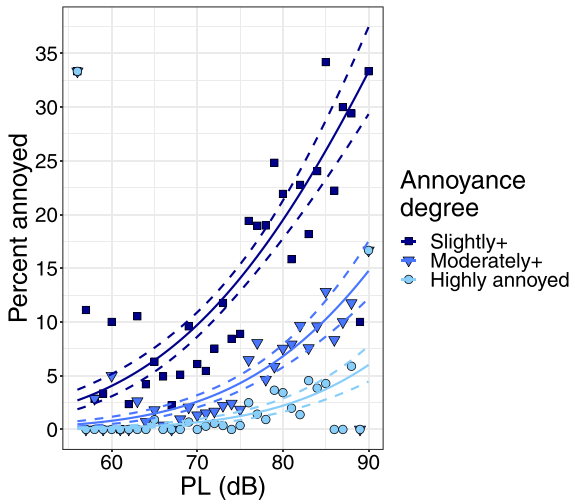

FIG. 5. (Color online) Estimated summary dose-response curves (solid) from the multilevel ordinal regression model compared to observed percentages for three degrees of annoyance, with 95% credible intervals (dashed).

IV. DISCUSSION

A. Comparison of the two models

The goal of this paper is not to downselect from the two multilevel models, both of which are appropriate for this type of panel sample survey data, but to demonstrate and to compare them. Figure 6 compares the summary dose-response curves calculated from the multilevel logistic regression and the multilevel ordinal regression models. The estimate calculated from the multilevel ordinal regression is higher than that from the multilevel logistic regression model. The 95% credible interval for each curve contains the point estimate for the other.

We also examine the practical differences in the two models' estimates by calculating quantities that correspond to two ways the statistical models could be used for regulatory purposes. The first method is to fix the PL value and estimate the percent highly annoyed, and the second is to fix the percent highly annoyed and estimate a corresponding PL value. The following five quantities are calculated[14]:

(1) percent highly annoyed at 65 dB,
(2) percent highly annoyed at 75 dB,
(3) percent highly annoyed at 85 dB,
(4) PL at 1% highly annoyed, and
(5) PL at 2% highly annoyed.

Figure 7 compares the two models' estimates when calculating percent highly annoyed at 65, 75, and 85 dB PL in (a), and PL at 1% and 2% highly annoyed in (b) using violin plots. The violin plots compare the calculated posterior distributions from the two models for each quantity, with the mean marked by the point and 95% credible intervals marked by the bars. Each "violin" consists of the

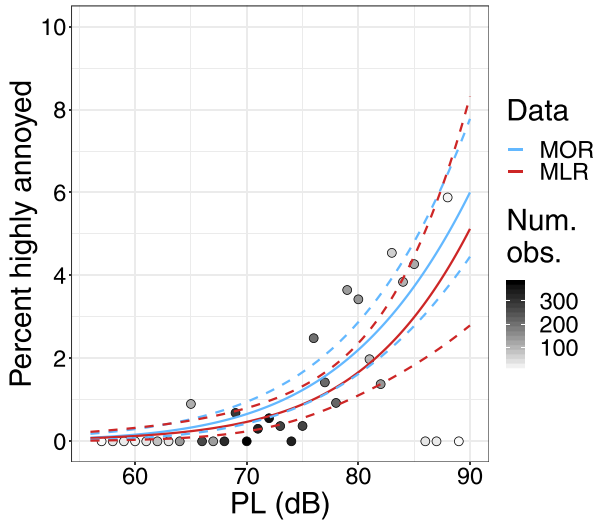

FIG. 6. (Color online) Estimated summary dose-response curves (solid) from the multilevel logistic regression model (MLR) and the multilevel ordinal regression model (MOR) compared to observed percent highly annoyed data, with 95% credible intervals (dashed). Shading of points indicates the sample size at each dose. Two points are not shown in this plot: 33% at 56 dB (1 of 3 observations is highly annoyed), and 16% at 90 dB (1 of 6 observations is highly annoyed).

https://doi.org/10.1121/10.0001021

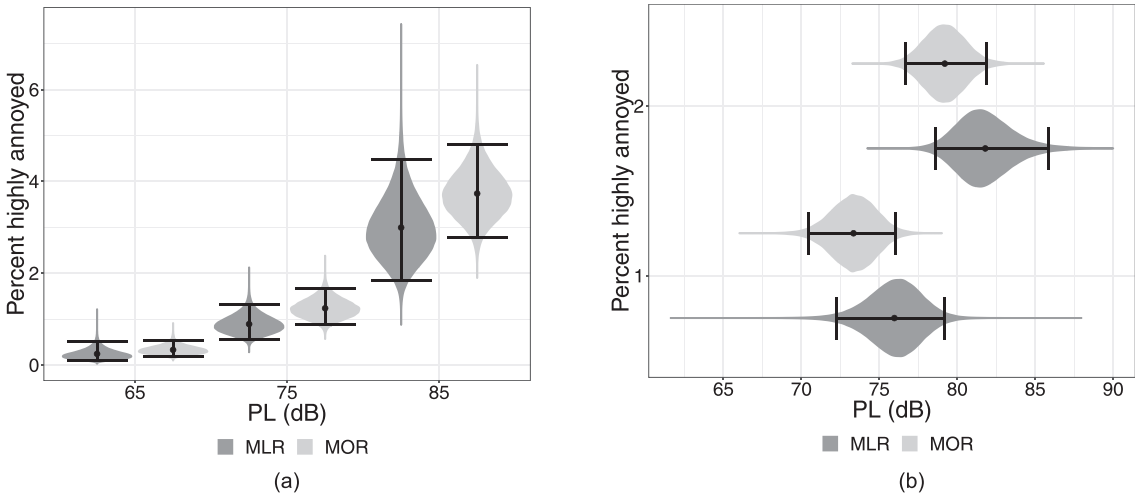

FIG. 7. Comparison of the posterior distributions for (a) percent highly annoyed fixing PL at 65, 75, and 85 dB, and (b) PL fixing percent highly annoyed at 1% and 2% calculated from the multilevel logistic regression (MLR) and the multilevel ordinal regression (MOR) models. The points indicate the point estimates, which are the posterior means, and the bars indicate the 95% credible intervals.

quantity calculated at each posterior sample for the particular model (i.e., for the multilevel logistic regression model, the "violin" for the percent highly annoyed at 65 PL dB consists of 400 000 data points). The two models' estimates of percent highly annoyed at fixed PL values are very similar with no practical difference. The largest difference is about 0.7% in the estimates at 85 dB, with an estimate of 3% highly annoyed from the multilevel logistic regression model and 3.7% from the multilevel ordinal regression model. The 95% credible intervals for the estimates of PL at fixed percent highly annoyed from the two models overlap, with each credible interval containing the mean from the other model. The point estimates differ by about 2.5 dB PL, which is a practical difference. A difference in residential noise exposure of 2.5 dB PL may or may not be audible. However, making an already quiet supersonic aircraft an additional 2.5 dB quieter would require significant additional research and development for an aircraft designer.

Although we detect practical differences when estimating PL at fixed percent highly annoyed values, there is no metric for directly comparing the goodness-of-fit of the two models. A common model comparison metric, the deviance information criterion (DIC), cannot be used to compare the relative fits of these two multilevel models because they are fit to different data: ordinal 1–5 ratings for the multilevel ordinal regression model and binary 0/1 data for the multilevel logistic regression model.

The multilevel logistic regression model is advantageous in that it is easier to interpret and to understand. It also requires fewer modeling assumptions than the multilevel ordinal regression model. However, the multilevel ordinal regression model makes use of the available data more fully as it models the ordinal responses and does not require dichotomizing the data, which leads to a loss of information. The multilevel ordinal regression model can also be used to easily calculate dose-response curves for different degrees of annoyance, which may be of interest if the number of highly annoyed responses becomes too sparse or rare. This is an advantage of the multilevel ordinal regression model compared to the logistic regression model because the logistic regression model must be refit when the dichotomization rule for annoyance changes. On the other hand, the same ordinal regression model fit can be used to estimate a dose-response curve for lower degree of annoyance without refitting the model. For example, instead of percent highly annoyed, the dose-response curve can be calculated for percent moderately to highly annoyed using the same statistical model fit. For the logistic regression model, on the other hand, the binary data are different and so the entire process from model fitting to dose-response calculations needs to be repeated for the new data. The goal of this paper is to describe and fit the two multilevel models to the QSF18 data rather than to downselect to either of the two models. Further analysis is needed if downselection is of interest. For example, a simulation study can be conducted to explore the bias of each model and to check the coverage of each model's credible intervals.

B. Comparison to previous pilot study

Another pilot study similar to QSF18, known as Waveforms and Sonic Boom Perception and Response (WSPR2011) study, was conducted in 2011 at Edwards Air Force Base (EAFB) (Page *et al.*, 2014). The key differences between WSPR2011 and QSF18 are: there were no reminder messages sent in WSPR2011 so all survey responses are assumed to be reported because the participant heard the event, the ordinal response scale for the annoyance survey question for WSPR2011 was from 0 to 10, and WSPR2011 had an order of magnitude fewer participants (49 versus 371). For the 11-point scale, a response of 8, 9, or 10 is considered as "highly annoyed" (Fields *et al.*, 2001). Recall that the ordinal responses scale for the annoyance survey question for QSF18 was from 1 to 5, and a response of a 4 or 5 is considered as "highly annoyed." In addition, the observed

PL range is different: the PL range for WSPR2011 was 63 to 106 dB whereas the PL range for QSF18 was 56 to 90 dB. The planned quiet sonic booms in WSPR2011 were also simulated using the F-18, and there were additional unplanned higher amplitude booms from the U.S. Air Force. Lee *et al.* (2019) describes a similar analysis of the single-event dose-response data from WSPR2011, and downselects to two models from seven candidate models, which are the multilevel logistic regression model and the multilevel ordinal regression model.

The multilevel ordinal regression model is refit to a subset of the "heard event" responses from QSF18 in order to compare with the WSPR2011 data, where all responses are assumed to be "heard." Thus, the data analyzed in Secs. III B 1 and III B 2 are different. Figure 8 compares the multilevel ordinal regression summary dose-response curves for the two studies. Note that they describe the percent highly annoyed, and the sample size at each dB is not shown in this plot. For the QSF18 data, the percent highly annoyed of 100% at 56 dB is not shown in order to focus on the majority of the data, which fall below 45%. The 100% at 56 dB represents only one "heard" response. The two dose-response curves are similar at the lower PL values. The difference in the slope of the WSPR2011 curve at high PL is likely driven by some of the "high annoyance" responses between 90 and 106 dB PL. Although the EAFB community members in the WSPR2011 study were acclimated to hearing higher amplitude sonic booms, they responded similarly to the Galveston community members in the QSF18 study. In the future, we would like to fit one model to the two datasets combined instead of fitting a separate model for each study.

C. Limitations and Future Work

For the future X-59 community tests, it is expected that panel sampling will be used again and testing will occur in multiple communities. Methods for combining results from multiple community studies to derive a single nationally representative summary dose-response curve have not been developed. The two multilevel models described could be extended by adding an additional level for the community in the model hierarchy similar to how Gille and Marquis-Favre (2019) consider multiple responses from participants from multiple studies. This approach would allow for a straightforward generalization to calculate one nationally representative summary dose-response curve. However, a variety of modeling approaches should be considered based on the specifics of the multiple community survey. The statistical models may then be used to estimate the minimum sample size requirements for the X-59 community surveys.

The current assumptions for the statistical models follow:

- no ordering effects based on boom sequence,
- no uncertainty in noise dose estimate,
- the effect of noise dose on annoyance is equivalent for all participants,
- data are missing at random,[15] and
- responses are not affected by the use of reminder messages.

Future work would be required to assess the validity of these assumptions.

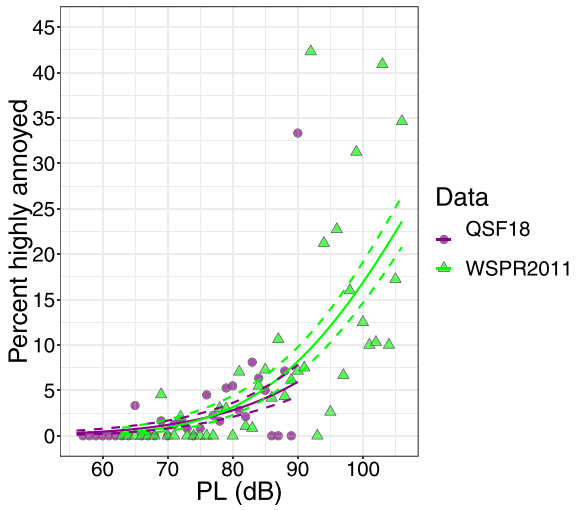

FIG. 8. (Color online) Estimated summary dose-response curves (solid) from the multilevel ordinal regression model for QSF18 "heard event" only data and WSPR2011 compared to observed percent highly annoyed data for each study, with 95% credible intervals (dashed). One point from the QSF18 data is not shown in this plot: 100% at 56 dB (calculated from 1 observation).

V. CONCLUDING REMARKS

For the panel sample data collected in Quiet Supersonic Flights 2018 (QSF18) test, both the multilevel logistic regression and multilevel ordinal regression models are appropriate statistical models for analyzing the data. A summary dose-response curve representative of the community can be calculated by averaging the sampled individuals' dose-response curves. The two models' estimates are very similar when estimating percent highly annoyed at fixed PL. However, the two models' estimates are practically different when estimating the PL at fixed percent highly annoyed because the point estimates differ by about 2.5 dB, which would be a difficult reduction in PL to achieve for an already quiet supersonic aircraft. Further research is needed if downselection of models is of interest.

ACKNOWLEDGMENTS

This work is funded by the NASA Commercial Supersonic Technology Project, and performed in the Structural Acoustics Branch at the NASA Langley Research Center. The authors would like to thank the National Institute of Aerospace for administering a Graduate Research Assistantship and post-graduate funding for J.L.

APPENDIX A: JAGS MODEL SPECIFICATIONS

In order to fit each of the models, the user must supply the JAGS model specification. Below is the JAGS model specification for the multilevel logistic regression model.

https://doi.org/10.1121/10.0001021

```
model{
# Model – multilevel logistic
  # regression model
  # likelihood
  for(j in 1:n){
    Y[j]~dbern(p[j])
    logit(p[j]) = beta_0i[subj[j]] + beta_1 * X[j]
  }

  # priors
  for(i in uniq_subj_id){
      beta_0i[i]~dnorm(beta_0, tau)
  }

  beta_0~dnorm(0, 1/100)
  beta_1~dnorm(0, 1/100)
  tau~dgamma(0.01, 0.01)
  sigma2 = 1/tau
}
```

The JAGS model specification for the multilevel ordinal regression model is based on the examples from Chapter 23 of Kruschke (2014). The model specification is shown below.

```
model{
#Model – multilevel ordinal regression
    #model
    #likelihood
    for(j in 1:n){
        Y[j]~dcat(pi[j,])

       # probabilities for pi
       pi[j,1]=pnorm(0-z[j],0,1)
       pi[j,K]=1-pnorm(gamma[K-1]-z[j],0,1)
       for(k in 2:(K-1)){
         pi[j,k]=max(0,pnorm(gamma[k]-z[j],0,1)
          -pnorm(gamma[k-1]-z[j],0,1))
    }

    # z is standardized Y* in latent var. model
    # (remember-sigma2_{Y*}=1)
    # used to find pnorm(gamma-z)
    z[j]=beta_0i[subj[j]]+beta_1*X[j]
}

#priors
for(i in subj_id){
    beta_0i[i]~dnorm(beta_0, tau)
}
for(k in 2:(K-2)){
    gamma[k]~dnorm(0, 0.1)
}
gamma[1] = 0
gamma[K-1]~dnorm(0, 0.1)
beta_0~dnorm(0, 0.01)
beta_1~dnorm(0, 0.01)
sigma2 = 1/tau #sigma2 here is for distrib.
    #of beta_0i's
tau~dgamma(0.01, 0.01)
}
```

APPENDIX B: COMPARISON OF NON-MULTILEVEL AND MULTILEVEL LOGISTIC REGRESSION MODELS

The multilevel logistic regression model is compared to a non-multilevel logistic regression model fit to the data because the latter model was recommended by FICON (Federal Interagency Committee on Noise, 1992) for previous cross-sectional studies. We were curious to what degree, if any, estimates from the multilevel model would differ from the estimates from the non-multilevel model, even though the non-multilevel model is not statistically appropriate for the panel data from QSF18. For the QSF18 data, the two models are neither statistically nor practically different when estimating the PL fixing percent highly annoyed, and estimating the percent highly annoyed fixing the PL. Nevertheless, the multilevel logistic regression model is recommended for analyzing this data because it fits the data better based on the goodness-of-fit metric, the deviance information criterion (DIC).

First, the non-multilevel logistic regression model is shown in Eq. (B1). Let H be the binary response, and p be the probability of high annoyance. The non-multilevel logistic regression model assumes these observations are independent. This model is analogous to modeling independent random coin flips with the probability of landing heads dependent on PL. The first two lines describe the standard logistic regression model, and the last two lines are the non-informative prior distributions assigned to the model parameters β_0 and β_1.

$$\begin{aligned} &H_i|p_i \sim Bernoulli(p_i)\\ &p_i|\beta_0,\beta_1 = logit^{-1}(\beta_0+\beta_1 PL_i) = \frac{\exp(\beta_0+\beta_1 PL_i)}{1+\exp(\beta_0+\beta_1 PL_i)}\\ &\beta_0 \sim N(0,100)\\ &\beta_1 \sim N(0,100). \end{aligned} \tag{B1}$$

The non-multilevel logistic regression model parameters are estimated using 50 000 posterior samples after discarding 4000 samples burn-in samples. Table III shows the posterior summaries for the two model parameters. To assess the model fit, three discrepancy statistics are calculated, none of which indicate lack of fit. The three statistics

TABLE III. Summary statistics of the non-multilevel logistic regression model parameters.

	Mean	SD	0.025 quant.	0.25 quant.	Median	0.75 quant.	0.975 quant.
β_0	−16.24	1.95	−20.18	−17.52	−16.20	−14.92	−12.52
β_1	0.153	0.025	0.105	0.136	0.152	0.169	0.202

are: the deviance, the proportion of highly annoyed responses, and the 0.1 quantile of the PL distribution where highly annoyed responses occur.

The summary dose-response curve estimate for the non-multilevel logistic regression consists of the pointwise posterior means of p_i calculated over the observed range of noise doses. A summary dose-response curve is calculated at each posterior sample to estimate a distribution of summary dose-response curves. To calculate each curve, the observed range of 56 to 90 dB is evenly divided into 1000 values. At each of the 1000 PL values, p_i is calculated as defined in Eq. (B1) and a curve connects the 1000 p_i values. The mean estimate of the summary dose-response curve is then the pointwise means of the 50 000 curves. The 95% credible intervals are the pointwise 0.025 and 0.975 quantiles of the 50 000 p_i values. Figure 9 compares the summary dose-response curves from the non-multilevel and multilevel logistic regression models. Note that the summary dose-response curve calculated from the non-multilevel logistic regression model is a population representative curve because the model parameters are estimated at a population level. The estimated summary dose-response curve for the non-multilevel logistic regression model is higher than that for the multilevel model above 80 dB PL. The 95% credible interval for each curve contains the point estimate for the other.

In addition to the visual comparison, the model comparison metric deviance information criterion (DIC) (Spiegelhalter *et al.*, 2002) is used to compare the two models. DIC can be broken down into two components: a goodness-of-fit measure and a penalty for the model complexity. Note that DIC does not have an absolute scale; rather, models are ranked relative to one another with the lowest DIC value indicating the best relative fit to the data. The DIC is 494.8 for the non-multilevel logistic regression model, and 385.4 for the multilevel logistic regression model. Thus, despite the model complexity, the multilevel version of the logistic regression model fits the data better.

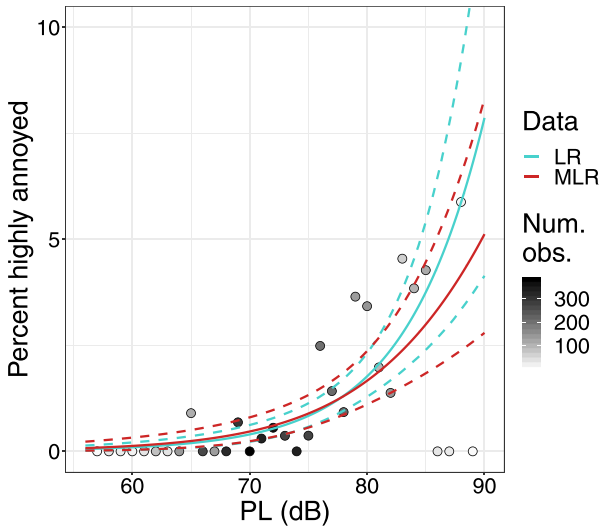

FIG. 9. (Color online) Estimated summary dose-response curves (solid) from the non-multilevel (LR) and multilevel (MLR) logistic regression models compared to observed percent highly annoyed data, with 95% credible intervals (dashed). Shading of points indicates the sample size at each dose. Two points are not shown in this plot: 33% at 56 dB (1 of 3 observations is highly annoyed), and 16% at 90 dB (1 of 6 observations is highly annoyed).

The quantities described in Sec. III B 2 are calculated for the non-multilevel logistic regression model as well to compare the practical differences between the non-multilevel and multilevel logistic regression models. Figure 10 compares the posterior distributions for the first three

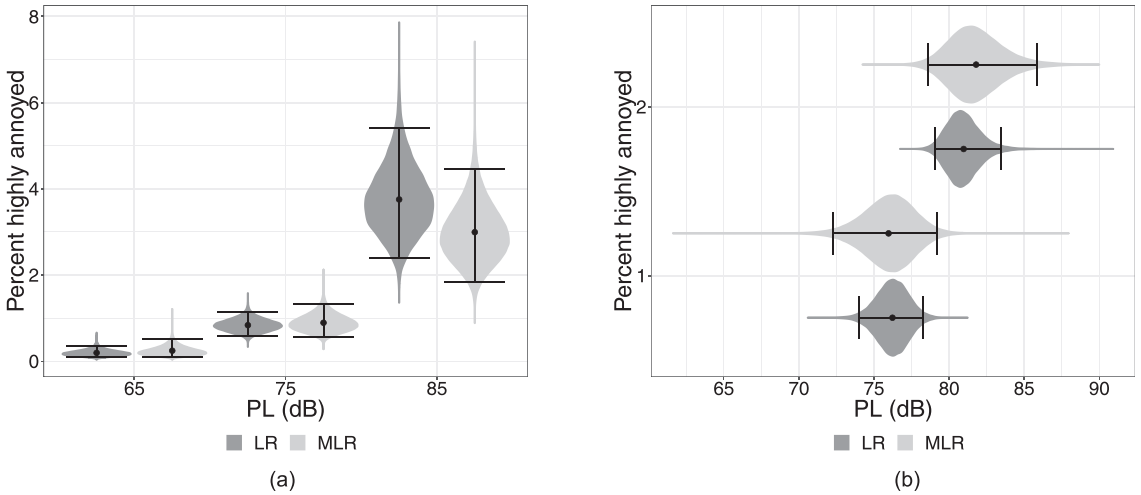

FIG. 10. Comparison of the posterior distributions for (a) percent highly annoyed fixing PL at 65, 75, and 85 dB, and (b) PL fixing percent highly annoyed at 1% and 2% calculated from the logistic regression (LR) and the multilevel logistic regression (MLR) models. The points indicate the point estimates, which are the posterior means, and the bars indicate the 95% credible intervals.

quantities in (a) and for the last two quantities in (b). When estimating the percent highly annoyed at the three fixed PL values, the estimates from the two models are quite similar: both model estimates are within the credible interval for the other. Practically, 3.75% highly annoyed at 85 dB estimated from the non-multilevel logistic regression model is not different from the 3% highly annoyed estimated from the multilevel logistic regression model. The PL estimates at fixed percent highly annoyed from the two models are also not practically different for this data: at 1% highly annoyed, the estimates differ by less than 1 dB.

[1]See Lee *et al.* (2019) for examples of these applications using statistical models fit to similar community response survey data.

[2]Another appropriate method commonly used to model panel sample data is a marginal model, which models the population-level instead of individual-level parameters. See Hu *et al.* (1998) and Fitzmaurice *et al.* (2012) for details on marginal models.

[3]This paper focuses on demonstrating the methods and modeling approaches. Analyzing the data to investigate whether the reminders introduced response bias is of interest but beyond the scope of this paper.

[4]The equal-tailed method, which is used in this paper, calculates the 0.025 and 0.975 quantiles of the posterior distribution for the lower and upper bounds of the 95% credible interval.

[5]See Gelman and Hill (2007) for details and examples about multilevel/hierarchical models, and Kruschke (2014) or Gelman *et al.* (2013) for details specific to using a Bayesian approach.

[6]We considered a random intercept and random slope model for a previous pilot study dataset. Based on the DIC criterion (Spiegelhalter *et al.*, 2002), we selected the simpler model, and we used this to inform our model for the QSF18 data. See Appendix B for a brief explanation on the DIC criterion.

[7]The prior distributions for β_0 and β_1 are both $N(0, 100)$. Relative to the likelihood, this choice of prior distributions is noninformative and has minimal impact on the model parameter estimation.

[8]The interval regression model proposed in Groothuis-Oudshoorn and Miedema (2006) assumes that scaled annoyance categories are equally spaced between 0 to 100. The ordinal regression model relaxes this assumption because the thresholds of the annoyance categories, γ, are estimated from the data instead of fixed. The estimated thresholds provide an opportunity to check the assumption of equally spaced annoyance categories. For the five-point annoyance scale in this dataset, the interval widths calculated from the median γ estimates in Table II tend to decrease with interval ($\gamma 2 - \gamma 1 = 0.97$, $\gamma 3 - \gamma 2 = 0.78$, and $\gamma 4 - \gamma 3 = 0.61$). This finding suggests that the annoyance categories may not be equally spaced on the annoyance scale. The authors recommend estimating interval thresholds from the data via the ordinal regression model in favor of assuming they are equally spaced via the interval regression model.

[9]See Lee *et al.* (2019) for additional suggested posterior predictive checks.

[10]See supplementary material at https://doi.org/10.1121/10.0001021 for a comma separated file of the data analyzed and the description of the data.

[11]The data cleaning process eliminated survey responses based on the following criteria: (1) inability to estimate a noise dose because participant location was not available, participant was outside the study area, or insufficient signal to noise ratio, (2) incomplete demographic information for the participant from the background survey, (3) the response referenced an event more than 15 min after the start time of the report because the selections for reporting boom times were in 15 min increments, (4) responses in which the participant did not respond whether he or she heard the event, and (5) responses in which the participant did not provide an annoyance rating. There were instances where a participant responded multiple times to the same boom event; for responses that were the same (every field in the survey was the same), only one survey response was retained. For responses that were different (not every field in the survey was the same), only the survey response with the latest completion time was kept. This follows the method documented in Page *et al.* (2014). Since the purpose of this paper is to demonstrate the two statistical models rather than substantive analysis, inclusion of these responses does not affect the objective. Last, there are five responses where the participant was given the opportunity to provide an annoyance rating even though he/she reported not hearing the event due to an error in the survey system. These responses are given a "not at all annoyed" rating to be consistent with other "not heard" responses, which were also assigned the "not at all annoyed" rating.

[12]See Kruschke (2014) for examples of traceplots and Gelman-Rubin plots.

[13]The Monte Carlo standard error (MCSE) accounts for simulation accuracy and is used to determine the number of decimal places to report for the posterior summaries by approximating a 95% interval around the posterior mean using $\pm$(2*MCSE). This interval is used to quantify the sampling variability in the posterior summary estimates. Since the posterior summaries are approximated using a random sample from the posterior distribution, we expect slightly different values if we drew a different random sample. The time-series standard error is one method for estimating the MCSE for Markov chains (Kruschke, 2014), but Flegal *et al.* (2008) describe other methods. The time-series SE is $s/\sqrt{ESS}$, where s is the sample standard deviation and ESS is the effective sample size, or the number of independent samples that the correlated MCMC samples is equivalent to. It is calculated as $ESS = L/[1 + 2\sum_{h=1}^{\infty}\rho(h)]$, where L is the number of posterior samples and $\rho(h)$ is autocorrelation at lag h. For example, the posterior summaries for β_0 are reported to the second decimal place in Table III because (2*time-series SE) is 0.004. Note that the MCSE is not equivalent to the posterior standard deviation, which quantifies the spread of the posterior distribution.

[14]Note that these PL and percent highly annoyed values are chosen because they are all within the observed range. Lee *et al.* (2019) advise against extrapolating beyond the observed ranges.

[15]One reason why the assumption of data missing at random may not hold is that for the participants who did not receive reminders, we expect fewer responses if the signal was not audible at lower noise exposure levels. In addition, missingness may depend on ordering if, for example, participants dropped out of the survey partway through.

Federal Interagency Committee on Noise (**1992**). "Federal agency review of selected airport noise analysis issues," technical report.

Fidell, S., and Horonjeff, R. D. (**2019**). "Field evaluations of sampling, interview, and flight tracking of NASA's Low Boom Flight Demonstrator aircraft," Technical Report No. NASA/CR-2019-22057, ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20190001426.pdf (Last viewed 10/03/2019).

Fidell, S., Mestre, V., Schomer, P., Berry, B., Gjestland, T., Vallet, M., and Reid, T. (**2011**). "A first-principles model for estimating the prevalence of annoyance with aircraft noise exposure," J. Acoust. Soc. Am. **130**(2), 791–806.

Fidell, S., Schultz, T., and Green, D. M. (**1988**). "A theoretical interpretation of the prevalence rate of noise-induced annoyance in residential populations," J. Acoust. Soc. Am. **84**(6), 2109–2113.

Fields, J., De Jong, R., Gjestland, T., Flindell, I., Job, R. S., Kurra, S., Lercher, P., Vallet, M., Yano, T., Guski, R., Felscher-Suhr, U., and Schumer, R. (**2001**). "Standardized general-purpose noise reaction questions for community noise surveys: Research and a recommendation," J. Sound Vib. **242**(4), 641–679.

Fitzmaurice, G., Laird, N., and Ware, J. (**2012**). *Wiley Series in Probability and Statistics, Applied Longitudinal Analysis*, 2nd ed. (Wiley, Hoboken, NJ), pp. 1–752.

Flegal, J. M., Haran, M., and Jones, G. L. (**2008**). "Markov chain Monte Carlo: Can we trust the third significant figure?," Stat. Sci. **23**(2), 250–260.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (**2013**). *Texts in Statistical Science, Bayesian Data Analysis*, 3rd ed. (Chapman and Hall/CRC, Boca Raton, FL), pp. 1–639.

Gelman, A., Goegebeur, Y., Tuerlinckx, F., and Van Mechelen, I. (**2000**). "Diagnostic checks for discrete data regression models using posterior predictive simulations," J. R. Stat. Soc.: Ser. C (Appl. Stat.) **49**(2), 247–268.

Gelman, A., and Hill, J. (**2007**). *Analytical Methods for Social Research, Data Analysis Using Regression and Multilevel/Hierarchical Models* (Cambridge University Press, Cambridge), pp. 1–625.

Gelman, A., and Rubin, D. (**1992**). "Inference from iterative simulation using multiple sequences," Stat. Sci. **7**(4), 457–511.

Gille, L.-A., and Marquis-Favre, C. (**2019**). "Estimation of field psychoacoustic indices and predictive annoyance models for road traffic noise combined with aircraft noise," J. Acoust. Soc. Am. **145**(4), 2294–2304.
Gille, L.-A., Marquis-Favre, C., and Weber, R. (**2016**). "Noise sensitivity and loudness derivative index for urban road traffic noise annoyance computation," J. Acoust. Soc. Am. **140**(6), 4307–4317.
Groothuis-Oudshoorn, C. G. M., and Miedema, H. M. E. (**2006**). "Multilevel grouped regression for analyzing self-reported health in relation to environmental factors: The model and its application," Biometr. J. **48**(1), 67–82.
Haering, E. A., Jr., Smolka, J. W., Murray, J. E., and Plotkin, K. J. (**2006**). "Flight demonstration of low overpressure n-wave sonic booms and evanescent waves," AIP Conf. Proc. **838**, 647–650.
Hu, F. B., Goldberg, J., Hedeker, D., Flay, B. R., and Pentz, M. A. (**1998**). "Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes," Am. J. Epidemiol. **147**(7), 694–703.
Kruschke, J. K. (**2014**). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*, 2nd ed. (Elsevier Science, London), pp. 1–759.
Lee, J., Rathsam, J., and Wilson, A. (**2019**). "Statistical modeling of quiet sonic boom community response survey data," Technical Report No. NASA/TM-2019-220427, ntrs.nasa.gov/search.jsp?R=20190033466 (Last viewed 10/03/2019).
Long, J. (**1997**). *Advanced Quantitative Techniques in the Social Sciences Series, Regression Models for Categorical and Limited Dependent Variables* (SAGE Publications, Thousand Oaks, CA), pp. 114–147.
Maglieri, D. J., Bobbitt, P. J., Plotkin, K. J., Shepherd, K. P., Coen, P. G., and Richwine, D. M. (**2014**). "Sonic boom: Six decades of research," Technical Report No. NASA/SP-2014-622, ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20150006843.pdf (Last viewed 10/03/2019).
Miedema, H. M. E., and Vos, H. (**1998**). "Exposure-response relationships for transportation noise," J. Acoust. Soc. Am. **104**(6), 3432–3445.
Miller, N. P., Cantor, D., Lohr, S., Jodts, E., Boene, P., Williams, D., Fields, J., Gettys, M., Basner, M., and Hume, K. (**2014**). *Research Methods for Understanding Aircraft Noise Annoyances and Sleep Disturbance* (The National Academies Press, Washington, DC), pp. 1–172.
Page, J., Plotkin, K., and Wilmer, C. (**2010**). *PCBoom Version 6.6 Technical Reference and User Manual*, Wyle Laboratories, Arlington, VA.
Page, J. A., Hodgdon, K. K., Krecker, P., Cowart, R., Hobbs, C., Wilmer, C., Koening, C., Holmes, T., Gaugler, T., and Shumway, D. L. (**2014**). "Waveforms and sonic boom perception and response (WSPR): Low-boom community response program pilot test design, execution, and analysis," Technical Report No. NASA/CR-2014-218180, ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20140002785.pdf (Last viewed 10/03/2019).
Plummer, M. (**2003**). "JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling," in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, Vienna, Austria, Vol. 124, pp. 1-10.
Rathsam, J., Hayward, M., Gille, L.-A., Nykaza, E., and Wayant, N. (**2018a**). "Multilevel modeling of recent community noise annoyance surveys," Proc. Meet. Acoust. **33**, 040002.
Rathsam, J., Klos, J., Loubeau, A., Carr, D. J., and Davies, P. (**2018b**). "Effects of chair vibration on indoor annoyance ratings of sonic booms," J. Acoust. Soc. Am. **143**(1), 489–499.
Schäffer, B., Pieren, R., Mendolia, F., Basner, M., and Brink, M. (**2017**). "Noise exposure-response relationships established from repeated binary observations: Modeling approaches and applications," J. Acoust. Soc. Am. **141**(5), 3175–3185.
Schultz, T. J. (**1978**). "Synthesis of social surveys on noise annoyance," J. Acoust. Soc. Am. **64**(2), 377–405.
Shepherd, K. P., and Sullivan, B. M. (**1991**). "A loudness calculation procedure applied to shaped sonic booms," Technical Report No. NASA-TP-3134, ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19920002547.pdf (Last viewed 10/03/2019).
Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (**2002**). "Bayesian measures of model complexity and fit," J. R. Stat. Soc.: Ser. B **64**(4), 583–639.
Stevens, S. S. (**1972**). "Perceived level of noise by Mark VII and decibels (E)," J. Acoust. Soc. Am. **51**(2B), 575–601.
Stevens, S. S. (**1975**). *Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects* (Wiley, New York), pp. 1–329.
Trollé, A., Marquis-Favre, C., and Klein, A. (**2014**). "Short-term annoyance due to tramway noise: Determination of an acoustical indicator of annoyance via multilevel regression analysis," Acta Acust. Acust. **100**(1), 34–45.
U.S. Environmental Protection Agency (**1982**). "Guidelines for noise impact analysis," Technical Report No. EPA-550/9-82-105.
Wilson, D. K., Wayant, N. M., Nykaza, E. T., Pettit, C. L., and Armstrong, C. M. (**2017**). "Multilevel modeling and regression as applied to community noise annoyance surveys," J. Acoust. Soc. Am. **141**(5), 3727–3728.
Xiang, N., and Fackler, C. (**2015**). "Objective Bayesian analysis in acoustics," Acoust. Today **11**(2), 54–61.