scientific reports

OPEN

Using mobile EEG to study auditory work strain during simulated surgical procedures

Marc Rosenkranz[1,2], Thorge Haupt[1], Manuela Jaeger[1], Verena N. Uslar[2] & Martin G. Bleichner[1,3✉]

Surgical personnel face various stressors in the workplace, including environmental sounds. Mobile electroencephalography (EEG) offers a promising approach for objectively measuring how individuals perceive sounds. Because surgical performance does not necessarily decrease with higher levels of distraction, EEG could help guide noise reduction strategies that are independent of performance measures. In this study, we utilized mobile EEG to explore how a realistic soundscape is perceived during simulated laparoscopic surgery. To examine the varying demands placed on personnel in different situations, we manipulated the cognitive demand during the surgical task, using a memory task. To assess responses to the soundscape, we calculated event-related potentials for distinct sound events and temporal response functions for the ongoing soundscape. Although participants reported varying degrees of demand under different conditions, no significant effects were observed on surgical task performance or EEG parameters. However, changes in surgical task performance and EEG parameters over time were noted, while subjective results remained consistent over time. These findings highlight the importance of using multiple measures to fully understand the complex relationship between sound processing and cognitive demand. Furthermore, in the context of combined EEG and audio recordings in real-life scenarios, a sparse representation of the soundscape has the advantage that it can be recorded in a data-protected way compared to more detailed representations. However, it is unclear whether information get lost with sparse representations. Our results indicate that sparse and detailed representations are equally effective in eliciting neural responses. Overall, this study marks a significant step towards objectively investigating sound processing in applied settings.

Surgical personnel often experience high levels of stress, which can lead to severe health problems such as burnout[1] or hypertension[2,3]. One cause of stress is distractions due to the environment of the operating room (OR)[4,5]. The soundscape in the OR is highly complex[6], comprising of sounds that are crucial to the surgery (e.g, communication, tool usage), as well as sounds that are not crucial to the surgery (e.g., instrument clattering, phone ringing) and could be minimized to improve the work environment[7]. The accumulation of the different sounds sources, leads to high sound levels which are often perceived as distracting and stressful[8–13]. To guide interventions that aim at reducing stress induced by auditory distractions it is important to understand and measure how ongoing sounds in the OR affect the personnel.

An objective evaluation of the subjectively experienced burden of the OR soundscape is challenging. Previous studies that have focused on the effect of sound on performance (i.e., surgery task performance) have found that only under extreme and unnatural conditions, such as dichotic listening to two pieces of music[14,15], but not under more natural conditions, such as a single stream of OR sounds[16–21], more mistakes in the surgery task were observable. However, medical personnel may strive to perform at high levels under adverse working conditions because mistakes can have serious consequences for the patient. Thus, the work strain surgeons experience can not necessarily be inferred from the surgery task performance. Therefore, additional measures are required to objectively assess the auditory strain in the OR.

To measure sound perception in the OR objectively, electroencephalography (EEG) is a promising method. Given its temporal resolution, it is particularly useful for measuring responses that are time-locked to sounds. By analyzing event-related potentials (ERPs) the neural response to individual sound events can be examined,

[1]Neurophysiology of Everyday Life Group, Department of Psychology, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany. [2]Pius-Hospital Oldenburg, University Hospital for Visceral Surgery, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany. [3]Research Center for Neurosensory Science, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany. ✉email: martin.georg.bleichner@uni-oldenburg.de

revealing perceptual and cognitive processes such as whether a sound is considered task-relevant[22]. Another approach involves the use of temporal response functions (TRFs) to study responses to continuous sound streams[23], which expands the use of EEG to naturalistic soundscapes. ERPs and TRFs can be measured outside of the laboratory while individuals are freely moving[24,25] and to complex soundscapes while an audio-visual-motor task is being performed[26]. The development of head-mounted EEG systems and wireless data transmission allows for unrestricted mobility for the wearer, making it applicable in various work environments[27] and enabling measurements over extended periods of time[28].

How we process distractors depends on contextual factors, including the demands of the task[29], and is reflected in the neural responses measured by EEG[30]. Thus, EEG signals can provide valuable insights into the cognitive demand experienced by an individual, independent of their task performance. This makes EEG a potentially useful tool for continuously monitoring a surgeon's perceived demand in real time. Such monitoring could guide the implementation of sound interventions in the OR by addressing not only the acoustic environment (e.g., loudness)[7], but also the individual's perceived demand. Furthermore, cognitive demand influences the ability of a person to shield themselves from auditory distractions[31]. Personnel in the OR report to be more vulnerable to distractions from the soundscape during periods of high compared to low demand[12]. It is reasonable to expect that both neural and subjective responses to the OR soundscape will differ based on the cognitive demands of the task being performed. By varying the overall cognitive demand, we can study how it impacts perception and processing of soundscapes, which in turn helps us to understand auditory distraction.

We investigated demand-dependent changes in EEG responses to the soundscape, as well as subjective and behavioural measures. We focused on ERP and TRF time-windows typically found in response to sounds, namely the N1, P2, and N2 time-window[32–34]. Previous studies reported mixed findings regarding the effect of varying demand on these time-windows[30]. Therefore, we investigated the effect of varying demand on each time-window.

Surgeons are frequently required to perform complex procedures for extended periods. This naturally leads to fluctuations in cognitive demand and how the environment is perceived. For example, sounds that were once easily ignored can become sources of distraction, or vice versa[12]. This change in processing, compounded by varying demand, underscores the critical need to consider the temporal dynamics during the performance of surgical tasks. Therefore, we explored changes over time for the ERP and TRF time-windows in response to the soundscape, as well as for the subjective and behavioral measures.

At last, there are practical considerations when opting to use TRFs in an applied context. Since the derivation of TRFs require information of the soundscape, sound recordings have to adhere to privacy concerns. To test whether representations of the soundscape void of personal information, produce similar results as rich representations we derived sound onset marker[35] in addition to the commonly used acoustic envelope[23]. Sound onsets only indicate sound occurrence, providing a data protected way of sound recording. Previous research has demonstrated that onsets and envelopes yield comparable results for computing TRFs[36,37]. To replicate this finding and strengthen the applicability of TRFs in real-life settings we compared the TRFs computed from the onsets and the envelopes of the soundscape.

The objective of this research was twofold. Firstly, it aimed to examine the influence of a realistic task under varying levels of cognitive demand on the individual processing of task-irrelevant auditory stimuli, reflecting an OR soundscape. This investigation integrated both subjective assessments (i.e., self-reports) and objective metrics (i.e., neural responses to irrelevant sounds and behavioral performance) to provide a comprehensive understanding of sound and distractor processing in work environments. Secondly, the study investigated the processing of continuous versus discrete acoustic features, offering insights into auditory responses to different aspects of the soundscape.

Methods

This study involving human participants was approved by the Medizinische Ethikkommission, Carl von Ossietzky Universität Oldenburg, Oldenburg (2021-031) and performed in accordance with the Declaration of Helsinki. The participants provided their written informed consent to participate in this study. This study was preregistered after data collection https://doi.org/10.17605/OSF.IO/AE3UY. All changes to the preregistration can be found in the supplementary material.

Participants

23 medical students were recruited through an online announcement on the university board and word-of-mouth (age range: 19−24; 16 female). One participant was excluded due to data loss during recording resulting in 22 included participants. We based the sample size on our previous studies in which we measured reliable ERPs and TRFs during a complex and dynamic task[26]. Four participants had previous experience with a laparoscopic simulator and 17 attended at least once a surgery as an observer. All participants received monetary reimbursement. Eligibility criteria included: enrollment as a medical student and self-reported normal hearing, normal or corrected vision, no psychological or neurological condition, and right-handedness.

Paradigm

Figure 1 illustrates the setup and paradigm. To mimic an actual surgical environment, participants performed a laparoscopic surgery task while presented with the playback of a continuous OR soundscape. To vary the cognitive demand during the surgery task, we adapted a serial recall paradigm: Prior to the surgery task, participants had to remember two or eight letters, simulating low and high cognitive demand, respectively. They received the instruction to silently repeat those letters while performing the surgery task. After the surgery task, they should recall the letters. The memory task was selected for the following reasons. First, to simulate a varying general demand of surgeons that might not be necessarily related to the surgery. Second, the task was chosen as
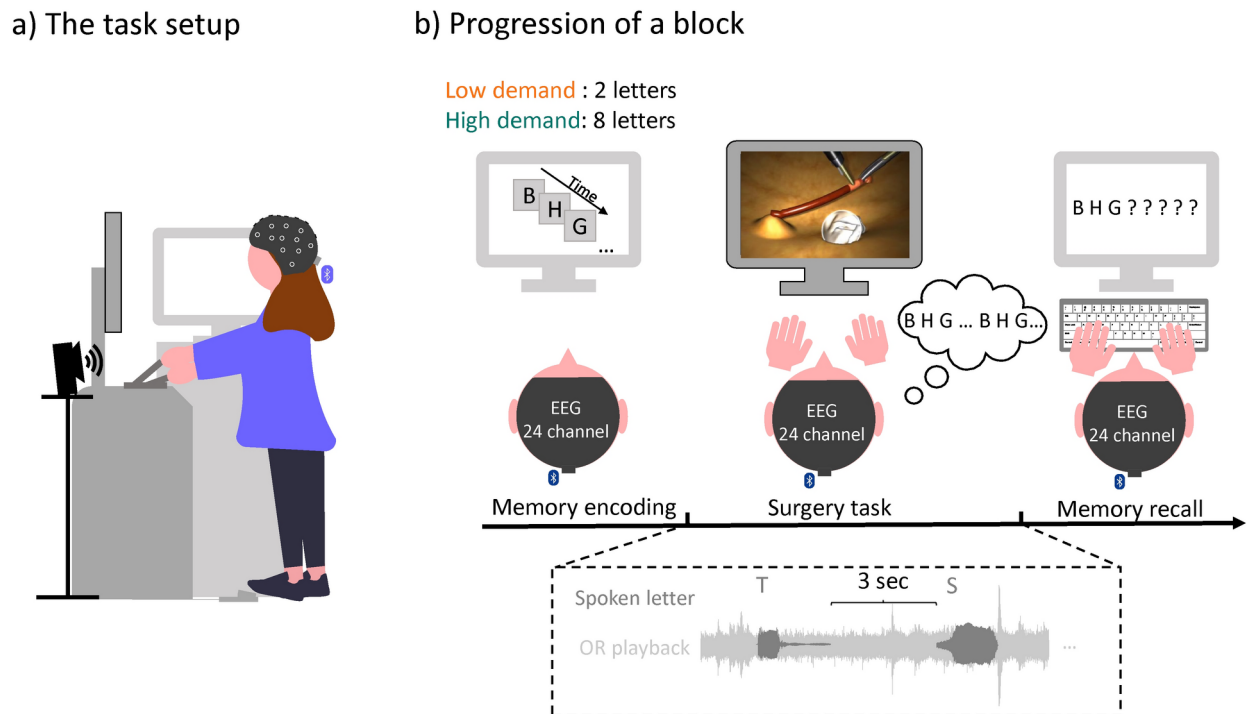
a) The task setup   b) Progression of a block

Figure 1. (**a**) Participants were standing throughout the entire experiment either in front of the surgery simulator or the screen for the memory task which were positioned in $90°$ to each other. The speaker were positioned at chest height and participants were equipped with a mobile EEG cap. (**b**) At the start of the block participant were presented with either two (low-demand condition) or eight (high-demand condition) letters which they had to remember. During the surgery they were presented with a continuous realistic soundscape of an operation room and with overlayed spoken letters. They were instructed to silently repeat the memorized letters during the surgery task. After finishing the surgery task they should enter the memorized letters and fill out a workload questionnaire. An experimental block ended with feedback regarding their memory performance.

visual working memory abilities correlate with surgical task performance[38,39], i.e., we manipulated the demand on the surgery task by manipulating the demand on visual memory. Third, the memory task did not require an overt response and hence did not interrupt the surgical procedure. Finally, we could increase the distractive potential of the task-irrelevant soundscape, by presenting spoken letters which are commonly used as distractors in classic serial recall paradigms[40,41]. The spoken letters were presented in the low- and high-demand condition. Overall this should simulate situations in which the surgeon must effectively shield themselves from irrelevant and potentially distracting information. Participants did not receive any instruction on what to do with the soundscape, and were free to disregard it. After retrieval of the letters, participants rated their perceived demand during the task and received feedback regarding their memory performance. Each participant performed a total of 28 experimental blocks, 14 of each condition, in a randomized order, with the restriction that seven blocks of each condition were presented during the first and second half of the experiment. During the entire experiment, participants were standing either in front of the screen for the memory task (Samsung, SyncMaster P2470) or in front of the surgery simulator. The monitor and simulator were positioned in $90°$ to each other. Further details can be found below.

Surgery task For the surgery task, the LabSim® (Surgical Science, Sweden) simulator was used. This simulator is used to train hand-eye coordination for laparoscopic surgeries and provides realistic graphics and tactile feedback. The setup includes specialized instruments for both hands, a foot pedal for additional controls, and a screen that displays the virtual surgical instruments in action.

The specific task which the participant performed throughout the experiment is called "cutting". In this task, participants needed to grasp a vessel with the right-hand instrument, stretch it, and cut it with an electrical cutting device which was handled with the left hand. To activate the cutter, a foot paddle must be pressed. The vessel must then be placed in a small bag. The task ends when three parts of the vessel are removed.

Participants were informed that their performance would be evaluated based on the duration to complete the surgery task, mistakes, and tissue damage. Mistakes included rupturing the vessel (e.g., by exerting too much pull on the vessel) or dropping the cut vessel, thus a block could have zero to three mistakes. Touching the tissue with the instrument caused immediate visual feedback as the screen received a red shade. After each task, they received visual feedback about their overall performance.

Memory task A serial recall task was adapted to vary the cognitive demand during the surgery task. Although the serial recall task is a classic working memory task, we refrain from calling it one, as the retention interval

(i.e., the performance of the surgery task) was too long. Prior to the surgery task, participants were asked to remember the visually presented letters in the correct order. For the low-demand condition two letters and for the high-demand condition eight letters had to be remembered. The letters were randomly selected from a set of twelve letters (B,C,D,F,H,K,L,M,P,Q,S,T) without replacement. Letters were presented in black on a gray screen for 1000 ms each with an inter-stimulus-interval of 500 ms. After the surgery task, either two or eight question marks indicated that the letters should be entered using a keyboard. Participants could enter an "X" for letters they could not remember and could correct themselves.

Soundscape The OR playback was recorded using a field recorder which was positioned close to the surgery table during a visceral surgery at the University Hospital Oldenburg[42]. The recording contains a variety of sounds, such as ventilation noise, beeps from monitoring devices, instrument clatter, and instrument sounds. Intelligible speech was removed after the recording for privacy reasons, however, unintelligible muttering and non-vocal sounds such as coughing were preserved. The recording lasts approximately 1 hour. In the first block, the recording starts at a random time point and continuous chronologically for every subsequent block. If the recording reached the end, it started at the beginning. The soundscape started automatically after the letters of the memory task were presented. To prevent clicking noise when the audio starts the starting 500 milliseconds were faded in.

In addition to the realistic soundscape we presented a sequence of individual spoken letters in both conditions. The method of presentation of the spoken letters was consistent across both conditions: Four letters were drawn from the same set of letters as the memory task, but never coincided with the to-be-remembered letters of a block. To ensure that the letters were presented equally often within a block, the four letters were presented as consecutive groups but were randomized within a group. The same letter was never presented consecutively. The inter-trial-interval between letters was three seconds. Letters were generated using a text-to-speech program (Notevibes, accessed 2021) and spoken by the same female voice. The letters were clearly audible in the soundscape recording and had an on- and offset ramp of ten milliseconds.

All sounds (i.e., the recording and letters) were sampled at a rate of 48 kHz and presented to the participant using Psychtoolbox 3 for MATLAB[43] (v3.0.17), a t.amp E4-130 amplifier (Thomann GmbH, Burgebrach, Germany) and presented as a stereo signal using two iLoudMTM loudspeakers (IK Multimedia Production srl, Modena, Italy). The loudspeakers were vertically tilted upwards $(20°)$ and located in front of the participant to the right and to the left of the LabSim˚ at chest height. The distance between loudspeakers was 0.5 m and the distance between loudspeaker and ear was 1.2 m. The sound pressure level was $45-55$ dB SPL, measured at the place of the participants head.

Subjective workload assessment To assess the subjectively perceived demand during each block we included three workload related questions. For a good representation of our research question we chose two items from the NASA-TLX[44], namely effort ("How hard did you have to work to accomplish your level of performance?") and frustration ("How insecure, discouraged, irritated, stressed, and annoyed were you?"). The effort item was chosen to investigate whether remembering eight compared to two words was perceived as more effortful, thereby linking it to our investigation of differences in demand. The frustration item was included to assess the emotional response to the high-demand task, particularly considering the potential stress induced to remember eight compared to two letters. Furthermore, one question from the SURG-TLX[45] was chosen, namely distraction ("How distracting was the operating environment"). This item was selected to investigate how the perception of the soundscape interacts with task-related demand. Each question was answered on a visual analog scale ranging from 0 to 20[44,45].

Training To familiarize themselves with the tasks, participants engaged in a series of practice blocks. During these sessions, the experimenter remained nearby, providing guidance and support to help participants navigate through the tasks. First, they performed the memory task twice, once with two letters and once with eight letters. Second, they performed a basic surgery task (i.e., instrument navigation) without the memory task or sounds to familiarize themselves with the LabSim˚. Third, they performed the cutting task without the memory task or sounds. Lastly, participants performed two training blocks that were identical with the blocks of the main experiment, first for the low-demand condition and then for the high-demand condition. After the training, the experiment started and participants performed the experimental blocks on their own.

Data acquisition
Participants were asked to wash their hair on the day of the recording and to not use hair styling products. EEG data were recorded using a wireless 24-channel amplifier (SMARTING, mBrainTrain, Belgrade, Serbia) attached to the back of the EEG cap (EasyCap GmbH, Hersching, Germany) with Ag/AgCl passive electrodes at $10-20$ layout positions (Fp1 Fp2 AFz Fz F3 F4 F7 F8 Cz C3 C4 T7 T8 CPz Pz M1 M2 P3 P4 P7 P8 POz O1 O2) with the reference and ground electrode at position FCz and Fpz, respectively. The data were recorded using a sampling rate of 500 Hz, and transmitted via Bluetooth from the amplifier to a Bluetooth dongle (BlueSoleil) that was plugged into a computer.

After fitting the cap, the skin was cleaned using 70% alcohol. Abrasive gel (Abralyt HiCl, Easycap GmbH, Germany) was used for reducing electrical impedance and ensuring high-quality signal. Impedances were kept below 20 kΩ at the beginning and again checked at the end of the recording using the SMARTING Streamer software (v3.4.3; mBrainTrain, Belgrade, Serbia).

ECG data were also recorded on a laptop but were not part of the current analyses.

Experimental markers (e.g., sound markers) were generated using the lab streaming layer library[46] (v1.14). The ECG recording laptop and EEG recording computer were connected via Lan. To synchronize all data streams, EEG data, ECG data, and experiment marker were collected in the Lab Recorder software (v1.14, https://github.com/labstreaminglayer/App-LabRecorder) and saved as one .xdf file on the EEG recording computer.

Preprocessing

EEG The EEG data were analyzed using EEGLAB[47] (v2022.0) in MATLAB R2020b (The MathWorks, Natick, MA, United States). For each participant, the continuous data were filtered with Hamming windowed FIR filter using the EEGLAB default settings: (1) high-pass: passband edge = 0.5 Hz (filter order = 3300, transition bandwidth = 0.5 Hz, cutoff frequency (−6dB) = 0.25 Hz); (2) low-pass: passband edge = 30 Hz (filter order = 220, transition bandwidth = 7.5 Hz, cutoff frequency (−6dB) = 33.75 Hz). For further analysis, only the EEG data during sound presentation were included, i.e., data during memory encoding and retrieval were removed. To minimize artifacts from switching between the surgery task screen and the memory task screen we removed the first and last five seconds of EEG data during sound presentation. Bad EEG channels were automatically rejected using the EEGLAB function *clean_artifacts* from the *clean raw data* plugin[48] following a procedure described by Klug & Gramann (2021)[49]. The function was executed over ten iterations with the following parameters: *ChannelCriterion* (0.8); *ChannelMaxBrokenTime* (0.5). The remaining parameters were turned off. If a channel was rejected in at least 50% of iterations, it was removed. A maximum of 5 channels could be removed to ensure that enough data were available for channel reconstruction. On average 0.72 (± 1.28) channels were rejected.

After bad channel removal, the data were cleaned from artifacts using infomax independent component analysis (ICA). For ICA, a copy of the preprocessed data were created and high-pass filtered (passband edge = 1 Hz, filter order = 825, transition bandwidth = 1 Hz, cutoff frequency (−6dB) = 0.5 Hz), and cut into consecutive epochs of one second. Improbable epochs with a global (all channels) or local (single channel) threshold exceeding 5 standard deviations were automatically rejected using the *jointprob* function. ICA decomposition was applied to the remaining epochs. The resulting components were back-projected on the original preprocessed, but uncleaned data. The components were then classified using the EEGLAB toolbox *ICLabel*[50] with the 'lite' classifier which is better at detecting muscle artifacts than the default classifier[51]. Components belonging to the categories eye blink and movement or muscle movement with 60% confidence were removed. Note, that the ICLabel classifier did not classify all components correctly because it was trained on stationary data with a larger electrode setup than ours. Therefore, we manually checked the components and made the following adjustments: We detected ICs located at the mastoids, probably from muscle movement (see Supplementary Figure S2 for an example). As the mastoids were used for re-referencing we manually removed these ICs. Furthermore, lateral eye movement also required manual removal in some cases. Afterwards, previously rejected channels were interpolated using spherical interpolation. Lastly, channels were re-referenced to the linked mastoids (M1/M2).

Audio For technical reasons, a constant delay between the marker that indicates a sound onset and the actual sound presentation was quantified beforehand. To correct for this constant delay the marker was shifted by 30 ms. Furthermore, the onsets of the spoken letters were also corrected. The letters were embedded in a constant sound stream which might lead to an energetic masking effect of the first few milliseconds of a letter. This affects when a letter becomes audible and therefore, the time when a brain response occurs. To obtain a better estimate when the participants could hear the letters, we used the OnsetDetector app[35] implemented in MATLAB which determined the first energetic peak of the letters. The markers were shifted between 0 to 12.83 ms (Supplementary figure S1).

In order to relate the ongoing soundscape to the ongoing neural response, acoustic features were extracted. For this, we only used the OR playback (i.e., without the letters). From the OR playback we extracted and compared three feature vectors, namely the envelope of the raw OR playback, the envelope of the noise-reduced OR playback, and the onsets of the OR playback.

The envelope of the OR playback was extracted using the *mTRFenvelope* function[23,52] with default inputs. To reduce the noise in the OR playback from ventilation and running machines, we used a Wiener filter implemented in MATLAB[53,54]. For this, we first high-pass filtered the OR playback at 1 Hz (filter order = 1000, transition bandwidth = 0.5 Hz, cutoff frequency (-6dB) = 0.00004 Hz). We then estimated the power spectral density of the noise using the first second of the OR playback, as it was representative of the static noise in the OR playback. The noise estimate was then subtracted from the remaining signal. Afterwards, we extracted the envelope from the noise-reduced OR playback using the *mTRFenvelope* function.

Onsets were calculated using the OnsetDetector App[35] implemented in MATLAB. As we aim to detect onsets in naturalistic settings the raw audio was used. The resulting feature vector contained zeros (i.e., no onset) and ones (i.e., onsets).

EEG analysis

We performed two types of analysis. An ERP analysis to study the responses to the onset of the spoken letters, and a TRF based analysis to study the response to the ongoing OR playback.

ERP calculation ERP analysis was performed for the spoken letters. For each letter, epochs from -200 to 600 ms with respect to the stimulus onset were extracted. A baseline correction from -200 to 0 ms prior to stimulus onset was performed. Improbable epochs with a global (all channels) or local (single channel) threshold exceeding 3 standard deviations were automatically rejected using the *jointprob* function. We then computed the average response of each participant and block.

TRF calculation A forward modeling approach was used to compute a temporal response function (TRF) that characterizes the brain's temporal response to a feature vector representing the auditory stimulus. To calculate the TRF we used the mTRF toolbox[23]. For the TRF analyses, EEG data were multiplied by factor .0313 for scaling (as suggested in the provided code by Crosse et al. (2016)[23]).

To evaluate which acoustic features best predicts the neural response we implemented a forward model based on individual EEG data using a 10-fold cross-validation approach. For this, we separated the blocks into 10 segments. With 28 blocks in total, each segment consisted of two to three successive blocks. We split the segments into training and testing data such that each segment was once test data and iterated through the following procedure. For the training procedure we determined a shrinkage regularization parameter λ using

the *mTRFcrossval* function with a time range from 0 to 450 ms time lag and a lambda range from 10^{-8} to 10^8. This resulted in a correlation value for each fold (i.e., number of TRF training blocks), lambda value, and channel. We averaged over folds and channels and used the lambda value that maximized the correlation for the subsequent TRF calculation. We then trained a forward model using the training blocks and the *mTRFtrain* function with time lags from 0 to 450 ms and the optimal lambda. The resulting model was used to predict the response of the test blocks using the *mTRFpredict* function. This resulted in a prediction value for each test block, channel, and iteration, which were averaged, leaving one prediction value per participant and acoustic feature.

Lastly, we computed a forward model for each block using *mTRFtrain* with time lags from −220 to 500 ms. For the individual optimal lambda value we chose the most frequently occurring one during the cross-validation procedure. As the scale of TRFs varied across participants, biasing statistical comparisons of amplitudes, we z-scored each participant's TRF weights across time-points, channel, and blocks.

GED analysis To evaluate amplitude differences of the ERP and TRF and avoid channel selection, we use generalized eigenvalue decomposition (GED) as a spatial filter following guidelines by Cohen (2022)[55]. In short, for each ERP and TRF time-window (i.e., N1/P2/N2) we computed a generic spatial filter across subjects that was then applied to the data. GED maximizes the contrast between a signal covariance matrix S and a reference covariance matrix R. For S, we computed N1, P2, and N2 time-windows and contrasted each time-window separately against the baseline period (i.e., −200 to 0 ms).

In detail, we first average ERPs and TRFs across blocks, resulting in one time-series for each participant. We then determined the N1/P2/N2 peak of each ERP and TRF. Regarding the ERPs, and TRFs calculated from the onsets we searched for the N1 peak between 80-150 ms, for the P2 peak between 150-250 ms, and for the N2 peak between 200-300 ms. The TRF calculated from the envelope showed earlier peaks, therefore we searched for the N1 peak between 50-120 ms, for the P2 peak between 120-220 ms, and for the N2 peak between 220−320 ms.

Second, we determined the channel with the largest amplitude for the N1/P2/N2 peak. Around the peak, we then calculated a time-window of ±25 ms for the N1 and ±50 ms for the P2/N2 peak.

Third, to obtain the corresponding GED filter weights we contrasted the ERP and TRF time-windows with the baseline period. Specifically, we mean-centered the data of the time-windows and baseline period and computed the covariance matrices for either time periods. For each participant, the covariance matrix of interest (S) was computed for the time-windows and the reference covariance matrix (R) was computed for the baseline period.

We cleaned the covariance matrix S and R by first computing the average covariance matrix $\bar{S}/\bar{R}$ across participants. We then computed for each participant the Euclidean distance between the covariance S and $\bar{S}$, and R and $\bar{R}$. Covariance matrices which deviated from $\bar{S}$ and $\bar{R}$ with more than 3 standard deviations were removed and $\bar{S}$ and $\bar{R}$ were computed again across the remaining covariance matrices. The $\bar{R}$ matrix was then regularized using a shrinkage regularization parameter of $\lambda = 0.01$.

The resulting spatial filter maximizes the contrast between the time-windows and baseline period. For each ERP and TRF time-window, this resulted in a number of eigenvectors that is equal to the number of channels. The eigenvector with the largest eigenvalue should best separate the baseline and time-window. However, this does not guarantee physiological plausibility. We therefore investigated the GED components and reported when not using the GED component with the strongest eigenvalue. As a result we received a GED component time-series for each participant, ERP/TRF time-window, and block, and for the TRF additionally for each acoustic feature. Additionally, we computed a forward model for each component to investigate the physiological interpretability of the filter.

As eigenvectors are sign uncertain, we set appropriate signs for the GED components (negative for N1, N2, and positive for P2). Lastly, we averaged the amplitude of the GED component across the ERP and TRF time-windows to extract one amplitude value per time-window, participant, and block and for the TRF also per acoustic feature.

Statistical analysis

All statistical analyses were performed in RStudio (v. 2021.09.0).

Prediction values of the acoustic feature were compared using a Wilcoxon signed rank test.

We computed regression models for the subjective workload questions, behavioral responses, and GED component mean amplitudes for the ERP and TRF time-windows (i.e., N1/P2/N2). For these analyses, we started with a model including 'participant' as a random intercept and 'condition' as a fixed effect. 'Condition' contained two categories, i.e., low- and high-demand which were coded 0 and 1, respectively. We explored time-on-task effects (i.e., they were not part of the preregistration) by adding the block number as a continuous predictor 'time' in a second model. In the third and most complex model an interaction term between condition and time was added:

$$\hat{y} \sim condition + (1|participant) \tag{1}$$

$$\hat{y} \sim condition + time + (1|participant) \tag{2}$$

$$\hat{y} \sim condition * time + (1|participant) \tag{3}$$

To test whether the models improved by adding predictors we used likelihood ratio testing. If the model fit did not improve by adding the predictor 'time', results from the first model were reported.

A linear mixed model (LMM) was estimated for the ERP and TRF amplitudes, for each subjective workload question, and for the time to complete the surgery task. As the number of mistakes during the surgery task

---

(zero to three mistakes) and the number of times the tissue was damaged provided count data, a generalized linear model (GLM) with a poisson distribution was estimated. The LMM and GLM were estimated using the R package lmer4 (v. 1.1-30). Fixed effects were evaluated using Satterthwaite approximations within the R package lmerTest, which estimates the degrees of freedom to calculate p-values.

For the memory task, memory scores were first calculated using edit distance scoring[56] (i.e., scores range from 0 to 1). Then, the scores received ordinal values which were fit using a cumulative link mixed model using the clmm function of the ordinal package (v. 12-4).

Evidence for an effect was assumed for α =.05. We corrected for multiple comparisons using α =.05/3 = 0.017 for the ERP/TRF amplitudes (i.e., N1, P2, N2), the comparison between prediction values of the acoustic features (i.e., envelope, noise-reduced envelope, onsets), subjective workload questions (i.e., effort, frustration and distraction), and surgery task performance (i.e., duration, mistakes, and tissue damage).

Results
For each response we evaluated a condition difference using regression models. We explored whether adding time as a predictor improved the model fit. Table 1 shows the selected model for each response and whether the corresponding beta values were significant.

Subjective measures of demand
After each block, we asked participants how effortful, frustrating, and distracting they perceived the task. There was a significant increase in scores from the low- to high-demand condition for all three measures (Fig. 2a−c; effort: $b = 3.73$, $SE = 0.22$, $p < .001$; frustration: $b = 2.71$, $SE = 0.28$, $p < .001$; distraction: $b = 1.31$, $SE = 0.19$, $p < .001$). There was no significant change in theses measures over time, i.e., for all three questions the model fit did not improve by adding time as a predictor (supplementary table S2).

Objective measures of performance
The memory score, which could range between 0 to 1, was in the low-demand condition on average 0.97 (SD=0.13) and in the high-demand condition 0.75 (SD=0.24). The first model, which included only condition as a predictor, was used (supplementary table S2) and revealed a significant decrease in memory score from the low- to high-demand condition (Fig. 2d, $b = -4.63$, $SE = 0.39$, $p < .001$).

To evaluate the performance during the surgery task we used three measures: task duration (i.e., time to finish the surgery task), the number of mistakes (i.e., dropping or rupturing a vessel), and the number of times the tissue was damaged (i.e., tissue that was touched with either of the tools). Figure 2e−g illustrates the results of the behavioral performance. The performance was already high at the beginning of the experiment, which is evident in the low number of mistakes in most of the blocks. The average number of mistakes is 0.33 (SD=0.48) in the first block and 0.18 (SD=0.39) in the last block. The number of times tissue was damaged changed from 6.2 (SD=3.96) in the first block to 5.4 (SD=4.88) in the last block. Furthermore, task duration appeared to improve, particularly in the first ten blocks, after which it remained relatively stable. For neither of these measures, we found a significant difference between conditions (task duration: $b = -1.842$, $SE = 1.57$, $p = .228$; mistakes: $b = 0.14$, $SE = 0.19$, $p = .445$; tissue damage: $b = -0.02$, $SE = 0.034$, $p = .556$). However, for all measures adding time as a predictor significantly improved the model fit (supplementary table S2) and revealed a significant

	Response	Model	Condition	Time
Subjective workload questions	tlx-effort	1	3.73*	
	tlx-frustration	1	2.71*	
	tlx-distraction	1	1.31*	
Behavioral performance	Memory score	1	−4.63*	−0.91*
	Duration	2	−1.842	−0.03*
	Mistakes	2	0.14	−0.01*
	Tissue damage	2	−0.02	
ERPs	N1	2	0.4	0.3*
	P2	2	0.21	0.38*
	N2	1	0.169	
TRF$_{envelope}$	N1	1	−0.94	
	P2	1	−0.37	
	N2	1	0.75	
TRF$_{onsets}$	N1	2	0.49	0.07 .
	P2	1	0.025	
	N2	1	−0.12	

Table 1. The table shows the selected model for each subjective, behavioral, and neural response. 'Model' lists the chosen model for each response. Model 1 included 'condition' as a predictor and model 2 'condition' and 'time' as predictors. Beta values are listed below each predictor. The stars indicate predictors below the Bonferroni-corrected *p*-value. A trend is marked with a dot.
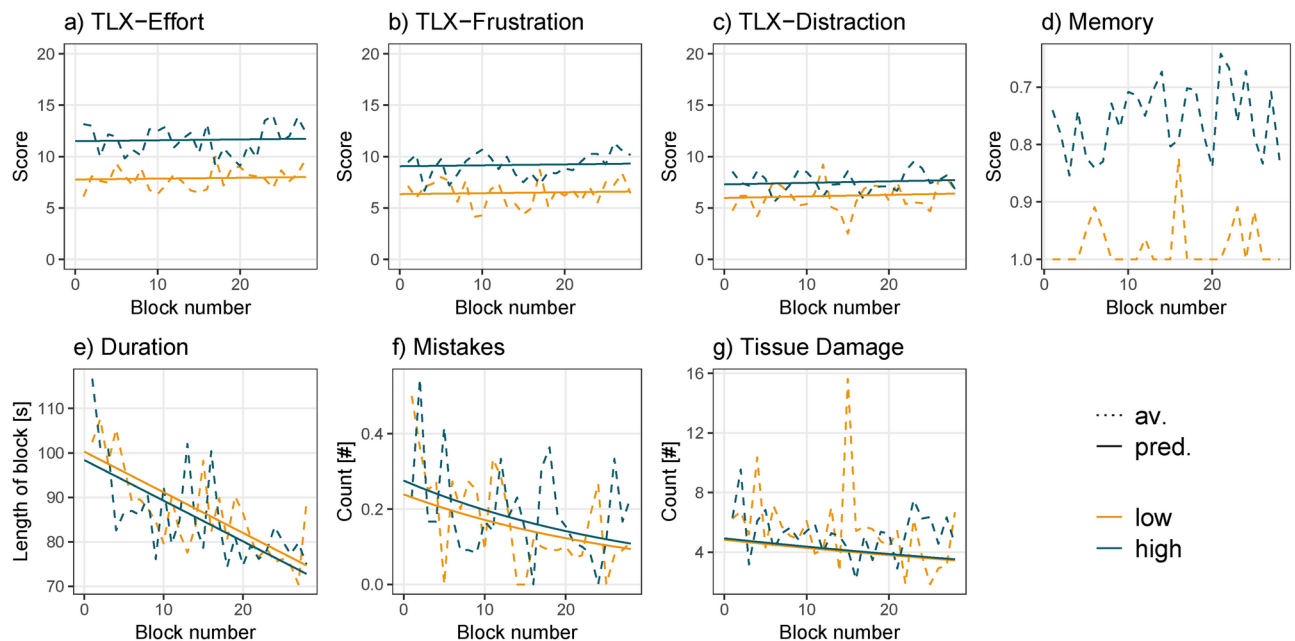
Figure 2. (**a–c**) Subjective responses. (**d**) Memory task performance. (**e–f**) Surgery task performance. Dotted lines show averaged data over participants for each condition and block. Solid lines represent predicted responses that were calculated using the second model. Predicted responses for the memory task were not computed as a cumulative link model was used.

decrease over time (task duration: $b = -0.91, SE = 0.1, p < .001$; mistakes: $b = -0.03, SE = 0.01, p = .0046$; tissue damage: $b = -0.01, SE = 0.02, p < .001$), in other words, participants became faster and made fewer mistakes.

Demand and time-on-task effects for ERPs in response to the spoken letters

The response to the task-irrelevant spoken letters was evaluated by averaging GED component amplitudes for the N1 (Fig. 4a), P2 (Fig. 4b), and N2 (Supplementary Figure S4a) over their respective time-window. Both the morphology as well as the topographies of the GED components with the strongest eigenvalues (supplementary Figure S3a) are physiologically plausible.

Contrary to our expectation, we observed no significant effect of demand on the N1 ($b = 0.4, SE = 1.84, p = .827$), the P2 ($b = 0.21, SE = 1.91, p = .256$), or the N2 ($b = 0.169; SE = 2.05; p = .411$) amplitude. Exploring time as a predictor significantly improved the models for the N1 and P2 amplitudes but not for the N2 amplitude (supplementary table S3) and revealed a significant N1 amplitude decrease ($b = 0.3, SE = 0.11, p = .0012$) and a P2 amplitude increase ($b = 0.38, SE = 0.11, p = .0014$) over time.

Prediction accuracies for the TRFs in response to the OR playback

The TRFs were calculated by relating the continuous OR playback to the EEG signal. We used three acoustic features to calculate the TRF, namely the envelope extracted from the raw audio, the envelope extracted from the noise-reduced audio, and the onsets extracted from the raw audio (Fig. 3a). To compare these features, we calculated the prediction accuracy for unseen neural data. Overall, the prediction values were small (Fig. 3b) which is common for this measure. Importantly, the envelope of the noise-reduced audio as well as the onsets showed significantly higher prediction accuracies compared to the envelope from the raw audio (noise-reduced envelope: $W = 3, p < .001$; onsets: $W = 1, p < .001$). The noise-reduced envelope and onsets did not significantly differ from each other ($W = 118, p = .799$).

Demand and time-on-task effects for TRFs computed from the noise-reduced envelope and onsets.

We computed the TRF in response to the noise-reduced envelope (hereafter $TRF_{envelope}$) and in response to the onsets (hereafter TRF_{onsets}). Although there was no difference between prediction accuracies between these two acoustic features, they might still represent different aspects of the soundscape and, therefore, different responses. To enhance the signal-to-noise ratio and to spatially filter the data GED components were extracted. The mean amplitudes over the N1, P2, and N2 time-windows were used for evaluation. Figure 4c and d shows the distinct time-series and topographies for the N1 $TRF_{envelope}$ and the N1 TRF_{onsets}, respectively. The remaining GED time-series, as well as GED eigenvalues can be found in supplementary figure S3 and S4. Visually there are considerable differences in the temporal evolution between the $TRF_{envelope}$ and the TRF_{onsets}. While the $TRF_{envelope}$ follows a trajectory similar to the ERPs its response peaks occur earlier than that of the ERPs and the TRF_{onsets}.
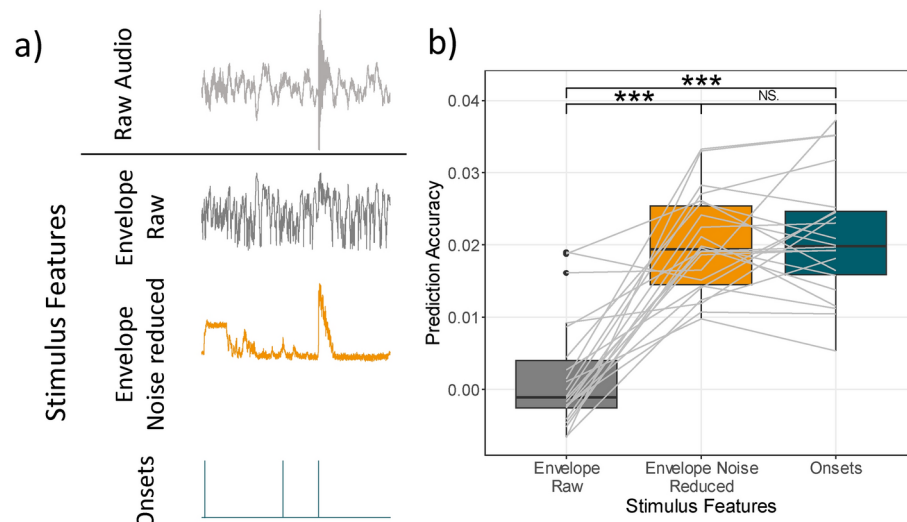
Figure 3. (**a**) The acoustic features used for TRF model estimation. The same stimulus snippet is shown as the raw audio, envelope of the raw audio, envelope of the noise-reduced audio, and onsets of the raw audio. (**b**) Prediction values of each acoustic feature. Each line represents the change in prediction value between acoustic features for each participant. *** $p < .001$.
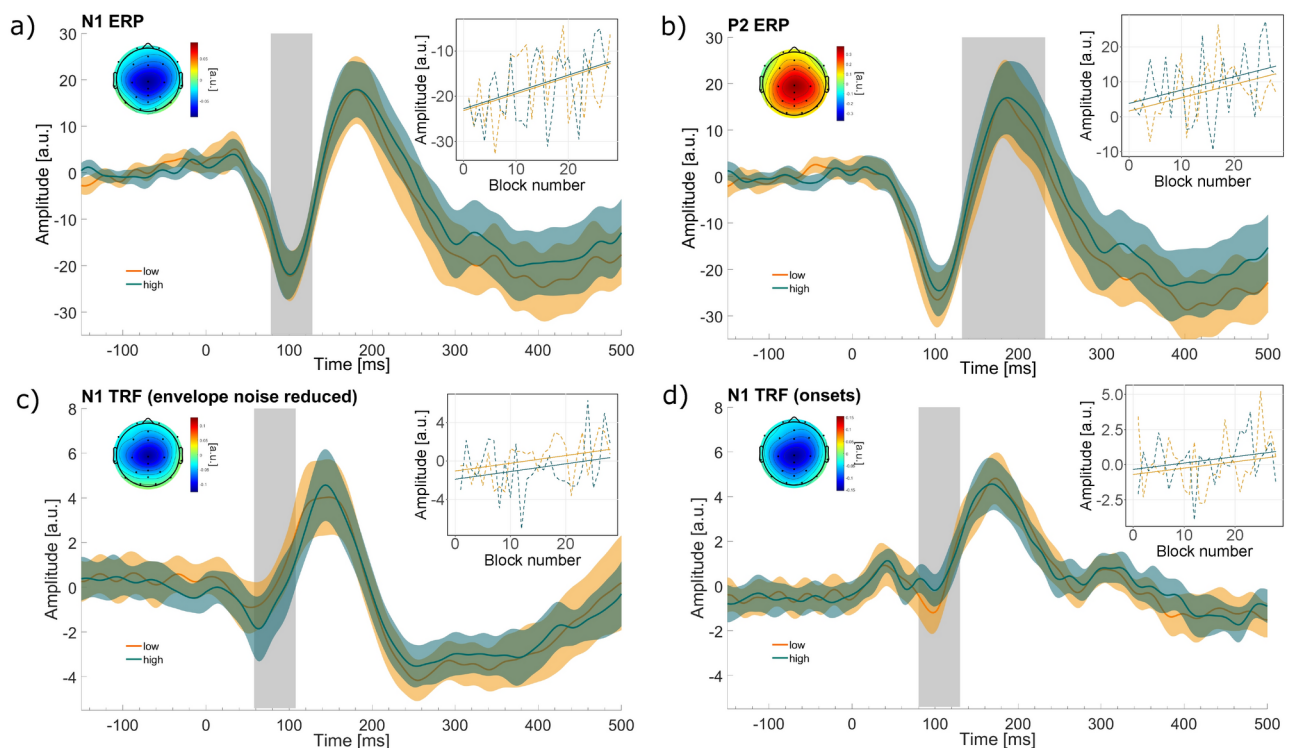

Figure 4. The first GED components of the (**a**) N1 ERP (demand: n.s., time: $p = .0012$), (**b**) P2 ERP (demand: n.s., time: $p = .0014$), (**c**) N1 TRF$_{envelope}$ (demand: n.s., time: n.s.), and (**d**) N1 TRF$_{onsets}$ (demand: n.s., time: n.s.). The morphology shows the GED time-series for each condition, averaged over participants. The shaded area represents the confidence interval across participants. The grey bar shows the time-window that was used to derive the amplitudes. The topographies depict the forward model of the GED weights and shows which channels contributed most to the GED component. The graph to the right, shows the change in amplitude across time and conditions. Dotted lines show averaged data over participants for each condition and block. Solid lines represent predicted responses that were calculated using the second model.

Regarding the N2 TRF$_{envelope}$, the GED component with the second strongest eigenvalue showed more plausible trajectories and topographies than the first GED component (supplementary figure S4). As TRF$_{envelope}$ and TRF$_{onset}$ showed similar prediction values, we evaluated amplitudes of both TRFs, and corrected for additional multiple comparisons using

α =.05/3 (responses) / 2 (acoustic feature) = 0.0083. For all TRF$_{envelope}$ amplitudes, we found no condition difference (N1: $b = -0.94, SE = 0.57, p = .102$; P2: $b = -0.37, SE = 0.49, p = .449$; N2: $b = 0.75, SE = 0.45, p = .095$) and adding time as a predictor did not lead to a better model fit compared to the first model (supplementary table S3).

Regarding the TRF$_{onsets}$ amplitudes we found no condition differences (N1: $b = 0.49, SE = 0.41, p = .23$; P2: $b = 0.025, SE = 0.39, p = .948$; P2: $b = -0.12, SE = 0.35, p = .716$). Adding time as a predictor yielded a better model fit for the N1 amplitude, but not for the other amplitudes. This revealed a trend towards significance for the predictor time for the N1 TRF$_{onsets}$ ($b = 0.07, SE = 0.02, p = .011$).

Discussion

In this study, we examined the impact of varying demand and time-on-task on subjective responses, task performance, and EEG responses within a naturalistic setting, utilizing a laparoscopic simulator and an operating room soundscape. We found divergent subjective and objective behavioral responses. Participants perceived the high-demand condition as more demanding than the low-demand condition, which remained stable over time. However, surgery task performance did not differ between demand conditions but demonstrated improvement over time.

The overall perceived demand of the two conditions was evaluated with subjective workload questions. The observed effect that high demand compared to low demand increases subjective workload ratings during a surgery task aligns with prior research[57–59]. A possible explanation for the perceived differences in demand across conditions, despite consistent performance levels across conditions, is that participants may have prioritized their performance of the surgery task. By putting more cognitive effort into the surgery task in the high-demand condition, they were able to achieve similar surgery task performance across both conditions. Likewise, a surgeon would prioritize performance during an actual surgery for the benefit of the patient, even if it means an increase in perceived demand. This implies that task performance and subjective experience are not necessarily related and may also explain the contradictory findings reported in the literature. Some studies report a negative impact of increased demand on performance[20,57], while others show no effect[60], or suggest that the effect depends on the investigated performance measure[58,61]. Most studies, did not assess the subjective demand, and only concentrate on task performance. However, this limits our understanding how perceived demand and surgical task performance relate and should be addressed in future research by incorporating subjective demand measures.

Regarding the impact of time-on-task on performance, previous research has also demonstrated that task performance improves over time[62–65], highlighting visible training effects within a short amount of practice when using the same surgery task. We did not find a change in perceived demand over time, similar to Suarez et al. (2022)[62] who found that training did not reduce demand for the same NASA-TLX items as used in this study (i.e., effort and frustration) or the total score. Others reported that surgical training lowers the NASA-TLX total score[63,65]. This once again demonstrates that the relationship between task performance and perceived demand is not straightforward. The use of various methods to vary demand and surgical tasks, each with different performance parameters and subjective and objective measures, also complicates the generalization of findings across studies[66]. To improve generalization, it would be useful to evaluate how the demand experienced during actual surgeries reflects the demand experienced during surgical simulations. Our results suggest that situational demand, rather than actual surgical skill, influences the perceived demand in inexperienced participants. This suggests that surgical training programs should integrate high-demand simulations to more effectively prepare novice surgeons, who are particularly susceptible to changes in task demand[4,60], for the cognitive and emotional demand they will encounter in real surgeries. By systematically exposing novice surgeons to challenging scenarios, training programs can enhance their ability to manage stress and maintain performance under pressure, potentially reducing errors and improving patient outcomes. Furthermore, novice surgeons may benefit from training to reflect on their individual susceptibility to noise. This is because noise reduction interventions are especially beneficial for surgeons who perceive the soundscape as disturbing[7]. Such training could also incorporate communication strategies to reduce avoidable noise sources in the operating room, especially at the start of critical and high-demand phases of the surgery. This would help to reduce the overall experienced demand for the surgeons.

As an alternative objective measure to capture the perceived task demand, we investigated two neural response measures to the task-irrelevant soundscape. Specifically, we analyzed ERPs in response to spoken letters and TRFs in response to the OR playback. In general, we find a clear neural response to the letters as well as for the naturalistic OR playback, which is in line with our previous study showing significant responses for naturalistic soundscapes[26]. This further expands the use of mobile EEG in applied settings[27], as both measures can be used to study the neural response to a complex natural soundscape in a work-like environment. Contrary to our hypothesis, we found no significant effect of task demand on the neural measure of sound processing, neither in response to the letters nor to the OR playback. Participants' subjective ratings indicate that our intended manipulation - to increase the scenario's challenge - had an effect, with participants perceiving the soundscape as more distracting in the high-demand condition. However, this perceptual shift is not reflected in our neural measurements. Our exploratory analyses revealed time-on-task effects, for the N1 and P2, and a trend for the

TRF$_{onsets}$ N1. This was apparently unrelated to the subjectively perceived distraction, which remained stable over time.

Our results should be considered in light of the neural measures used to investigate different aspects of the soundscape, i.e., the ERPs and TRFs. The ERPs captured responses to the regularly occurring letters associated with the memory task. To increase the distracting nature of the task-irrelevant soundscape and simulate demanding internal cognitive processes we adapted a serial-recall paradigm. We observed no effect on any ERP time-window, however, previous studies using a serial-recall paradigm have shown that an increase in distraction leads to an increase in N1 ERP amplitude[67]. In contrast to a classic serial-recall paradigm, we used the serial-recall task as a secondary task with a long retention interval (more than 60 seconds), and presented distractors at a rate of 3 seconds. Furthermore, we presented the letters in the presence of background noise, which may have reduced distraction effects in our paradigm[68]. Our choice of distractor presentation (long inter-trial-intervals and in the presence of noise), in combination with an engaging surgery task, may have enabled participants to easily ignore the letters over time, independent of the number of items to recall, resulting in similar responses across conditions. Notably, the results showed a decrease in N1 amplitude and an increase in P2 amplitude, indicating neural adaptation due to repeated stimulus presentation. This may be similar to the adaptation observed for brief repetitive tones[69,70]. The decrease in ERP N1 amplitude could indicate that distraction due to the letters decreased[67], independent of the conditions. Participants may have initially attended to the letters, e.g., due to their deviation from the continuous soundscape, and learned that the letters are task-irrelevant and can be ignored as such[71]. The P2 amplitude increase is more difficult to interpret, as only few studies investigated it in the context of distraction. One study showed an amplitude increase with increased distraction[72], which would contradict our interpretation, while another study found no effect of distraction on P2 amplitude[73]. In summary, the results suggest that the task-irrelevant spoken letters did not sufficiently distract participants to elicit distinct neural responses under varying demands. However, the changes in amplitude across time indicate that the responses can capture some form of modulation. As this was an exploratory finding, further investigation is necessary to reveal how responses to task-irrelevant acoustic events change over time when presented in a realistic work-like setting. This information could shed light on how the brain adapts to such events in real life.

The TRFs captured responses to the realistic OR playback. To the best of our knowledge this is the first study investigating demand effects on responses to task-irrelevant ongoing soundscapes. Despite the fact that we measured a clear TRF to the soundscape, we did not find a difference between conditions. One explanation for this null-finding is that TRFs calculated from the entire soundscape are not sensitive enough to detect demand effects. A small portion of sounds may have elicited demand effects similar to ERPs[30]. However, these effects may have been reduced by mostly demand-unrelated responses. Support for this explanation comes from our previous study[26], where we found attention-dependent responses for specific highly salient sounds, but only marginal differences in the TRF to the background soundscape. This might also explain why we found no time-on-task effects for the TRF, except for the trend of a decrease for the TRF N1 calculated from the onsets (TRF$_{onsets}$), despite the effect for the ERPs. While the ERPs captured responses to similar sounds, the TRFs captured responses to the OR playback which was much more diverse than the letters. This might have decreased a potential time-on-task effect for the TRFs compared to ERPs, as different and overlapping responses are less sensitive to detect effects compared to similar and isolated responses[37]. For example, an audio signal, that contains a constant high level of noise, would result in an essentially flat envelope, for which no reliable TRF could be computed. Any acoustic event, that is embedded in this background sound, though perceivable to the human ear, would not be reflected in the envelope. This is evident in our finding that TRFs computed from the raw envelope, which captured little variation of the soundscape, provided low prediction values. The individual acoustic events, such as the beeps of machinery or the clatter of tools, were only adequately captured by reducing the noise in the raw envelope. By highlighting the individual acoustic events, the envelope became more similar to the computation of the onsets. The similarity between these two features is expressed in similar prediction values. This suggests that the brain's processing of sounds goes beyond simply following the acoustic envelope[36,74]. Instead, it actively distinguishes and monitors individual acoustic events in the presence of background noise, similar to how speech is perceived in noisy environments[75]. Note, that similar prediction values, do not necessarily imply a similar neural representation of the soundscape[37]. This is hinted by the trend in the TRF$_{onsets}$ N1 which was not observed in the TRF$_{envelope}$. However, the similar prediction accuracies indicate that onset information about acoustic events is enough to estimate the neural response to complex soundscapes. This can have practical implications when privacy is important (e.g., in an actual OR context where medical information about a person are discussed). It would be sufficient to extract acoustic onsets automatically, without the need to record the raw audio[35,76]. Overall, we have demonstrated that it is possible to obtain responses from continuous task-irrelevant soundscapes. However, future studies may benefit from describing the soundscape not only in the form of an envelope or by extracting onsets, but to differentiate between different sound sources to obtain a more fine grained picture.

An alternative explanation for the discrepancy between our objective measures (i.e., task performance and neural measures) and the subjective reports is that the subjective reports did not accurately reflect participants' perceived demand throughout an entire block. As the subjective responses were collected at the end of a block, they may reflect the perceived demand of the memory task during encoding. In other words, participants found it difficult to remember two compared to eight letters, but experienced the surgery task as similar demanding across conditions. It is important to note that although performance on the memory task differed between conditions, the difference was generally small. This may be attributed to the homogeneity of the participant pool and their high cognitive capacities, comprising medical students of a similar young age. Although participants still rated the high-demand condition to be more demanding, the high-demand condition has not depleted their cognitive resources enough, which could explain the absence of differences in our objective measures. Furthermore, given the complexity of the task, accurately pinpointing the source and amount of demand

- whether it's remembering the letters, the surgery task, or a combination of both - may have been difficult for participants. Indeed, the NASA-TLX has been criticized that different participants might rate different aspects of the task[77]. Collecting separate subjective responses for each task may have improved interpretation. However, in natural environments, the source of demand, whether it is the environment or the task, can rapidly change, making subjective responses at a single point in time difficult to interpret. This once again emphasizes the importance of continuously and objectively monitoring sources of demands, such as the soundscape, to understand how they affect the individual.

In order to improve the understanding of varying demands on responses to ongoing, task-irrelevant soundscapes, we offer three suggestions for future work. Firstly, a different manipulation of demand could clarify whether neural responses to irrelevant sounds are indeed insensitive to detect demand effects in complex settings. The measurement of EEG during a standing, bi-manual task presents a greater challenge than in traditional stationary setups. Such complex scenarios also engage a variety of cognitive processes. These two aspects likely increase individual variability in EEG responses and reduce sensitivity to isolated demand effects. Furthermore, it is possible that the demand manipulation may have not been strong enough to elicit differences in the neural response. For these reasons, it might be necessary to implement manipulations with larger differences in cognitive demand. Secondly, we used a general representation of the soundscape, by computing the envelope and onsets of the playback. A more detailed description of the soundscape, for example by differentiating between specific sounds may lead to a better estimate of the neural response[37]. Finally, our study involved a small sample of young medical students with limited OR experience. This offers valuable insights into the feasibility of using EEG in complex settings. However, the generalizability to the clinical personal is limited. Future studies would benefit from increasing heterogeneity and sample size. Tailoring the sample to the target population by including experienced surgeons would provide deeper understanding of how personnel accustomed to complex environments process and perceive the soundscape under varying demand. Implementing these suggestions will advance the application of EEG in work environments and enhance our understanding of whether and how cognitive states can be inferred from responses to natural, task-irrelevant soundscapes.

Our ultimate goal is to study neural responses in real-world settings and workplaces, such as the OR. These settings are characterized by a complex and varying soundscape, while individuals engage in tasks of varying complexity, which the researchers have limited control over. In order to learn from our results for other studies that are interested in real-world recordings, it is important to discuss some properties of the soundscape we have used here. We used the playback of a natural soundscape that was recorded in an OR using a static microphone positioned at the center of the room. We presented this soundscape via two loudspeakers to a stationary participant. This setup ensured that the soundscape maintained realistic spatial properties from the listener's perspective, reflecting the acoustic environment of an operating room. Despite the naturalness of this soundscape, there are important differences between the OR playback we used and an actual OR soundscape experienced in real-life situations. One key difference is sound expectancy. For instance, in an OR setting, there is an expectation that specific actions one can see, such as laying down a tool, will produce a particular sound. In other words, there is a congruence between visual and auditory information. In our study, this audio-visual congruence is absent; all the sounds our participants hear are not congruent with the actual situation they are in. One might expect these differences to affect the neural response we have measured, although the direction of this effect is difficult to predict. Some speech tracking studies have shown an enhanced tracking of the speech envelope when congruent visual information is provided[78]. On the other hand, a mismatch of visual and auditory information could also lead to a larger neural response[79]. A second difference between the soundscape experienced in the OR and our OR playback is the relevance of the sounds for the participant. In an OR setting, many sounds carry contextual information, for example the sound of an alarm, the feedback sound of a tool that is in use, or relevant communication between personnel, that are critical to the surgery and sometimes require task-related actions. In contrast, the soundscape we presented was entirely task-irrelevant for our participants, essentially transforming all sounds into a meaningless noise background. One can expect that for personally relevant or salient information within a soundscape influences its processing and thereby the neural response[25,80,81]. Taken together, these factors may significantly change the neural response to the naturalistic soundscape we have presented, compared to when this naturalistic soundscape would have been experienced in its actual context. Moving forward, future studies should be aware of these important factors when studying naturalistic soundscapes.

To conclude, our results demonstrated the feasibility and potential of combining mobile EEG and audio recordings to study cognitive processes in complex, real-world environments. Our findings showed a temporal change in neural responses to specific sounds, but not to the entire soundscape. This suggests that responses to isolated sounds may be more sensitive indicators of alterations in auditory processing than responses to the entire soundscape. Furthermore, we observed that increased task demand resulted in a higher perceived demand and distraction, but these changes were not reflected in the neural responses or task performance measures. This discrepancy indicates that a high performance can be maintained even in demanding environments, but potentially at a cost to a person's wellbeing. These insights have important implication for the surgical field and other workplaces. They highlight the need to consider both subjective experiences and objective measures when evaluating cognitive demand in high-stakes environments. Our approach contributed to methodological advancements in neurophysiological research and opens new avenues for optimizing auditory environments in surgical workplaces, ultimately contributing to both performance and the wellbeing of personnel.

Data availability
The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

References

1. Etheridge, J. C. *et al.* Trends in Surgeon Burnout in the US and Canada: Systematic Review and Meta-Regression Analysis. *Journal of the American College of Surgeons* **236**, 253. https://doi.org/10.1097/XCS.0000000000000402 (2023).
2. Marrelli, M., Gentile, S., Palmieri, F., Paduano, F. & Tatullo, M. Correlation between Surgeon's Experience, Surgery Complexity and the Alteration of Stress Related Physiological Parameters. *PLOS ONE* **9**, e112444. https://doi.org/10.1371/journal.pone.0112444 (2014).
3. Rieger, A., Stoll, R., Kreuzfeld, S., Behrens, K. & Weippert, M. Heart rate and heart rate variability as indirect markers of surgeons' intraoperative stress. *International Archives of Occupational and Environmental Health* **87**, 165–174. https://doi.org/10.1007/s00420-013-0847-z (2014).
4. Arora, S. *et al.* The impact of stress on surgical performance: A systematic review of the literature. *Surgery* **147**, 318-330.e6. https://doi.org/10.1016/j.surg.2009.10.007 (2010).
5. Kern, M. *et al.* Psychological stressors, resources and well-being of surgeons in Germany: A cross-sectional study. *Chirurg* **90**, 576–584. https://doi.org/10.1007/s00104-018-0780-5 (2019).
6. Baltin, C. T. *et al.* Noise patterns in visceral surgical procedures: Analysis of second-by-second dBA data of 599 procedures over the course of one year. *Scientific Reports* **10**, 1–10. https://doi.org/10.1038/s41598-020-59816-4 (2020).
7. Engelmann, C. R., Neis, J. P., Kirschbaum, C., Grote, G. & Ure, B. M. A noise-reduction program in a pediatric operation theatre is associated with surgeon's benefits and a reduced rate of complications: A prospective controlled clinical trial. *Annals of Surgery* **259**, 1025–1033. https://doi.org/10.1097/SLA.0000000000000253 (2014).
8. Padmakumar, A. D. *et al.* Effect of noise on tasks in operating theatres: A survey of the perceptions of healthcare staff. *British Journal of Oral and Maxillofacial Surgery* **55**, 164–167. https://doi.org/10.1016/j.bjoms.2016.10.011 (2017).
9. Keller, S. *et al.* Noise in the Operating Room Distracts Members of the Surgical Team. An Observational Study. *World Journal of Surgery* **42**, 3880–3887. https://doi.org/10.1007/s00268-018-4730-7 (2018).
10. Healey, A. N., Primus, C. P. & Koutantji, M. Quantifying distraction and interruption in urological surgery. *Quality and Safety in Health Care* **16**, 135–139. https://doi.org/10.1136/qshc.2006.019711 (2007).
11. Tsiou, C., Efthymiatos, G. & Katostaras, T. Noise in the operating rooms of Greek hospitals. *The Journal of the Acoustical Society of America* **123**, 757–765. https://doi.org/10.1121/1.2821972 (2008).
12. van Harten, A., Gooszen, H. G., Koksma, J. J., Niessen, T. J. & Abma, T. A. An observational study of distractions in the operating theatre. *Anaesthesia* **76**, 346–356. https://doi.org/10.1111/anae.15217 (2021).
13. Gülşen, M., Aydıngülü, N. & Arslan, S. Physiological and psychological effects of ambient noise in operating room on medical staff. *ANZ Journal of Surgery* **91**, 847–853. https://doi.org/10.1111/ans.16582 (2021).
14. Conrad, C. *et al.* The effect of defined auditory conditions versus mental loading on the laparoscopic motor skill performance of experts. *Surgical Endoscopy* **24**, 1347–1352. https://doi.org/10.1007/s00464-009-0772-0 (2010).
15. Conrad, C. *et al.* A quality improvement study on avoidable stressors and countermeasures affecting surgical motor performance and learning. *Annals of Surgery* **255**, 1190–1194. https://doi.org/10.1097/SLA.0b013e318250b332 (2012).
16. Hodge, B. & Thompson, J. F. Noise pollution in the operating theatre. *The Lancet* **335**, 891–894. https://doi.org/10.1016/0140-6736(90)90486-O (1990).
17. Moorthy, K., Munz, Y., Sarker, S. K. & Darzi, A. Objective assessment of technical skills in surgery. *British Medical Journal* **327**, 1032–1037. https://doi.org/10.1136/bmj.327.7422.1032 (2003).
18. Moorthy, K., Munz, Y., Undre, S. & Darzi, A. Objective evaluation of the effect of noise on the performance of a complex laparoscopic task. *Surgery* **136**, 25–30. https://doi.org/10.1016/j.surg.2003.12.011 (2004).
19. Suh, I. H. *et al.* The negative effect of distraction on performance of robot-assisted surgical skills in medical students and residents. *The International Journal of Medical Robotics and Computer Assisted Surgery* **6**, 377–381. https://doi.org/10.1002/rcs.338 (2010).
20. Moorthy, K., Munz, Y., Dosis, A., Bann, S. & Darzi, A. The effect of stress-inducing conditions on the performance of a laparoscopic task. *Surgical Endoscopy and Other Interventional Techniques* **17**, 1481–1484. https://doi.org/10.1007/s00464-002-9224-9 (2003).
21. Han, Y. *et al.* Impact of background music on the performance of laparoscopy teams. *BMC Medical Education* **22**, 439. https://doi.org/10.1186/s12909-022-03503-7 (2022).
22. Polich, J. Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology* **118**, 2128–2148. https://doi.org/10.1016/j.clinph.2007.04.019 (2007).
23. Crosse, M. J., Di Liberto, G. M., Bednar, A. & Lalor, E. C. The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience* **10**, 1–14. https://doi.org/10.3389/fnhum.2016.00604 (2016).
24. Debener, S., Minow, F., Emkes, R., Gandras, K. & de Vos, M. How about taking a low-cost, small, and wireless EEG for a walk?. *Psychophysiology* **49**, 1617–1621. https://doi.org/10.1111/j.1469-8986.2012.01471.x (2012).
25. Straetmans, L., Holtze, B., Debener, S., Jaeger, M. & Mirkovic, B. Neural tracking to go: Auditory attention decoding and saliency detection with mobile EEG. *Journal of Neural Engineering* **18**, 066054. https://doi.org/10.1088/1741-2552/ac42b5 (2021).
26. Rosenkranz, M., Cetin, T., Uslar, V. N. & Bleichner, M. G. Investigating the attentional focus to workplace-related soundscapes in a complex audio-visual-motor task using EEG. *Frontiers in Neuroergonomics* **3**. https://doi.org/10.3389/fnrgo.2022.1062227 (2023).
27. Wascher, E. *et al.* Neuroergonomics on the Go: An Evaluation of the Potential of Mobile EEG for Workplace Assessment and Design. *Human Factors* **00**, 1–21. https://doi.org/10.1177/00187208211007707 (2021).
28. Hölle, D., Meekes, J. & Bleichner, M. G. Mobile ear-EEG to study auditory attention in everyday life: Auditory attention in everyday life. *Behavior Research Methods* **53**, 2025–2036. https://doi.org/10.3758/s13428-021-01538-0 (2021).
29. Lavie, N. Distracted and confused?: Selective attention under load. *Trends in Cognitive Sciences* **9**, 75–82. https://doi.org/10.1016/j.tics.2004.12.004 (2005).
30. Brockhoff, L., Schindler, S., Bruchmann, M. & Straube, T. Effects of perceptual and working memory load on brain responses to task-irrelevant stimuli: Review and implications for future research. *Neuroscience & Biobehavioral Reviews* **135**, 104580. https://doi.org/10.1016/j.neubiorev.2022.104580 (2022).
31. Sörqvist, P., Dahlström, Ö., Karlsson, T. & Rönnberg, J. Concentration: The Neural Underpinnings of How Cognitive Load Shields Against Distraction. *Frontiers in Human Neuroscience* **10** (2016).
32. Kong, Y.-Y., Mullangi, A. & Ding, N. Differential modulation of auditory responses to attended and unattended speech in different listening conditions. *Hearing Research* **316**, 73–81. https://doi.org/10.1016/j.heares.2014.07.009 (2014).
33. Horton, C., D'Zmura, M. & Srinivasan, R. Suppression of competing speech through entrainment of cortical oscillations. *Journal of Neurophysiology* **109**, 3082–3093. https://doi.org/10.1152/jn.01026.2012 (2013).
34. Effects of attention. Picton, T. W. & Hillyard, S. A. Human auditory evoked potentials. II. *Electroencephalography and Clinical Neurophysiology* **36**, 191–200. https://doi.org/10.1016/0013-4694(74)90156-4 (1974).
35. Hölle, D., Blum, S., Kissner, S., Debener, S. & Bleichner, M. G. Real-time audio processing of real-life soundscapes for EEG analysis: ERPs based on natural sound onsets. *Frontiers in Neuroergonomics* **3**, https://doi.org/10.3389/fnrgo.2022.793061 (2022).
36. Drennan, D. P. & Lalor, E. C. Cortical Tracking of Complex Sound Envelopes: Modeling the Changes in Response with Intensity. *eNeuro* **6**, https://doi.org/10.1523/ENEURO.0082-19.2019 (2019).

37. Haupt, T., Rosenkranz, M. & Bleichner, M. G. Exploring relevant Features for EEG-Based Investigation of Sound Perception in Naturalistic Soundscapes. https://doi.org/10.31234/osf.io/nuy7e (2024).
38. Schlickum, M. *et al.* Surgical Simulation Tasks Challenge Visual Working Memory and Visual-Spatial Ability Differently. *World Journal of Surgery* **35**, 710–715. https://doi.org/10.1007/s00268-011-0981-2 (2011).
39. Hedman, L., Klingberg, T., Enochsson, L., Kjellin, A. & Felländer-Tsai, L. Visual working memory influences the performance in virtual image-guided surgical intervention. *Surgical Endoscopy* **21**, 2044–2050. https://doi.org/10.1007/s00464-007-9287-8 (2007).
40. Colle, H. A. & Welsh, A. Acoustic masking in primary memory. *Journal of Verbal Learning and Verbal Behavior* **15**, 17–31. https://doi.org/10.1016/S0022-5371(76)90003-7 (1976).
41. Jones, D. & Morris, N. Irrelevant speech and serial recall: Implications for theories of attention and working memory. *Scandinavian Journal of Psychology* **33**, 212–229. https://doi.org/10.1111/j.1467-9450.1992.tb00911.x (1992).
42. Rennies, J. *et al.* Analyse und algorithmische Optimierung von Geräuschkulissen und Sprachkommunikation im OP- Saal. *In Fortschritte der Akustik - DAGA* **2023**, 646–649 (2023).
43. Kleiner, M. *et al.* What's new in psychtoolbox-3. *Perception* **36**, 1–16 (2007).
44. Hart, S. G. & Staveland, L. E. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology* **52**, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9 (1988).
45. Wilson, M. R. *et al.* Development and Validation of a Surgical Workload Measure: The Surgery Task Load Index (SURG-TLX). *World Journal of Surgery* **35**, 1961–1969. https://doi.org/10.1007/s00268-011-1141-4 (2011).
46. Kothe, C. *et al. The Lab Streaming Layer for Synchronized Multimodal Recording.* https://doi.org/10.1101/2024.02.13.580071 (2024).
47. Delorme, A. & Makeig, S. EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics. *Journal of Neuroscience Methods* **13**, 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009 (2004).
48. Mullen, T. R. *et al.* Real-time Neuroimaging and Cognitive Monitoring Using Wearable Dry EEG. *Physiology & behavior* **176**, 100–106. https://doi.org/10.1109/TBME.2015.2481482 (2015).
49. Klug, M. *et al. The BeMoBIL Pipeline for automated analyses of multimodal mobile brain and body imaging data.* https://doi.org/10.1101/2022.09.29.510051 (2022).
50. Pion-Tonachini, L., Kreutz-Delgado, K. & Makeig, S. ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage* **198**, 181–197. https://doi.org/10.1016/j.neuroimage.2019.05.026 (2019).
51. Klug, M. & Gramann, K. Identifying key factors for improving ICA-based decomposition of EEG data in mobile and stationary experiments. *European Journal of Neuroscience* **54**, 8406–8420. https://doi.org/10.1111/ejn.14992 (2021).
52. Crosse, M. J. *et al.* Linear Modeling of Neurophysiological Responses to Speech and Other Continuous Stimuli: Methodological Considerations for Applied Research. *Frontiers in Neuroscience* **15**, 705621. https://doi.org/10.3389/fnins.2021.705621 (2021).
53. Scalart, P. Wiener filter for noise reduction and speech enhancement. https://www.mathworks.com/matlabcentral/fileexchange/24462-wiener-filter-for-noise-reduction-and-speech-enhancement (2023). [Online; accessed April, 2023].
54. Plapous, C., Marro, C. & Scalart, P. Improved Signal-to-Noise Ratio Estimation for Speech Enhancement. *IEEE Transactions on Audio, Speech, and Language Processing* **14**, 2098–2108. https://doi.org/10.1109/TASL.2006.872621 (2006).
55. Cohen, M. X. A tutorial on generalized eigendecomposition for denoising, contrast enhancement, and dimension reduction in multichannel electrophysiology. *NeuroImage* **247**, 118809. https://doi.org/10.1016/j.neuroimage.2021.118809 (2022).
56. Gonthier, C. An easy way to improve scoring of memory span tasks: The edit distance, beyond "correct recall in the correct serial position". *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01908-2 (2022).
57. Gao, J. *et al.* Subjective and Objective Quantification of the Effect of Distraction on Physician's Workload and Performance During Simulated Laparoscopic Surgery. *Medical Science Monitor* **25**, 3127–3132. https://doi.org/10.12659/MSM.914635 (2019).
58. Modi, H. N., Singh, H., Darzi, A. & Leff, D. R. Multitasking and Time Pressure in the Operating Room: Impact on Surgeons' Brain Function. *Annals of Surgery* **272**, 648–657. https://doi.org/10.1097/SLA.0000000000004208 (2020).
59. Zander, T. O. *et al.* Automated Task Load Detection with Electroencephalography: Towards Passive Brain-Computer Interfacing in Robotic Surgery. *Journal of Medical Robotics Research* **02**, 1750003. https://doi.org/10.1142/s2424905x17500039 (2017).
60. Hsu, K. E., Man, F. Y., Gizicki, R. A., Feldman, L. S. & Fried, G. M. Experienced surgeons can do more than one thing at a time: Effect of distraction on performance of a simple laparoscopic and cognitive task by experienced and novice surgeons. *Surgical Endoscopy and Other Interventional Techniques* **22**, 196–201. https://doi.org/10.1007/s00464-007-9452-0 (2008).
61. Poolton, J. M. *et al.* Multitask training promotes automaticity of a fundamental laparoscopic skill without compromising the rate of skill learning. *Surgical Endoscopy* **30**, 4011–4018. https://doi.org/10.1007/s00464-015-4713-9 (2016).
62. Suárez, J. X. *et al.* Changes in brain activity of trainees during laparoscopic surgical virtual training assessed with electroencephalography. *Brain Research* **1783**, 147836. https://doi.org/10.1016/j.brainres.2022.147836 (2022).
63. Hu, J. S. L., Lu, J., Tan, W. B. & Lomanto, D. Training improves laparoscopic tasks performance and decreases operator workload. *Surgical Endoscopy* **30**, 1742–1746. https://doi.org/10.1007/s00464-015-4410-8 (2016).
64. Maimon, N. B. *et al.* Continuous Monitoring of Mental Load During Virtual Simulator Training for Laparoscopic Surgery Reflects Laparoscopic Dexterity: A Comparative Study Using a Novel Wireless Device. *Frontiers in Neuroscience* **15** (2022).
65. Zakeri, Z., Mansfield, N., Sunderland, C. & Omurtag, A. Physiological correlates of cognitive load in laparoscopic surgery. *Scientific Reports* **10**, 12927. https://doi.org/10.1038/s41598-020-69553-3 (2020).
66. Georgiou, K., Larentzakis, A. & Papavassiliou, A. G. Surgeons' and surgical trainees' acute stress in real operations or simulation: A systematic review. *The Surgeon* **15**, 355–365. https://doi.org/10.1016/j.surge.2017.06.003 (2017).
67. Campbell, T., Winkler, I., Kujala, T. & Näätänen, R. The N1 hypothesis and irrelevant sound: Evidence from token set size effects. *Cognitive Brain Research* **18**, 39 (2003).
68. Haapakangas, A., Hongisto, V. & Liebl, A. The relation between the intelligibility of irrelevant speech and cognitive performance—A revised model based on laboratory studies. *Indoor Air* **30**, 1130–1146. https://doi.org/10.1111/ina.12726 (2020).
69. Wong, B. W. L., Huo, S. & Maurer, U. Adaptation and Mismatch Negativity (MMN): Comparing First vs. Subsequent Repetitions in a Roving EEG Paradigm with Minimized Expectations. https://doi.org/10.22541/au.168323405.54729335/v1 (2023).
70. Hari, R., Sams, M. & Järvilehto, T. Auditory evoked transient and sustained potentials in the human EEG: I. Effects of expectation of stimuli. *Psychiatry Research* **1**, 297–306. https://doi.org/10.1016/0165-1781(79)90011-8 (1979).
71. Schröger, E. On the detection of auditory deviations: A pre-attentive activation model. *Psychophysiology* **34**, 245–257. https://doi.org/10.1111/j.1469-8986.1997.tb02395.x (1997).
72. Regenbogen, C. *et al.* Auditory Processing under Cross-Modal Visual Load Investigated with Simultaneous EEG-fMRI. *PLOS ONE* **7**, e52267. https://doi.org/10.1371/journal.pone.0052267 (2012).
73. Mahajan, Y., Kim, J. & Davis, C. Does working memory protect against auditory distraction in older adults?. *BMC Geriatrics* **20**, 515. https://doi.org/10.1186/s12877-020-01909-w (2020).
74. Di Liberto, G. M., O'Sullivan, J. A. & Lalor, E. C. Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. *Current biology: CB* **25**, 2457–2465. https://doi.org/10.1016/j.cub.2015.08.030 (2015).
75. Khalighinejad, B., Herrero, J. L., Mehta, A. D. & Mesgarani, N. Adaptation of the human auditory cortex to changing background noise. *Nature Communications* **10**, 2509. https://doi.org/10.1038/s41467-019-10611-4 (2019).
76. Hölle, D. & Bleichner, M. G. Smartphone-based ear-electroencephalography to study sound processing in everyday life. *European Journal of Neuroscience* **n/a**, 1–15, https://doi.org/10.1111/ejn.16124 (2023).
77. McKendricka, R. D. & Cherry, E. A deeper look at the NASA TLX and where it falls short. *Proceedings of the Human Factors and Ergonomics Society* **1**, 44–48. https://doi.org/10.1177/1541931218621010 (2018).

78. Crosse, M. J., Butler, J. S. & Lalor, E. C. Congruent Visual Speech Enhances Cortical Entrainment to Continuous Auditory Speech in Noise-Free Conditions. *Journal of Neuroscience* **35**, 14195–14204. https://doi.org/10.1523/JNEUROSCI.1829-15.2015 (2015).
79. Ullsperger, P., Erdmann, U., Freude, G. & Dehoff, W. When sound and picture do not fit: Mismatch negativity and sensory interaction. *International Journal of Psychophysiology* **59**, 3–7. https://doi.org/10.1016/j.ijpsycho.2005.06.007 (2006).
80. Roye, A., Jacobsen, T. & Schröger, E. Discrimination of personally significant from nonsignificant sounds: A training study. *Cognitive, Affective, & Behavioral Neuroscience* **13**, 930–943. https://doi.org/10.3758/s13415-013-0173-7 (2013).
81. Holtze, B., Jaeger, M., Debener, S., Adiloğlu, K. & Mirkovic, B. Are They Calling My Name? Attention Capture Is Reflected in the Neural Tracking of Attended and Ignored Speech. *Frontiers in Neuroscience* **15**, https://doi.org/10.3389/fnins.2021.643705 (2021).

Author contributions

M.R. and M.B. conceptualized the experiment. M.R. performed the data acquisition and analyzed the data, which was assisted by T.H., M.J., and M.B.. M.R. wrote the manuscript to which T.H., M.J., V.U., and MB contributed with critical revisions. All authors reviewed the manuscript.

Additional information

Competing interests

The author(s) declare no competing interests. During the preparation of this work the authors used ChatGPT (v. 3.5) and DeepL Write (acadamic style) to improve grammar and wording. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-74946-9.

Correspondence and requests for materials should be addressed to M.G.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.