Bayesian analysis of paired-comparison sound quality ratings

Arne Leijon,[1,a)] Martin Dahlquist,[2] and Karolina Smeds[2]
[1]*School of Electrical Engineering, KTH, Stockholm, Sweden*
[2]*ORCA Europe, Widex A/S, Björns Trädgårdsgränd 1, SE-11621 Stockholm, Sweden*

(Received 22 April 2019; revised 5 October 2019; accepted 7 October 2019; published online 11 November 2019)

This paper presents a method to analyze paired-comparison data including either binary or graded ordinal responses, with or without ties. The proposed method can use either of two classical choice models: (1) Thurstone case V, which assumes a Gaussian distribution of the sensory variables underlying listener decisions, or (2) the Bradley-Terry-Luce (BTL) model, which assumes a logistic distribution. The analysis method was validated using simulated paired-comparison experiments with known distributions of the sound-quality parameters in the simulated population from which "participants" were generated at random. The validation indicated that the Thurstone and BTL models give similar results close to the true values. The estimated credibility of a quality difference was slightly higher with the BTL model. The analysis results showed dramatically better precision when the response data included graded ordinal judgments instead of binary responses. Allowing tied responses also tended to improve precision. The method was also applied to data from a real evaluation of hearing-aid programs. The analysis revealed clinically interesting results with high statistical credibility, although the amount of test data was limited.

© *2019 Acoustical Society of America*. https://doi.org/10.1121/1.5131024

[VMR] Pages: 3174–3183

I. INTRODUCTION

Subjective quality evaluations are necessary in all applications of audio processing and audio coding systems that introduce some distortion. The present study is focused on sound quality evaluations of hearing aids or similar equipment, but subjective rating methods are also widely used in evaluations of speech coders for telecommunication (e.g., Grancharov and Kleijn, 2008) and other multi-media applications.

Standard procedures have been defined for quality evaluations in telecommunications (ITU, 2003). Two psychophysical methods are commonly used (IEEE, 1969): (1) absolute magnitude ratings, and (2) paired-comparison ratings. In these applications, the purpose is usually to measure quality degradations caused by coding and transmission, and the processed sound can be compared to the perfect original version.

In evaluations of subjective quality of, e.g., different signal processing strategies in a hearing aid, it is impossible to define a perfect "original" version. Here, the only possible way is to compare different systems, none of which has perfect quality for the user. Such comparisons are still sufficient to determine if a new hearing aid algorithm gives better subjective sound quality than some other existing state-of-the-art variant. In these applications, the paired-comparison procedure is useful.

In paired-comparison procedures, each presentation includes two test items, and the test participant indicates which of the items is better in terms of the specified *perceptual attribute* being investigated. The perceptual attribute could be, e.g., "speech clarity," "sound quality," or "general preference," or any other quality defined by the test instructions. This method can be applied even if the difference between tested systems is so small that it is just barely detectable.

Listeners may be required to make *graded* judgments of the perceived difference between the items in each presented pair, as proposed by Dillon (1984), or to give only a *binary* response, as assumed in the classical theoretical papers (Bradley and Terry, 1952; Thurstone, 1927). It has not yet been scientifically established which of these methods gives the most accurate results. Most computational analysis methods only allow binary data (Cattelan, 2012). It is also an open research question whether *forced-choice* judgments should be required. Allowing *tied* judgments, i.e., including a "No difference" response alternative, probably makes the procedure subjectively more pleasant for participants. However, a recent review (Pérez-Ortiz and Mantiuk, 2017) noted that the use of tied responses "is a controversial issue…, still disputed and researched."

The application of paired comparisons for hearing-aid evaluation was reviewed by Amlani and Schafer (2009). They presented the historical background, theoretical principles, and clinical usefulness of paired comparisons. Paired comparisons in the field using cellular-phone apps for data recording is an evolving trend for hearing-aid evaluations, by so-called Ecological Momentary Assessment (EMA) methods (e.g., Smeds *et al.*, 2019). Such experiments generate large amounts of data, including subsets of data from various listening conditions. There may be widely different amounts of data from different participants and situations.

A simple way to quantify the results of a paired-comparison evaluation is to record the *"win counts"*, i.e., how many times each tested object was ranked higher than a competing item. This method has two main problems

[a)] Current address: Strindbergsgatan 36, SE-11531 Stockholm, Sweden. Electronic mail: leijon@kth.se

(Pérez-Ortiz and Mantiuk, 2017): (1) In order to get fair and balanced win counts, every test object must be compared to every other object equally many times, and all test participants should make the same number of comparisons. (2) Win counts provide only an ordinal ranking of the tested objects but do not reflect the magnitude of the perceived difference between objects.

For these reasons, most paired-comparison studies have used probabilistic scaling models to analyze the data. The resulting quality measures are estimated on a well-defined objective interval scale, although all the primary paired-comparison data are necessarily subjective, with only ordinal properties. There is a vast amount of literature on such analysis models, all based on one of two classical variants: Thurstone (1927) presented the first model and emphasized its relation to psycho-physical discrimination tests. The other model was first proposed by Zermelo (1929) for the analysis of chess tournaments and was later independently re-invented by Bradley and Terry (1952), who exemplified its use in a taste-testing experiment with pork roasts from pigs fed on different diets. This later model variant is often referred to as the "Bradley-Terry-Luce (BTL)" model because Luce (1959) presented an axiomatic motivation for the model structure. The mathematical theory behind both models, estimation methods for model parameters, and some later model extensions will be briefly reviewed in Sec. II.

We now present a novel approach that extends previous methods in several ways. To our knowledge, this is the first fully Bayesian analysis method that allows binary or graded responses from paired comparisons, with or without ties, using either the Thurstone or the BTL framework, and also allows for the possibility that each participant might interpret the ordinal response categories in different ways. A separate model is adapted to the data for *each participant*, and these individual models are also hierarchically influenced by a model adapted for the *population* from which the participants were recruited. The individual models can include separate quality parameters for the tested objects in different *test conditions*. The proposed method has no requirements on the number of presentations. The Bayesian analysis result automatically quantifies the statistical reliability for the given amount of raw test data.

In addition to defining the analysis model, the present study will answer the following main research questions:

(1) Is there an advantage of using either the Thurstone or the BTL model for the analysis?
(2) Can more precise results be obtained from tests allowing graded differences rather than binary responses?
(3) Will the estimated results get more or less precise if judgment "ties" are allowed, i.e., responses like "No preference"?
(4) Can a limited amount of individual paired-comparison data be used to predict perceptual differences for the population from which the test participants were recruited?

The current study is focused on paired comparison for evaluation of sound-processing features. Procedures for individual fine-tuning of those features may also use paired comparisons, but that application is outside the scope of this study.

II. THEORY — MODELS AND ESTIMATION METHODS

This section briefly reviews the Thurstone and the BTL models for paired-comparison data. We extend both models to include graded difference magnitudes, and allow each test participant to use different criteria for their grading. We then propose a hierarchical Bayesian estimation procedure to estimate individual and population parameters for either of the two model variants.

In both the Thurstone and the BTL models, the listener's response to a presented pair (A, B) is determined by the outcome of a sensory random variable X_{AB} with a probability distribution that depends only on the difference between some unknown quality parameters μ_A and μ_B for the two objects. The probability for the event that the participant ranks object B higher than A in a binary forced-choice trial is defined by the cumulative distribution function $F(\,)$ of the decision variable, as

$$P[\text{“B} > \text{A”}|\mu_{\text{A}}, \mu_{\text{B}}] = P[X_{AB} > 0|\mu_A, \mu_B] = F(\mu_B - \mu_A). \quad (1)$$

Thus, the quality difference $\mu_B - \mu_A$ is objectively defined on an interval scale by the function $F(\,)$ and the corresponding response probability. Although the response probabilities cannot be directly observed, it is possible to estimate values for the model parameters μ_A and μ_B, and similar parameters for other tested objects, in agreement with the complete set of paired-comparison responses. It should be noted that this formal definition of the quality scale does not require that the participants give only binary responses. The quality parameters can be estimated in a similar way based on graded ordinal responses, as exemplified in Fig. 1 and formally defined in Sec. II C 1. A model with several decision thresholds was used already by Garner (1952) and Durlach and Braida (1969) for intensity discrimination experiments and later proposed by Agresti (1992) for paired comparisons. Figure 1 also illustrates how the use of threshold parameters conveniently allows for "Equal" responses, as proposed by Rao and Kupper (1967) using a different parameterization restricted to the BTL model.

Since paired-comparison tests can only reveal *differences* between objects, the zero point of the quality scale is arbitrary. One of the tested objects may always be placed at the zero point of the scale.

A. The Thurstone model

The Thurstone model is closely related to psychophysical measurements of Just Noticeable Differences (JND) in the sense that the quality measures for the tested objects are placed on a *"Cumulative-JND"* scale. The scale is sometimes called *"Cumulative-d-prime"* because the detectability index d' is used to define the JND. The detectability index d' is derived from general signal detection theory as applied to psychophysical experiments (Green and Swets, 1988). The

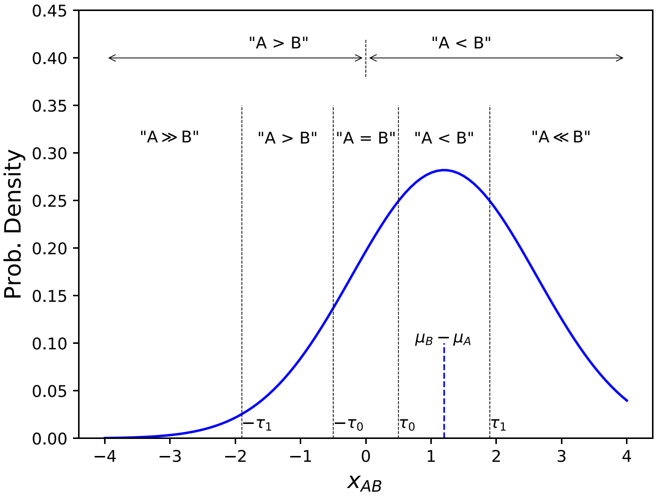


FIG. 1. (Color online) Example of a conditional probability density function of the sensory variable X_{AB} that determines the response to a single paired-comparison presentation (A, B), given a true quality difference $\mu_B - \mu_A = 1.2$ *d*-prime units. Decision intervals are indicated for one experiment with only forced-choice binary responses "A > B" or "A < B," and for another experiment allowing five graded response alternatives, "$A = B$", "$A < B$", "$A \ll B$," etc., with response thresholds defined by parameters τ_0 and τ_1.

Cumulative-JND scale also has a long history of use as an objective scale of loudness based on intensity discrimination (Allen and Neely, 1997; Durlach and Braida, 1969; Houtsma *et al.*, 1980; Garner, 1952; Riesz, 1933).

For the Thurstone (case V) model, we assume that the presented objects A and B yield two independent Gaussian sensory variables Z_A and Z_B with means μ_A and μ_B, both with unity variance. (The standard deviation is just an arbitrary scale factor, so assuming unity is no loss of generality.) The difference variable $X_{AB} = Z_B - Z_A$ is then also Gaussian with mean $\mu_{AB} = \mu_B - \mu_A$ and variance $\sigma^2_{AB} = 2$. Thus, the function in Eq. (1) is

$$F(x) = \Phi\left(x/\sqrt{2}\right), \tag{2}$$

where $\Phi()$ is the standard Gaussian cumulative distribution function.

Although Thurstone (1927) explicitly warned against interpreting the model parameters as real physical quantities, it is tempting to consider each decision as the result of some neural activity in the listener's brain. Since the decision is influenced by a very large number of independent random neural events, it is plausible to assume that the underlying sensory variable has a Gaussian distribution, according to the Central Limit Theorem in probability theory.

B. The BTL model

The BTL model (Luce, 1959; Bradley and Terry, 1952) differs from the Thurstone model only by using a *logistic* distribution for the decision variable, with the cumulative distribution function in Eq. (1) defined as

$$F(x) = \frac{1}{1 + e^{-x}}. \tag{3}$$

It may seem strange to assume that some neural activity in the listener's brain should follow a logistic distribution. This model assumption cannot be motivated by reference to the Central Limit Theorem. Instead, Luce (1959) showed that the BTL model structure follows logically from a plausible "choice axiom." He also noted the similarity between the Thurstone and BTL models and argued that the Gaussian and logistic distributions are so similar that the two models might be nearly equivalent for practical purposes. The similarity was also explored by Tsukida and Gupta (2011). Both models can also be equivalently formulated in the framework of cumulative link models for ordinal regression (Bürkner and Vuorre, 2019), using either a "probit" (Thurstone) or a "logit" (BTL) link function.

The standard Thurstone and BTL models both enforce a *uni-dimensional* scale for all judgments. This might be unrealistic in some applications where participants may focus on different quality aspects depending on which objects are presented (e.g., Zimmer *et al.*, 2004). If the judgments are actually determined by multi-dimensional sensory impressions, the responses may show systematic *intransitivity*, e.g., A > B, B > C, and C > A. The BTL model has been extended to allow several perceptual dimensions for each object (Tversky, 1972). This extension would be more difficult to apply to the Thurstone model structure, but, on the other hand, the BTL extension cannot handle graded responses. Wickelmaier and Schmid (2004) developed a MATLAB program to estimate model parameters using the multi-dimensional BTL model. This program was recently applied to evaluations of hearing instruments (Laugesen *et al.*, 2015), although it is not clear if the multi-dimensional feature of the model was actually needed in that application.

C. Bayesian parameter estimation

Conventional maximum-likelihood (ML) estimation of paired-comparison model parameters might have no well-defined solution for some input data. For example, what if a subject consistently responds "$B \gg A$" for every presentation of (A, B)? Then, the ML estimate of parameters as in Fig. 1 would be $\mu_B - \mu_A \to \infty$, and $\tau_1 \to \tau_0 \to 0$. Therefore, it is necessary to apply a weakly informative prior distribution for the parameters. A feasible point estimate of the model parameters is then the *maximum a posteriori probability* (MAP) solution, as proposed by Dahlquist and Leijon (2003).

In contrast, the goal of a fully Bayesian approach is to estimate the complete posterior distribution of all model parameters, given the observed data set, i.e., not only a single typical point of that distribution. Thus, the Bayesian result automatically includes a reliability measure. We now propose a new hierarchical Bayesian model with estimation of all model parameters jointly for each individual participant and for the population from which test participants were recruited.

1. Individual response probabilities

In a paired-comparison experiment, we have a set $\{A, B, C, \ldots\}$ of two or more sound "objects," i.e., classes of sound stimuli, from which the two sounds (S_{p1}, S_{p2}) in the

pth presented pair are selected. The graded ordinal response to the p-th presentation can always be encoded by an integer $R_p \in \{0, \pm 1, \pm 2, \ldots, \pm M\}$, where M represents the largest difference grade. When, e.g., $(S_{p1}, S_{p2}) = (A, B)$ as in Fig. 1, $R_p = 0$ would mean "$A = B$", and $R_p = +2$ could mean "$A \ll B$".

The log-probability of any response is calculated as

$$L(R_p) = \begin{cases} \ln P\left[\tau_{R_p - 1} < X_p \le \tau_{R_p}\right], & R_p > 0, \\ \ln P\left[-\tau_0 < X_p \le \tau_0\right], & R_p = 0, \\ \ln P\left[-\tau_{-R_p} < X_p \le -\tau_{-R_p - 1}\right], & R_p < 0, \end{cases} \tag{4}$$

where $\boldsymbol{\tau} = (\tau_0, \tau_1, \ldots, \tau_M = \infty)$ is a strictly increasing sequence of thresholds for the decision variable X_p, as exemplified in Fig. 1. In a *forced-choice* experiment, the response $R_p = 0$ is not allowed, and $\tau_0 \equiv 0$. Otherwise, except for the outermost limit $\tau_M = \infty$, the thresholds are considered as free model parameters, assumed to have the same values across all presentations for each participant, but the values may differ among individuals. Given the individual quality-parameter difference $\delta_p = \mu_{S_{p2}} - \mu_{S_{p1}}$ and the individual thresholds, the response log-likelihood for this listener is calculated as

$$L(R_p | \boldsymbol{\mu}, \boldsymbol{\tau}) = \ln\left(F(\tau_{R_p} - \delta_p) - F(\tau_{R_p - 1} - \delta_p)\right), \tag{5}$$

in case $R_p > 0$ and similarly for the other response cases. Assuming that responses are conditionally independent, given the model, the total log-likelihood for all responses from the listener is just the sum across presentations.

It should be noted that this model imposes no requirements on the number of presentations for each pair, or for any listener. If more than two systems are being compared in an experiment, it is not even necessary that all possible combinations are presented. For example, if three systems, A, B, C, are evaluated, it is computationally sufficient to have results for pairs (A, B) and (B, C), and no direct comparisons of (A, C). Of course, the reliability of the analysis results will improve if all combinations are included, and if there are many presentations of each combination. It is also advisable to balance the presentations, with blinding, such that (A, B) and (B, A) are presented equally often and in randomized order, unknown to the listener.

2. Individual response thresholds

To ensure that the thresholds τ_m form a strictly increasing sequence, and for numerical stability, it is convenient to map the response intervals to the range [0,1] using the logistic distribution function $F(\,)$ from Eq. (3), as

$$2F(\tau_m) - 1 = \frac{\sum_{i=0}^{m} e^{\eta_i}}{\sum_{i=0}^{M} e^{\eta_i}}. \tag{6}$$

Here, the parameters $\boldsymbol{\eta} = (\eta_0, \ldots, \eta_m, \ldots, \eta_M)$ are logarithms of the relative interval widths in this mapped range. The inverse mapping defines the boundaries as

$$\tau_m(\boldsymbol{\eta}) = \ln \frac{\left(\sum_{i=0}^{M} e^{\eta_i}\right) + \sum_{i=0}^{m} e^{\eta_i}}{\left(\sum_{i=0}^{M} e^{\eta_i}\right) - \sum_{i=0}^{m} e^{\eta_i}}. \tag{7}$$

Adding a constant to all elements in $\boldsymbol{\eta}$ does not change the resulting thresholds, so the $M+1$ values actually define only M free interval boundaries. In case of *forced-choice* trials, $\tau_0 = 0$, and $\eta_0 = -\infty$. With *binary* forced-choice data the interval limits are fixed at $\tau_0 = 0$ and $\tau_1 = \infty$, so the thresholds have no influence and can be omitted from the model.

3. Individual vs population models

Let $\boldsymbol{U}_n = (\ldots, \mu_{nit}, \ldots, \eta_{nm}, \ldots)^T$ denote the column vector of all model parameters for the nth listener in a group of N test participants. Here, μ_{nit} is the quality parameter for the ith tested object in the tth test condition, for $i = 1, \ldots, I - 1$, excluding the fixed values $\mu_{n,0,t} \equiv 0$. The parameters η_{nm}, $m = 0, \ldots, M$ define the individual listener's response thresholds by Eq. (7). A corresponding parameter vector is defined for the *mean* in the population from which test participants are recruited.

The distribution of individual parameters is adapted to the response data from each listener together with a Gaussian prior density conditional on the population parameters, as defined in Eq. (A1) in Appendix A. The population model is simultaneously adapted to all the individual results together with a Gaussian-gamma prior density defined in Eq. (A2) for the population parameters.

This model structure is somewhat similar to the hierarchical model proposed by Böckenholt (2001) for the BTL model. Tsai and Böckenholt (2002) used a similar approach with the Thurstone model and allowed ordinal responses but still assumed only a single set of threshold parameters for all participants. Cattelan (2012) reviewed several models and software packages allowing individual variations in quality values, still only deriving point estimates for all parameters and assuming identical response thresholds for all participants.

4. Variational model inference

As described in Appendix B, we use variational learning (e.g., Bishop, 2006, Chap. 10) to derive approximate posterior density functions $q(\boldsymbol{U}_n), n = 0, \ldots, N-1$, for the parameters of each individual test participant, as well as a separate posterior Gauss-gamma density function $q(\boldsymbol{V}, \Lambda)$ for the population mean $\boldsymbol{V}$ and precision (inverse variance) Λ, given all observed data. Although the participant models $q(\boldsymbol{U}_n)$ are formally independent, the population model has a regularizing influence on all the individual models.

5. Predictive distributions

The trained models are used to calculate three predictive distributions as defined in Appendix C:

(1) for a random individual in the *group of participants* (used in Fig. 4),
(2) for a random individual in the *population* from which the participants were recruited (used in Figs. 2, 3, and 5),
(3) for the *population mean* (used in Figs. 2 and 5)

The predictive distributions are used to evaluate the *joint credibility* for combinations of single hypotheses, as described by Leijon *et al.* (2016, Appendix C).

III. EXPERIMENTAL METHODS

In order to quantify the precision of the analysis results, it is necessary to evaluate the difference between estimated model parameters and the corresponding true parameter values. The only way to do this is to use simulated paired-comparison experiments, because the true values are, of course, never known in real experiments. An analysis of data from a real study is also included to exemplify how the proposed method can reveal interesting results even with a rather small amount of data.

A. Simulated paired-comparison trials

Several types of computer-simulated paired-comparison experiments were performed as described in the following sections. All the simulations included two or three different "hearing aids," evaluated by groups of $N \in \{5, 10, 20, 30\}$ simulated "listeners," each performing $K \in \{2, 4, 8, 10\}$ replicated judgments for each pair of objects. The "listeners" were drawn from a population in which the means of the true quality parameters were fixed, and the quality parameters of each individual were set to deviate from the population mean by a random normal-distributed amount with zero mean and a standard deviation of 0.3 *d*-prime units unless otherwise stated. Most of the simulated experiments allowed $C = 7$ response alternatives with four difference grades: "no difference," "slightly better," "better," and "much better."

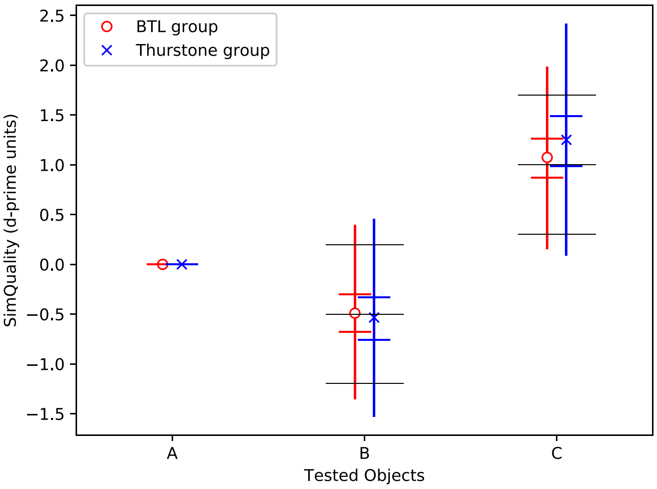


FIG. 2. (Color online) Example results from the proposed analysis method, estimated using the Thurstone analysis model [Eq. (2)]. Estimated medians (○,×) and 90% symmetric credible intervals are shown by vertical lines for a random individual in two simulated populations, one with responses generated by the Thurstone model, and one with the BTL model. Short horizontal lines show credible intervals for the population means. True population means and 90% ranges are shown by longer horizontal lines. Data were generated for $N = 20$ simulated "listeners" from each population, with $K = 10$ replicated judgments for each pair.

1. Illustrative example

Two groups of $N = 20$ "listeners" were drawn at random, one group from each of two separate populations. In one group, the simulated decisions were generated by the Thurstone model, and in the other group the BTL model was used to generate the responses.

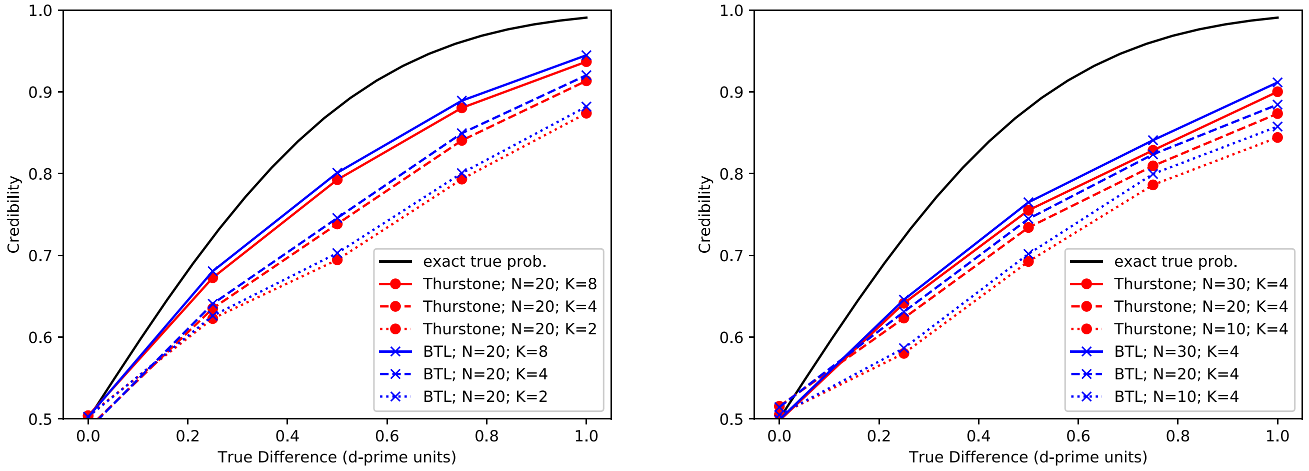
The simulation was designed to resemble a realistic scenario for a paired-comparison experiment. Subjects in the Thurstone population were assigned quality parameters $\boldsymbol{\mu}$ for three hearing aids with a population mean of (0, −0.5, 1.0). Because of the different scales of the Thurstone and the BTL models, as defined in Eqs. (2) and (3), the corresponding mean in the BTL population was set as (0, −0.57, 1.15) BTL scale units. The inter-individual standard deviation was 0.3 *d*-prime units in the Thurstone population and 0.34 scale units in the BTL population. These population parameters are equivalent in the sense that they would generate the same probability distribution of responses in a forced-choice binary trial for both populations. Responses were generated with $C = 7$ alternatives with four difference grades, specified by fixed thresholds $\boldsymbol{\tau} = (0.5, 1.5, 2.5)$ *d*-prime units in the Thurstone group. The corresponding thresholds for the BTL-generated responses were $\boldsymbol{\tau} \approx (0.57, 1.78, 3.22)$ BTL scale units. These two sets of thresholds are equivalent in the sense that they yield the same probability distribution of responses in both groups for a presented pair with zero quality difference.

2. Credibility vs experimental effort

In the planning of an evaluation experiment, one critical question might be the following: Assuming that our new device B is just slightly better than the reference, A, how many test participants do we need to engage, and how complex test procedure should we use, in order to show the improvement with good statistical reliability? To answer such questions, we simulated several group experiments comparing two objects A and B with true population means $\mu_A = 0$ and $\mu_B \in (0, \ldots, 1)$ *d*-prime units. The inter-individual standard deviation was 0.3 *d*-prime units in the population. All simulated responses were generated using the Thurstone model.

For each simulated data set, the posterior probability (credibility) was estimated for the hypothesis that $\mu_B > \mu_A$ for a random individual in the population, using both the Thurstone (Eq. 2) and the BTL (Eq. 3) analysis models. To show the effects of group size and test procedure, one series of experiments was done with groups of $N = 20$ "subjects," each performing $K \in \{2, 4, 8\}$ replicated judgments for each pair. Another series was done with groups of $N \in \{10, 20, 30\}$ "subjects," each performing $K = 4$ replicated judgments for each pair. This second series was also designed to reveal the potential errors caused by "lapse responses," in which the participants temporarily lose focus and just respond at random, regardless of the presented sounds. Lapse responses were generated with a probability of $p_l = 10\%$.

All experiments allowed $C = 7$ response alternatives with four difference grades, specified by fixed response



(A) No lapses. N = 20 subjects.

(B) Lapse probability 10%. K = 4 presentations.

FIG. 3. (Color online) Estimated credibility for the hypothesis that the quality difference between two objects is positive for a random individual in the population, plotted vs the true mean difference. The credibility was estimated using the Thurstone [Eq. (2)] (○) and the BTL [Eq. (3)] (×) analysis models for each simulated experiment with K pair presentations for each of N "subjects" in the test group. Each data point shows the average credibility across 100 group simulations. The solid curve without symbols is the proportion of individuals with positive true difference in the population.

thresholds $\boldsymbol{\tau} = (0.5, 1.5, 2.5)$ d-prime units. All experiments were replicated with 100 random group simulations for each test condition.

3. Number of response categories

An important experimental design issue, and one of our main research questions, is whether subjects should be required to make graded difference judgments or only give binary responses. Therefore, we simulated a series of experiments using forced-choice binary responses, i.e., $C = 2$ categories, forced-choice graded responses with three ordinal difference magnitudes, i.e., $C = 6$ response categories, as well as with three difference grades plus an "equal" grade, i.e., a total of $C = 7$ alternatives.

All experiments compared two objects A and B with true population means $\mu_A = 0$ and $\mu_B \in (0, \ldots, 1)$ d-prime units and an inter-individual standard deviation of 0.3 d-prime units. All simulated responses were generated using the Thurstone model and results were also estimated using this analysis model [Eq. (2)]. The response thresholds were $\boldsymbol{\tau} = (0, 1, 2)$ for the forced-choice trials with three grades $(C = 6)$ and $\boldsymbol{\tau} = (0.5, 1.5, 2.5)$ in the trials with $C = 7$ response categories. Each simulated group included $N = 20$ "subjects," each performing $K \in \{4, 8\}$ replicated judgments with zero lapse probability. The experiments were replicated with 100 random groups for each test condition.

For all groups, we calculated the squared difference between the estimated and the known true quality difference between the objects for each individual. The squared deviations were averaged across subjects and across the 100 group simulations, and the square root of the result is shown in Fig. 4 as a single root-mean-square (rms) error measure for each test condition.

B. Real evaluation of hearing aids

The subjective preference for two hearing-aid programs, called A and B, was evaluated with paired comparisons by $N = 10$ experienced hearing-aid users (Smeds *et al.*, 2019). The two programs were compared in the laboratory using listening situations guided by the Common Sound Scenarios (CoSS), selected to represent typical listening situations in everyday life (Wolters *et al.*, 2016). The listening situations were grouped into three main test conditions differing mainly in the subject's intention: (1) three samples of "Speech communication" (communication between the test person and one or two test leaders with and without cafeteria background noise), (2) one sample of "Focused listening" (watching and listening to a Youtube clip), and (3) one sample of "Other" condition (monitoring surroundings while performing vacuum cleaning). The set of five listening situations was repeated twice during a test session. Response categories were preference for "A," "B," or "Equal."

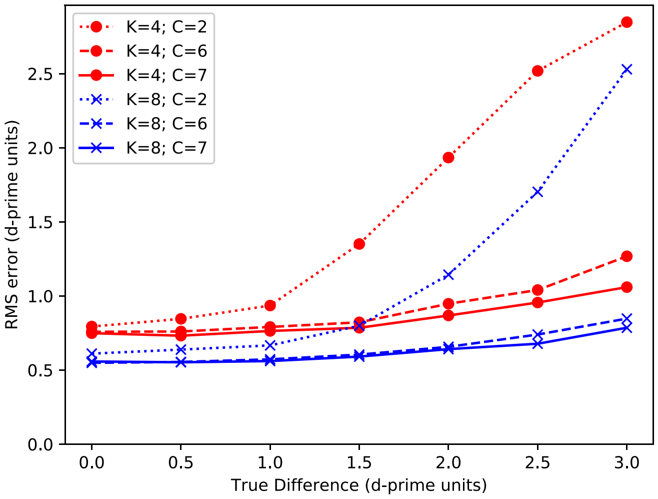


FIG. 4. (Color online) The rms deviation between the estimated and the true individual quality difference between two objects, plotted vs the mean difference in the population. The posterior parameter distributions were estimated using the Thurstone analysis model [Eq. (2)] for simulated experiments with forced-choice binary $(C = 2)$ and graded $(C = 6)$ as well as non-forced graded $(C = 7)$ response categories, with $K \in \{4, 8\}$ pair presentations for each of $N = 20$ "subjects" in each test group. Each data point shows the average across 100 group simulations.

IV. RESULTS

A. Simulation experiments

1. Illustrative example

Figure 2 shows that the Thurstone analysis model [Eq. (2)] yields quality estimates very close to the true values, regardless of whether the Thurstone or the BTL model was used to generate the simulated paired-comparison responses. The results from the BTL model are not shown because they are very similar, except for the different scale unit. The estimated *joint credibility* was 80% for the combined result that $\mu_B < \mu_A < \mu_C$ for a random individual in the population. The joint credibility measure includes the effect of multiple hypothesis tests so no further correction is needed.

In Fig. 2, the estimated 90% credible intervals for the population mean of objects B and C included the true value. This would be expected to happen in 90% of all simulated experiments. Among 100 complete simulations with $N = 20$ subjects, the true value was covered in 87% of the estimated intervals with the Thurstone analysis model and 82% with the BTL model. In other simulations with an inter-individual standard deviation of 1 *d*-prime unit, the empirical coverage with $N = 5$ subjects in each group was 87% with the Thurstone analysis model and 88% with the BTL model. With $N = 20$ subjects in each group, the empirical coverage was 90% and 89%, respectively. These results are all consistent with the nominal credibility of 90%.

2. Credibility versus experimental effort

The proposed analysis method can reveal quite small quality differences with high statistical credibility. The results in Fig. 3(A) show that a quality difference of 1 *d*-prime unit was indicated with credibility about 90% for a random individual in the population, when $K = 4$ paired-comparison responses were given by each listener in a group of $N = 20$ participants. The credibility is naturally slightly reduced if the test participants sometimes lose concentration and respond at random, as shown in Fig. 3(B). The BTL analysis model [Eq. (3)] consistently identified the quality difference with slightly higher credibility than the Thurstone analysis model [Eq. (2)].

3. Number of response alternatives

Individual root-mean-square deviations plotted in Fig. 4 show that allowing graded difference magnitudes can improve the precision dramatically compared to the common practice of using just binary responses, especially when there is a large difference between objects. The change from six to seven response alternatives also tended to reduce the deviation.

The plotted rms deviations include the variance of individual posterior parameter distributions, given the response data, as well as the random variability in the responses, given the true individual parameters. Results were very similar with the BTL analysis model.

B. Real evaluation of hearing aids

Results from a real evaluation of two hearing-aid programs are shown in Fig. 5 for users comparing the programs in three conditions with different listening intentions: “Speech Communication,” “Focused listening,” and “Other,” as defined in Sec. III B. For a *random individual* in the population from which test participants were recruited, program A is predicted to be better than program B in situations with Speech (credibility = 72%) and Focused listening (credibility = 88%), whereas program B is predicted to be better for the “Other” condition, which includes “Passive listening” (credibility = 87%). The estimated joint credibility was 99.8% for the result that the *population mean* preference for program B was higher in situation “Other” than in the two other situations. Using the average of ordinal ratings for B vs A by each subject in the three situations, a Friedman test indicated that the observed differences in the *population mean* are statistically highly significant ($N = 10, \chi^2 \approx 12.88, p \approx 0.16\%$).

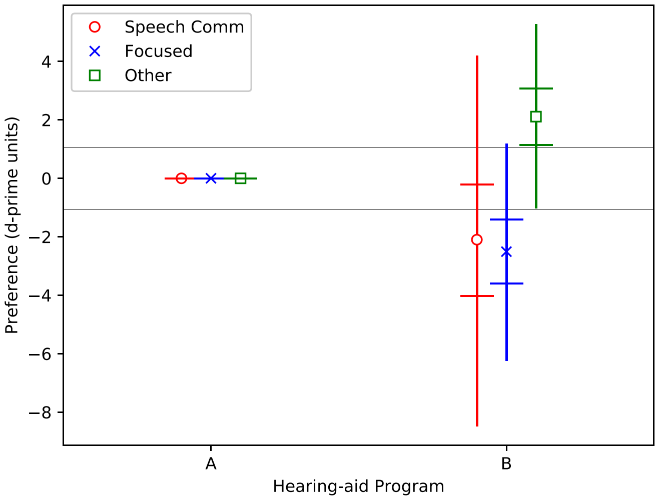


FIG. 5. (Color online) Preference for hearing-aid program A vs B, measured with paired comparisons by $N = 10$ hearing-aid users in three main test conditions. Medians (○, ×, □) and 90% symmetric credible intervals (vertical lines) are estimated for a random individual in the population from which the users were recruited. Short horizontal lines show 90% credible intervals for the population means. Long horizontal lines at about ±1 show the median estimated threshold between responses “Equal” and “Different” in either direction.

V. DISCUSSION

The presented new analysis method can estimate quality parameters (1) for a random individual in the group of test participants, (2) for a random individual in the population from which test participants are recruited, and/or (3) for the population mean. The validation with simulated paired-comparison data indicated that the proposed method can estimate quality parameters quite accurately. The estimated credible intervals included the true value with approximately the desired probability.

It is interesting to note that responses with graded differences improved result precision dramatically, especially when the difference between tested objects was large. Contrary to what might be expected, allowing tied (“Equal”) responses did not cause any loss of precision, but rather tended to improve the precision in the estimated results.

The posterior joint distribution of all model parameters, given the observed data, cannot be expressed in a closed form, so approximations are inevitable. One method is to draw random samples from the total joint distribution. This "all-sampling" approximation becomes asymptotically exact in the limit with infinitely many samples. The variational approximation, defined in Appendix B, separates the distribution of *population* parameters from that of *individual* model parameters. The main advantage with this approach is that a closed-form parametric distribution can be derived for the population parameters, while a sampling representation is used only for the individual parameters. This makes it easy to algebraically integrate out some parameters to derive the *predictive* population distributions, which are the main goals of the estimation, as defined in Sec. II C 5 and Appendix C. The "all-sampling" approach would have to implement the integration by an additional sampling step.

A known potential weakness in the proposed analysis model might be the prior assumption that model parameters are mutually uncorrelated in the population. We believe this choice of prior is defensible, because by using fewer population parameters the analysis model can handle experimental data sets including rather few participants, as discussed in a footnote[1] to App. A. The Bayesian analysis results include predictive error measures (credible intervals) for the given amount of data. The simulations indicated that those error measures were reasonably accurate. An extension allowing correlated population parameters is considered for a future version of the model.

As noted in Sec. II B, the present analysis method assumes a unidimensional decision space. It may be possible to extend the analysis of graded ordinal responses using some form of multidimensional scaling, but we must leave this for future research.

The proposed analysis method has been implemented as a python package `PairedCompCalc`, freely available at the Python Package Index. The implementation is more general than exemplified in this paper: the package can handle several objects, listener groups, test conditions, and perceptual attributes in a single analysis. The code package also includes simulation functions allowing the user to validate the performance of the method and to plan a practical experiment. Since the model illustrated in Fig. 1 can also be formulated in an ordinal-regression framework, the analysis might also be implemented using a general-purpose code package for regression (Bürkner and Vuorre, 2019).

VI. CONCLUSION

A new Bayesian parametric analysis method for paired-comparison data was presented. The method was evaluated with simulated experimental data and exemplified using data from a real experiment. We conclude the following answers to our main research questions:

(1) The Thurstone and BTL models gave similar results close to the true values. The estimated credibility of quality differences was slightly higher with the BTL model.

(2) Allowing graded ordinal responses in the experimental procedure improved the precision dramatically, compared to using only forced-choice binary responses, especially when the real difference was large.

(3) Allowing tied ("Equal") responses tended to improve precision.

(4) When applied to real evaluation data, the analysis revealed clinically interesting results with high statistical credibility for a random individual in the population, although the amount of test data was limited.

ACKNOWLEDGMENTS

This research was supported by Widex A/S. The authors would like to thank the editor and two anonymous reviewers for constructive critical comments on an earlier version of the paper.

APPENDIX A: PRIOR DISTRIBUTIONS

The prior distribution of individual model parameters in the population is assigned a conditional multivariate Gaussian density with independent[1] elements

$$p(\boldsymbol{U}_n|\boldsymbol{V},\Lambda)=\prod_{d=1}^{D}\sqrt{\frac{\lambda_d}{2\pi}}e^{(-1/2)(U_{nd}-V_d)^2\lambda_d}, \tag{A1}$$

given the population mean $\boldsymbol{V}=(\ldots,V_d,\ldots)$ and precision (inverse variance) matrix $\Lambda=\mathrm{diag}[\ldots,\lambda_d,\ldots]$. A conventional Gaussian-gamma prior is defined for the population parameters as

$$p(\boldsymbol{V},\Lambda)=\prod_{d=1}^{D}p(V_d|\lambda_d)p(\lambda_d); \tag{A2}$$

$$p(V_d|\lambda_d)=\sqrt{\frac{\beta'\lambda_d}{2\pi}}e^{(-1/2)(V_d-m'_d)^2\beta'\lambda_d}; \tag{A3}$$

$$p(\lambda_d)=\frac{{b'_d}^{a'}}{\Gamma(a')}\lambda_d^{a'-1}e^{-b'_d\lambda_d}. \tag{A4}$$

Here, $\Gamma(z)=\int_0^\infty x^{z-1}e^{-x}\,dx$ is the gamma function. In the absence of prior information, we assign all $m'_d=0$. The Jeffreys prior for the mean and precision of a Gaussian distribution would suggest $\beta'\to 0, a'\to 0, b'_d\to 0$. However, to avoid computational indeterminacy in case of extreme response patterns, we must use a weakly informative prior.

The effective weight of the prior, relative to the weight of one real test subject, is assigned as $\beta'=0.2$. We choose $a'=\beta'/2$ and $b'_d=\sigma_d^2/2$, with $\sigma_d=1$ for all d. This means that individual deviations from the population mean have a typical prior scale σ_d.

APPENDIX B: VARIATIONAL LEARNING

We use variational inference (VI) (Bishop, 2006, Chap. 10) to learn a good approximation $q(\underline{\boldsymbol{U}},\boldsymbol{V},\Lambda)\approx p(\underline{\boldsymbol{U}},\boldsymbol{V},\Lambda|\underline{\boldsymbol{R}})$ to the true posterior distribution of all individual models, $\underline{\boldsymbol{U}}$, and all population parameters, given all observed data $\underline{\boldsymbol{R}}$. The VI procedure maximizes a lower bound

$$\mathcal{L}(q) = \left\langle \ln \frac{p(\underline{\boldsymbol{R}}, \underline{\boldsymbol{U}}, \boldsymbol{V}, \Lambda)}{q(\underline{\boldsymbol{U}}, \boldsymbol{V}, \Lambda)} \right\rangle_q \leq \ln p(\underline{\boldsymbol{R}}), \tag{B1}$$

to the data log-likelihood and minimizes the Kullback-Leibler divergence between the approximate and exact posterior model distributions. Here, the symbol $\langle\cdot\rangle_q$ means the expectation calculated using the current estimate of $q()$ in each learning step. The procedure is guaranteed to converge. In this application, it is sufficient to enforce a partial factorization between the distributions of individual and population models as

$$q(\underline{\boldsymbol{U}}, \boldsymbol{V}, \Lambda) = q(\underline{\boldsymbol{U}})q(\boldsymbol{V}, \Lambda). \tag{B2}$$

The total log-likelihood of all observed data and all model parameters is

$$\begin{aligned}\ln p(\underline{\boldsymbol{R}}, \underline{\boldsymbol{U}}, \boldsymbol{V}, \Lambda) = \text{const.} &+ \sum_{n=0}^{N-1}\sum_{p=0}^{P_n-1} L(R_{np}|\boldsymbol{U}_n)\\ &+ \sum_{n=0}^{N-1}\sum_{d=1}^{D} \frac{1}{2}\ln\lambda_d - \frac{1}{2}(U_{nd} - V_d)^2\lambda_d\\ &+ \sum_{d=1}^{D} \frac{1}{2}\ln\beta'\lambda_d - \frac{\beta'}{2}(V_d - m'_d)^2\lambda_d\\ &+ \sum_{d=1}^{D}(a' - 1)\ln\lambda_d - b'_d\lambda_d,\end{aligned} \tag{B3}$$

where the likelihood function $L()$ was defined in Eq. (4). Using the standard VI solution for any factorized approximation (Bishop, 2006, Chap. 10), we find

$$\begin{aligned}\ln q(\boldsymbol{V}, \Lambda) = \langle \ln p(\underline{\boldsymbol{R}}, \underline{\boldsymbol{U}}, \boldsymbol{V}, \Lambda)\rangle_{q(\underline{\boldsymbol{U}})} = \text{const.}&\\ + \sum_{d=1}^{D} \frac{1}{2}\ln\lambda_d - \frac{1}{2}\underbrace{(N + \beta')}_{\beta} V_d^2\lambda_d&\\ + \sum_{d=1}^{D} V_d\lambda_d \underbrace{\left(\beta' m'_d + \sum_{n=0}^{N-1}\langle U_{nd}\rangle\right)}_{\beta m_d}&\\ + \sum_{d=1}^{D} - \frac{\beta'}{2} m'^2_d\lambda_d - \frac{1}{2}\sum_{n=0}^{N-1}\langle U_{nd}^2\rangle\lambda_d&\\ + \sum_{d=1}^{D}(N/2 + a' - 1)\ln\lambda_d - b'_d\lambda_d.&\end{aligned} \tag{B4}$$

This is the logarithm of a new Gaussian-gamma density with independent elements for each d, like Eq. (A2). The conditional Gaussian part is

$$q(V_d|\lambda_d) = \sqrt{\frac{\beta\lambda_d}{2\pi}} e^{(-\beta/2)(V_d - m_d)^2\lambda_d}, \tag{B5}$$

with updated parameters $\beta = \beta' + N$ and

$$\beta m_d = \beta' m'_d + \sum_{n=0}^{N-1}\langle U_{nd}\rangle. \tag{B6}$$

The remaining gamma density for the precision is

$$q(\lambda_d) \propto \lambda_d^{a-1} e^{-b_d\lambda_d}, \tag{B7}$$

with updated parameters $a = a' + N/2$ and

$$b_d = b'_d + \frac{\beta'}{2}(m'_d - m_d)^2 + \frac{1}{2}\sum_{n=0}^{N-1}\langle (U_{nd} - m_d)^2\rangle. \tag{B8}$$

As Eq. (B3) is a sum of terms involving each $\boldsymbol{U}_n$, the variational $q(\underline{\boldsymbol{U}}) = \prod_n q(\boldsymbol{U}_n)$ is naturally factorized without any further approximation, with

$$\begin{aligned}\ln q(\boldsymbol{U}_n) = \text{const.} &+ \sum_{p=0}^{P_n-1} L(R_{np}|\boldsymbol{U}_n)\\ &- \frac{1}{2}\sum_{d=1}^{D}(U_{nd} - m_d)^2\langle\lambda_d\rangle.\end{aligned} \tag{B9}$$

Here, the last sum represents the hierarchical influence of the current population model on the individual models. There is no closed form for $q(\boldsymbol{U}_n)$, but the density can be effectively represented by a large set of equally probable sample vectors, $\boldsymbol{u}_{ns}, s = 0, \ldots, n_s - 1$, drawn from the distribution by Hamiltonian sampling (Neal, 2011). The Hamiltonian sampler uses only the log-likelihood function [Eq. (B9)] and its gradient. For the results presented in this paper, $n_s = 1000$ vectors were sampled from each $q(\boldsymbol{U}_n)$. The expectation of any function of $\boldsymbol{U}_n$ is approximated by the average across samples. Thus, $\langle \boldsymbol{U}_n\rangle \approx (1/n_s)\sum_s \boldsymbol{u}_{ns}$, and $\langle (U_{nd} - m_d)^2\rangle \approx (1/n_s) \sum_s (u_{nsd} - m_d)^2$.

To monitor the progress of VI learning, we calculate the lower bound $\mathcal{L}(q)$ in Eq. (B1) most conveniently as

$$\begin{aligned}\mathcal{L}(q) = \langle \ln p(\underline{\boldsymbol{R}}|\underline{\boldsymbol{U}})\rangle_q + \langle \ln p(\underline{\boldsymbol{U}}|\boldsymbol{V}, \Lambda)\rangle_q&\\ - \langle \ln q(\underline{\boldsymbol{U}})\rangle_q - \left\langle \ln \frac{q(\boldsymbol{V}, \Lambda)}{p(\boldsymbol{V}, \Lambda)}\right\rangle_q.&\end{aligned} \tag{B10}$$

The first two terms are calculated during the sampling operation using Eq. (B9). The third term is the sum of entropy for each $\boldsymbol{U}_n$, which is calculated from the samples using a nearest-neighbour ("Kozachenko-Leonenko") estimator (Singh and Poczos, 2016). The last term subtracts the Kullback-Leibler divergence $\mathrm{KL}(q||p)$.

APPENDIX C: PREDICTIVE RESULTS

The learned individual and population models are used to calculate three predictive distributions:

(1) The predictive distribution for a random individual drawn from the *group of participants* is the mixture density

$$q(\boldsymbol{U}_N) = \frac{1}{N}\sum_{n=0}^{N-1} q(\boldsymbol{U}_n), \tag{C1}$$

represented by the joined sets of samples of all $q(\boldsymbol{U}_n)$.

(2) The predictive distribution for the dth parameter of a random unknown individual in the *population* from which the test group was recruited is the marginal density, integrated over all posterior parameters

$$p(U_{Nd}) = \int p(U_{Nd}|V_d, \lambda_d) q(V_d|\lambda_d) q(\lambda_d)\, dV_d\, d\lambda_d \propto \left(1 + \frac{(U_{Nd} - m_d)^2 \beta}{2b_d(\beta + 1)}\right)^{-(2a+1)/2}, \qquad \text{(C2)}$$

using the conditional Gaussian $p(\boldsymbol{U}_{Nd}|V_d, \lambda_d)$ in Eq. (A1).

(3) The predictive distribution of the *population mean* (equal to the median) is the marginal density

$$q(V_d) = \int q(V_d|\lambda_d) q(\lambda_d)\, d\lambda_d \propto \left(1 + \frac{(V_d - m_d)^2 \beta}{2b_d}\right)^{-(2a+1)/2}. \qquad \text{(C3)}$$

The distributions in Eqs. (C2) and (C3) are both univariate Student-t distributions with location m_d and degrees-of-freedom $\nu = 2a$. The Student-t scale parameter is $\sqrt{b_d(\beta+1)/a\beta}$ in Eq. (C2) and $\sqrt{b_d/a\beta}$ in Eq. (C3).

[1] Another possibility is to allow a full covariance matrix in (A1) and a Gaussian-Wishart population prior in (A2). However, this would require $D(D + 1)/2$ instead of D free precision parameters and a correspondingly larger $N \propto D^2$ to yield reliable estimates. A Wishart prior would require $N > D - 1$ participants to allow any proper population predictions at all, while the diagonal covariance only requires $N > 2$.

Agresti, A. (**1992**). "Analysis of ordinal paired comparison data," J. R. Stat. Soc. Ser. C **41**(2), 287–297.

Allen, J. B., and Neely, S. T. (**1997**). "Modeling the relation between the intensity just-noticeable difference and loudness for pure tones and wideband noise," J. Acoust. Soc. Am. **102**(6), 3628–3646.

Amlani, A. M., and Schafer, E. C. (**2009**). "Application of paired-comparison methods to hearing aids," Trends Amplif. **13**(4), 241–259.

Bishop, C. M. (**2006**). *Pattern Recognition and Machine Learning* (Springer, New York), Chap. 10, pp. 461–522.

Böckenholt, U. (**2001**). "Hierarchical modeling of paired comparison data," Psychol. Methods **6**(1), 49–66.

Bradley, R. A., and Terry, M. E. (**1952**). "Rank analysis of incomplete block designs. I. The method of paired comparisons," Biometrika **39**, 324–345.

Bürkner, P.-C., and Vuorre, M. (**2019**). "Ordinal regression models in psychology: A tutorial," Adv. Methods Pract. Psychol. Sci. **2**(1), 77–101.

Cattelan, M. (**2012**). "Models for paired comparison data: A review with emphasis on dependent data," Stat. Sci. **27**(3), 412–433.

Dahlquist, M., and Leijon, A. (**2003**). "Paired-comparison rating of sound quality using MAP parameter estimation for data analysis," in *[AQS-2003] First ISCA Tutorial and Research Workshop on Auditory Quality of Systems*, April 23–25, Mont-Cenis, Germany, pp. 79–84.

Dillon, H. (**1984**). "A procedure for subjective quality rating of hearing aids," NAL report No. 100 (Australian Government Publishing Service, Sydney, Australia).

Durlach, N., and Braida, L. (**1969**). "Intensity perception. I. Preliminary theory of intensity resolution," J. Acoust. Soc. Am. **46**(2), 372–383.

Garner, W. R. (**1952**). "An equal discriminability scale for loudness judgments," J. Exp. Psychol. **43**(3), 232–238.

Grancharov, V., and Kleijn, W. B. (**2008**). "Speech quality assessment," in *Springer Handbook of Speech Processing*, edited by J. Benesty, M. M. Sondhi, and Y. Huang (Springer Verlag, Berlin, Germany), Chap. 5, pp. 83–102.

Green, D. M., and Swets, J. A. (**1988**). *Signal Detection Theory and Psychophysics* (Peninsula Publ, Los Altos, CA).

Houtsma, A., Durlach, N., and Braida, L. (**1980**). "Intensity perception. XI. Experimental results on the relation of intensity resolution to loudness matching," J. Acoust. Soc. Am. **68**(3), 807–813.

IEEE (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **AU-17**(3), 227–246.

ITU (**2003**). ITU-R-BS1283, *Methods for the Subjective Assessment Of Sound Quality—A Guide To Existing Recommendations* (ITU, Geneva, Switzerland).

Laugesen, S., Jensen, N. S., Rønne, F. M., and Pedersen, J. H. (**2015**). "Can individualised acoustical transforms in hearing aids improve perceived sound quality?," in *Proceedings of the International Symposium on Auditory and Audiological Research*, May 27–30, Nyborg, Denmark, pp. 245–252.

Leijon, A., Henter, G. E., and Dahlquist, M. (**2016**). "Bayesian analysis of phoneme confusion matrices," IEEE Trans. Audio Speech Lang. Proc. **24**(3), 469–482.

Luce, R. D. (**1959**). *Individual Choice Behavior; A Theoretical Analysis* (Wiley, New York).

Neal, R. M. (**2011**). "MCMC using Hamiltonian dynamics," in *Handbook of Markov chain Monte Carlo*, edited by S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng (Chapman and Hall/CRC Press, Boca Raton, FL), Chap. 5, pp. 113–162.

Pérez-Ortiz, M., and Mantiuk, R. (**2017**). "A practical guide and software for analysing pairwise comparison experiments," arXiv:1712.03686.

Rao, P. V., and Kupper, L. L. (**1967**). "Ties in paired-comparison experiments: A generalization of the Bradley-Terry model," J. Am. Stat. Assoc. **62**(317), 194–204.

Riesz, R. R. (**1933**). "The relationship between loudness and the minimum perceptible increment of intensity," J. Acoust. Soc. Am. **4**(3), 211–216.

Singh, S., and Poczos, B. (**2016**). "Analysis of k-nearest neighbor distances with application to entropy estimation," arXiv:1603.08578.

Smeds, K., Dahlquist, M., Larsson, J., Herrlin, P., and Wolters, F. (**2019**). "LEAP, a new laboratory test for evaluating auditory preference," in *Proceedings of the 23rd International Congress on Acoustics*, September 9–13, Aachen, Germany, pp. 7608–7615, available at http://pub.dega-akustik.de/ICA2019/data/articles/000162.pdf.

Thurstone, L. L. (**1927**). "A law of comparative judgment," Psychol. Rev. **34**(4), 273–286.

Tsai, R.-C., and Böckenholt, U. (**2002**). "Two-level linear paired comparison models: Estimation and identifiability issues," Math. Soc. Sci. **43**(3), 429–449.

Tsukida, K., and Gupta, M. R. (**2011**). "How to analyze paired comparison data," UWEE Technical Report No. UWEETR-2011-0004 (University of Washington, Seattle, WA).

Tversky, A. (**1972**). "Elimination by aspects: A theory of choice," Psychol. Rev. **79**(4), 281–299.

Wickelmaier, F., and Schmid, C. (**2004**). "A MatLab function to estimate choice model parameters from paired-comparison data," Behav. Res. Methods Instrum. Comput. **36**(1), 29–40.

Wolters, F., Smeds, K., Schmidt, E., and Norup, C. (**2016**). "Common sound scenarios: A context-driven categorization of everyday sound environments for application in hearing-device research," J. Am. Acad. Audiol. **27**(7), 527–540.

Zermelo, E. (**1929**). "Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung" ("Calculating tournament results as a maximization problem in probability theory"), Math. Z. **29**(1), 436–460.

Zimmer, K., Ellermeier, W., and Schmid, C. (**2004**). "Using probabilistic choice models to investigate auditory unpleasantness," Acta Acust. united Ac. **90**(6), 1019–1028.