Author accepted version

Using a Sequence of Earcons to Monitor Multiple Simulated Patients

Anna Hickling [1]

Birgit Brecknell [1]

Robert G. Loeb [4,1]

Penelope Sanderson [1,2,3]

[1]School of Psychology

[2]School of ITEE

[3]School of Medicine

The University of Queensland

St Lucia, QLD Australia 4072

[4]School of Medicine

University of Arizona, Tucson AZ

Narrative: 8588 words

Abstract: 249 words

Hickling, A., Brecknell, Birgit, Loeb, R. G. and Sanderson, P. (2017). Using a sequence of earcons to monitor. multiple simulated patients. *Human Factors 59* (2) 268-288. https://doi.org/10.1177/0018720816670986

Abstract

Objective. To determine whether a sequence of earcons can effectively convey the status of multiple processes, such as the status of multiple patients in a clinical setting.

Background. Clinicians often monitor multiple patients. An auditory display that intermittently conveys the status of multiple patients may help.

Method. Non-clinician participants listened to sequences of 500 ms earcons that each represented the heart rate (HR) and oxygen saturation (SpO_2) levels of a different simulated patient. In each sequence, 1, 2, or 3 patients had an abnormal level of HR and/or SpO_2. In Experiment 1, participants reported which of 9 patients in a sequence were abnormal. In Experiment 2, participants identified the vital signs of 1, 2, or 3 abnormal patients in sequences of 1, 5, or 9 patients, where the interstimulus interval (ISI) between earcons was 150 ms. Experiment 3 used the 5-sequence condition of Experiment 2, but the ISI was either 150 ms or 800 ms.

Results. Participants reported which patient(s) were abnormal with median 95% accuracy. Identification accuracy for vital signs decreased as the number of abnormal patients increased from 1 to 3, $p<.001$, but accuracy was unaffected by number of patients in a sequence. Overall identification accuracy was significantly higher with an ISI of 800 ms (89%) compared with an ISI of 150 ms (83%), $p<.001$.

Conclusion. A multiple-patient display can be created by cycling through earcons that represent individual patients.

Application. The principles underlying the multiple-patient display can be extended to other vital signs, designs, and domains.

Keywords. Sonification, medical monitoring, pulse oximetry, auditory displays, neonatal medicine

Precis. A sequence of earcons, each earcon representing the vital sign levels of an individual patient, could support eyes-free monitoring of multiple patients. The design principles could be applied to clinical contexts and also to other contexts where multiple processes must be monitored.

Introduction

In many healthcare contexts, clinicians must monitor multiple patients. Examples include nurses in a ward, an emergency department, or an intensive care unit (ICU), respiratory therapists in an ICU, or supervising anesthesiologists in an operating suite. In such settings, patients may need attention when their vital signs are not at target levels. However, clinicians may have difficulties in monitoring multiple patients because of the information environment. Visual information is often only available bedside or at a central monitoring station. Additionally, auditory alarms often provide clinically unimportant information and, as a result, they are frequently ignored or turned off (Xiao, Seagull, Nieves-Khouw, Barczak, & Perkins, 2003). Furthermore, clinicians are more likely to miss changes in patient status when working in different areas, such as other patients' rooms, medication preparation areas, supply rooms, and staff areas.

We conducted three studies to test whether an auditory display that represents the vital signs of multiple patients could provide clinicians situation awareness of monitored values. We aimed to improve current monitoring situations, where clinicians rely solely on visual displays supplemented by auditory alarms. Our auditory display uses an intermittent version of the principle of continuous informing (Ferris & Sarter, 2011; Watson & Sanderson, 2004). Specifically, it provides background reassurance when all is well, but alerts the listener when the vital signs of one or more patients depart from desired states. The vital signs tested—heart rate (HR) and oxygen saturation (SpO_2)—are represented by a single earcon, and a set of patients is represented by a sequence of earcons played in close succession. We tested an initial implementation of this concept using earcons developed by Janata and Edwards (2013) for use in a Neonatal Intensive Care Unit (NICU) or a neonatal nursery, but the design principles can be generalized to many other contexts, both clinical and non-clinical.

Earcons

The purpose of auditory displays, as for any displays, is to provide a better understanding of the relationships and dynamics of a process or system (Herrmann, Hunt, & Neuhoff, 2011). Conventional auditory alarms are often regarded as uninformative, annoying, and ineffective (Rayo & Moffatt-Bruce, 2015; Ruskin & Dirk, 2015). Therefore, researchers have been developing more informative auditory displays, such as earcons. Earcons are short, abstract auditory motifs that are played intermittently to convey information (Brewster, Wright, & Edwards, 1992). The sound of an earcon changes when the information that it represents changes.

Earcons are useful when continuous monitoring is impractical or misleading (Csapo & Wersenyi, 2013). New users must learn to associate an earcon with a given event, because there is no natural relationship between an earcon and the message it represents, as there would be with an auditory icon (Spain & Bliss, 2008). Despite this, learning to identify earcons requires little training (Brewster et al., 1992; Brewster, Wright, & Edwards, 1993; Herrmann et al., 2011). Evidence suggests that earcons do not fatigue a user, are reasonably pleasant, and are relatively undemanding on the user's memory (Blattner, Sumikawa, & Greenberg, 1989). Earcons have been used in portable devices, as sound features on user interfaces for the visually impaired (Herrmann et al., 2011), and in healthcare settings (Janata & Edwards, 2013; Watson, 2006; Watson & Gill, 2004).

For healthcare, earcons have been designed to help clinicians monitor vital signs, such as blood pressure (Watson, 2006; Watson & Gill, 2004). Watson and colleagues found that non-clinician participants could identify 9 different levels of hypertension or hypotension with high accuracy, particularly when an initial reference tone or "beacon" indicating normal blood pressure levels was included.

More recently, Janata and Edwards (2013) found that earcons were a promising means for monitoring neonates in a nursery or NICU. They designed a set of earcons that

signaled target or non-target HR and SpO_2 for premature neonates on oxygen support—the SpO_2 level for such neonates must be kept below maximum levels to avoid tissue damage (Stenson et al., 2013). HR and SpO_2 were represented by tremolo and timbre respectively. Subjectively, tremolo is the amount of corrugation or vibration in a sound, making it semantically congruent to a heart beat, whereas timbre is the amount of sharpness or brightness in a sound, making it semantically congruent with the effervescence of oxygen as a gas. The set of earcons represented all permutations of five levels of HR and SpO_2 (Very Low; Low; Normal; High; and Very High, for each vital sign). Janata and Edwards tested clinical practitioners' ability to (a) discriminate pairs of earcons as same or different, (b) identify which vital sign differentiated the pair, and (c) classify any difference as none, small or large. Most participants quickly learned to use earcons for these tasks.

Cycling Earcons for Patient Monitoring

The Janata and Edwards (2013) earcons have five levels on two dimensions, so they could in principle be applied to a wide variety of monitored signals. We hypothesized that the earcons could be placed in a sequence, to represent multiple patients. In the designs we tested, the first sound in the sequence is the earcon representing normal HR and normal SpO_2, acting as both an alert and as a reference tone. It is followed by a series of earcons, where each earcon represents the HR and SpO_2 levels of one patient. The design is such that after a period of silence, the HR and SpO_2 values for each patient would be updated, and the reference tone and earcons would sound again. We refer to this as a "cycling" approach.

We have already identified contexts in which clinicians might have to monitor multiple patients. Aiken et al. (2010) reported mean patient-to-nurse ratios per shift in the range of 4.8–6.8 patients to 1 nurse for medical-surgical wards, and 4.5–5.9 patients to 1 nurse for telemetry units. Ratios occasionally rise even higher when staff are off the floor. We decided to test across these ranges of multiple-patient monitoring, centered around a ratio of 5 patients to 1 clinician, but also testing a ratio of 1 patient to 1 clinician as a

baseline, and 9 patients to 1 clinician as an upper level. We do not anticipate that a load of 9 patients is normal or typical. As pointed out by Kantowitz (1992), it is important to know whether performance will be robust under unexpectedly high demands, and to know at what limit performance will degrade. We wanted to ensure that demands on performance would fall well within those boundaries. For a new display such as cycling earcons, it was appropriate to make such a check.

We conducted three studies to test the effectiveness of our auditory display. Experiment 1 tested whether participants could accurately report the ordinal position of any patients with abnormal vital sign levels in a sequence of 9 patients. Experiment 2 assessed whether participants monitoring either 5 or 9 patients could accurately identify abnormal vital sign levels in 1, 2, or 3 patients. Experiment 3 examined whether the rate at which the earcons were presented affected participants' abilities to encode information.

Memory Capacity for Auditory Stimuli

We anticipated that working memory would impose limitations on participants' ability to identify and report abnormal patients (Baddeley, 2012; Baddeley & Hitch, 1974). People's ability to retrieve information declines rapidly as the number of independent items to be held simultaneously in working memory increases (Cowan, 2000; Oberauer & Kliegl, 2006). We also anticipated that identifying and reporting abnormal patients would be more difficult amongst larger, rather than smaller, numbers of patients, because processing incoming patient information might interfere with the maintenance of items already in working memory (Wilsch & Obleser, 2015).

Information retrieval depends on how information was encoded and consolidated (Baddeley, 1997). *Encoding* refers to how a stimulus is initially registered in working memory, after which information is *consolidated*, or processed further, to make its representation more resistant to forgetting (Ricker & Cowan, 2014). Sounds that are consolidated into verbal labels can be maintained by covertly repeating the labels in the

phonological loop—a brief memory store in which information is maintained by vocal or subvocal rehearsal (Baddeley, 2012; Li, Cowan, & Saults, 2013). Rehearsal can build resistance to interference or decay (Baddeley, 2012; Cowan, 1988, 2000; Mercer & McKeown, 2010a), and therefore improve a user's ability to correctly identify earcons.

Alternatively, nonverbal auditory stimuli can only be retained through their sound properties rather than through phonological or semantic properties (Li et al., 2013). Participants' accuracy may be compromised if they encode and retain a series of sounds acoustically, because such stimuli are vulnerable to interference (see Nairne, 1990; Oberauer, 2009; Oberauer & Kliegl, 2006; Oberauer & Lange, 2008). Interference can occur if attention is distracted or if stimuli are partially overwritten by overlapping memory representations, a phenomenon called *feature overwriting* (Oberauer, Lange & Engle, 2004). Evidence for feature overwriting with nonverbal auditory items was demonstrated by Mercer & McKeown (2010a; 2010b). Overall, the capacity of nonverbal auditory working memory is lower than for verbal auditory items (Golubock & Janata, 2013), but people's capacity to retain nonverbal items can be increased if tones are encoded as auditory imagery (Hubbard, 2010).

Applying the Research to The Current Study

In the literature cited, only recall or recognition was examined and the materials were either purely verbal or non-verbal. Our experiments imposed different demands. In Experiment 1, participants reported only which patients had abnormal vital signs. However, in Experiments 2 and 3 participants were required to encode the Janata and Edwards (2013) earcons and then, for the abnormal vital signs, consolidate them in terms of the verbal categories of HR and SpO_2 so they could be recalled

Figure 1 shows three possible strategies for performing the latter task: verbal, auditory, or visual. Using a verbal strategy, participants immediately consolidated the sounds into verbal categories after encoding, and then rehearsed the verbal categories in the

phonological loop. With an auditory or visual strategy, after encoding the sounds would have to be consolidated into verbal categories after being retrieved from working memory.

Information may be more vulnerable to interference or decay if a participant uses auditory or visual strategies. With a verbal strategy, accuracy depends on whether the sounds can be properly encoded and consolidated between the onset of one tone and the next. As shown in Figure 2, it is unclear how long the encoding and consolidation process takes. We addressed this issue in Experiment 3 by manipulating the inter-stimulus interval (ISI) to see if it affected how accurately participants could identify abnormal vital sign levels.

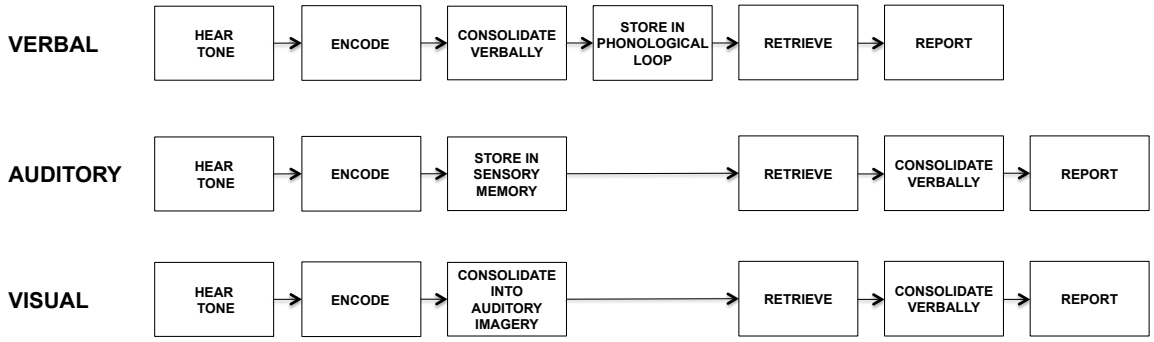

Figure 1. Three alterative strategies that participants may follow when presented with a patient in the sequence with abnormal HR and/or SpO_2 levels. Time runs left to right.

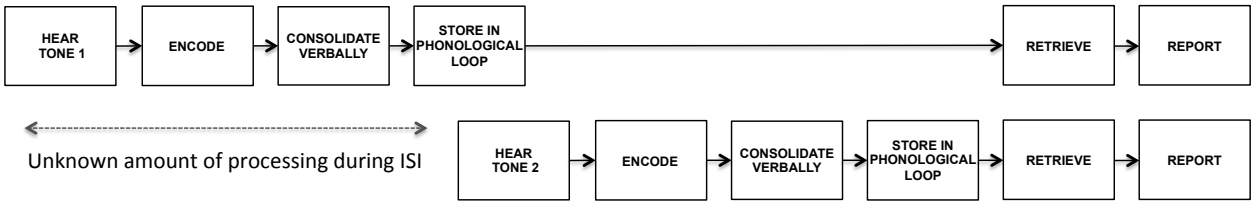

Figure 2. Diagram shows that by the time the next tone sounds; it is unclear how much processing the participant will have achieved for the first tone during the inter-stimulus interval (ISI) and therefore how effectively the next tone will be processed.

Experiment 1

In Experiment 1 we tested whether participants could accurately report which patient(s) were abnormal, when patients were represented by a sequence of earcons, each earcon representing one patient. Participants may have access to considerable information about stimuli, but the memory load associated with retaining and reporting that information in the format required by an experimenter can interfere with the memory contents (Sperling, 1960)—a phenomenon known as output interference. As a result, when participants are asked too many questions about memory contents, their performance under-represents the actual information available to consciousness. Accordingly we separated the report of which patients had abnormal vital signs (Experiment 1) and what the levels of those vital signs were (Experiments 2 and 3).

To test the limits of performance, participants were asked to monitor 9 patients and were told that one or more vital signs for 0, 1, 2, or 3 of the 9 patients would become abnormal. Their task was to report the ordinal position of any abnormal patient(s) in the sequence. We tested whether participants' accuracy changed as the number of abnormal patients increased.

The participants in Experiment 1, 2 and 3, were non-clinicians. Our goal was to test the effectiveness of the design for perception and memory with a nonclinical population before moving to more clinically-specific tests with clinicians. Simply testing clinician participants with unrepresentative tasks would not guarantee generalizability (Araújo, Davids, & Passo, 2007).

Method

Participants. Following ethics approval from The University of Queensland (15-PSYCH-4-56-AH), 13 first-year psychology students were tested. Participants included 10 females and 3 males. Their ages ranged from 17 to 33 years (*Md* = 19).

Design. We used a within-participants design and manipulated the number of abnormal patients (1, 2, or 3) in a sequence of 9 patients. The dependent variable was participants' accuracy at reporting the ordinal position of each abnormal patient in the sequence.

Participants completed 60 trials. There were 18 trials each with 1, 2 and 3 abnormal patients. We also included 6 trials with 0 abnormal patients to check that participants could recognize a completely normal sequence. The trials were presented in a random order. Participants were tested individually in a single one-hour session.

Apparatus and stimuli. The sections below describe the software and earcons used in Experiment 1.

Software. The experiment was run on a MacBook Pro laptop with a 13-inch screen. Participant responses were recorded in MS Excel™. Sounds were played through Edirol MA-7A stereo monitor speakers.

Heart Rate

Tremolo	J&E (2013)		Present study
5 cycles	VERY HIGH	*-2 SDU*	**HIGH**
4 cycles	HIGH	*-1 SDU*	*Not used*
3 cycles (6 Hz)	NORMAL	*0 SDU*	**NORMAL**
2 cycles	LOW	*+1 SDU*	*Not used*
1 cycles	VERY LOW	*+2 SDU*	**LOW**

SpO_2

Timbre	J&E (2013)		Present study
Very bright tone	VERY HIGH	*-2 SDU*	**HIGH**
Bright tone	HIGH	*-1 SDU*	*Not used*
Moderately bright tone	NORMAL	*0 SDU*	**NORMAL**
Less bright tone	LOW	*+1 SDU*	*Not used*
Pure tone	VERY LOW	*+2 SDU*	**LOW**

Figure 3. Mapping of heart rate and SpO_2 to Janata and Edwards' (2013) earcon sounds, showing the subset of Janata and Edwards' levels that was used for the earcons in Experiments 1, 2 and 3. Levels were renamed High, Normal, and Low for clarity. SDU = vital sign change from normal in standard deviation units.

Earcons. Each earcon represented one patient. We used a subset of Janata and Edwards' (2013) earcons to represent HR and SpO_2. In Janata and Edwards' full sound set, the earcon sound was designed to change when HR or SpO_2 changed by one or two standard deviation units (SDU), resulting in a total of five levels for each vital sign (see Figure 3). We used only the -2 SDU, 0 SDU and +2 SDU sounds because early piloting indicated that participants found it difficult to differentiate the -1/+1 SDU sound changes. We labeled the sounds as High, Normal, and Low, dropping the word 'Very' to avoid confusion.

Each earcon was 500 ms long, and had a fundamental frequency of 329.6 Hz. HR was represented by tremolo, which was created by modulating the amplitude of the earcon with 1, 3 or 5 cycles per earcon (see Figure 3). SpO_2 was represented by timbre, which was created using frequency modulation synthesis to modify the harmonics of the sound. A -2 SDU earcon (SpO_2 Low in our experiment) had the carrier frequency of a pure tone (flute-like sound), and a +2 SDU earcon (SpO_2 High) had a modulator frequency that was twice the carrier tone (very sharp and bright sound). The timbre of the 0 SDU earcon (SpO_2 Normal) lay mid-way between the other two.

Multiple-patient sequences. Earcons were played in a sequence that represented multiple patients. The order of the sounds (Figure 4) was (1) an initial reference tone, which was an earcon that represented normal HR and SpO_2 levels, (2) 1000 ms of silence, and (3) a sequence of earcons representing patients. Based on early piloting, we set an ISI of 150 ms between earcons in the patient sequence. The earcons were heard in close succession so that they could be easily compared to the reference tone, and so the sequence would not require a long listening time.

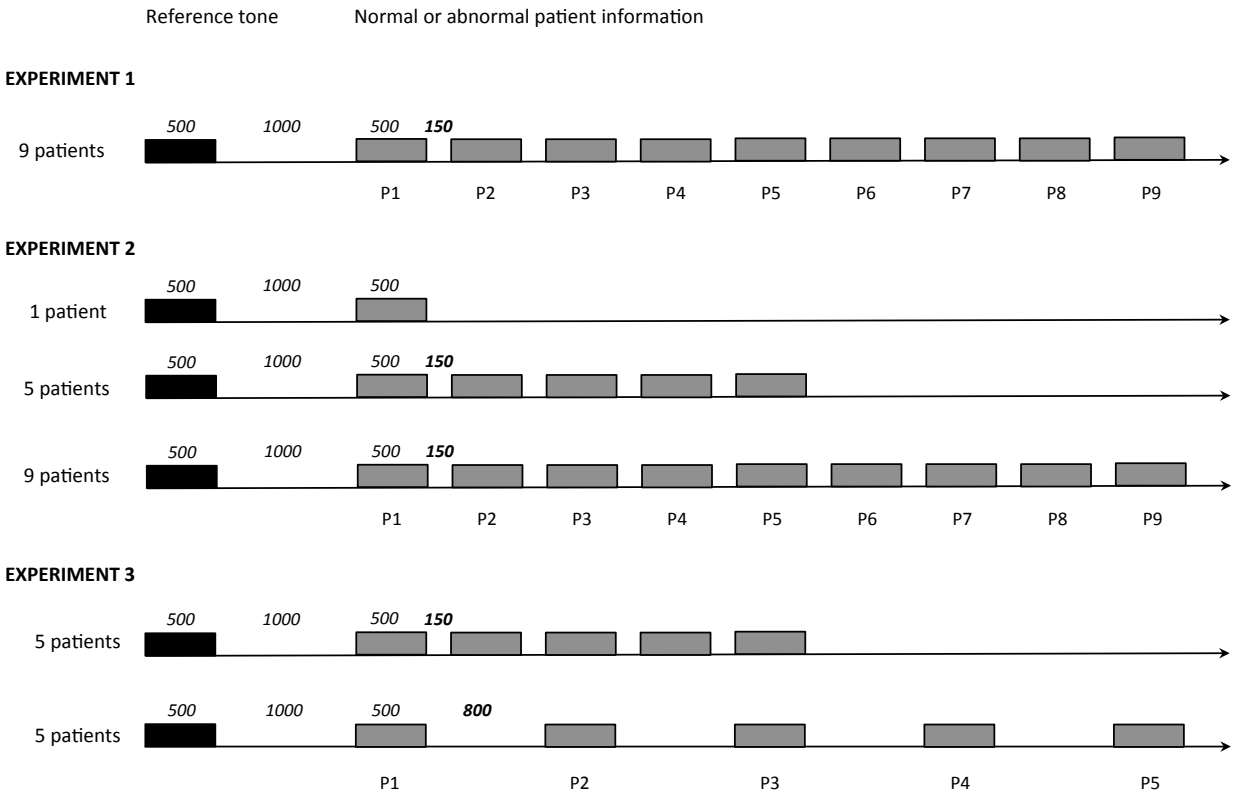

Figure 4. Earcon displays for Experiments 1, 2 and 3. After the reference tone played (black bar), representing HR Normal / SpO_2 Normal, a sequence of earcons played (gray bars), one for each patient (P) being monitored. Small numbers show durations in milliseconds of earcons and silences. For Experiment 1 and 2, diagram shows ISI of 150 ms between adjacent patient earcons, and for Experiment 3, diagram shows ISI of 150 ms or 800 ms between adjacent patient earcons.

In each block of the experiment, each trial presented one sequence of earcons. The number of abnormal patients in a sequence, their positions in the sequence, and the kind of abnormal status were restricted by a set of rules. The resulting combinations ensured that participants (a) were presented with a large range and unbiased spread of patient abnormalities across trials and (b) were not presented with three abnormal patients adjacent to each other (e.g., the 6^{th}, 7^{th}, and 8^{th} patients), a special case that might bias results. Each sequence of earcons was independent of all previous and following sequences.

Procedure. Participants were tested individually by the experimenter. Following a brief introduction, and after given written informed consent, the participant answered a demographic questionnaire. The experimenter then trained the participant to report the ordinal position of abnormal patients in the sound sequence. The participant completed 10 practice trials, and was given feedback on their accuracy after each trial. Then the participant completed 60 test trials, providing the experimenter with verbal responses. No feedback was given. A two-minute break was provided after 20 and 40 trials.

Results

Report accuracy. Inspecting the residuals revealed that the data were not normally distributed and homogeneity of variance assumptions were not met. Therefore, we used non-parametric tests. We report medians and non-parametric 95% confidence intervals based around the median.

Participants' accuracy in reporting the ordinal position of abnormal patient(s) was high (*Mdn* = 95%, CI [93%, 99%]). A Friedman test determined whether report accuracy changed as the number of abnormal patients in the sequence increased. Accuracy differed significantly across the 1, 2, and 3 abnormal patient conditions, χ^2 (2) = 10.26, p = .006. Pairwise comparisons showed that participants reported the ordinal position of 1 abnormal patient (*Mdn* = 100%, CI [94%, 100%]) more accurately than the ordinal positions of 3 abnormal patients (*Mdn* = 91%, CI [87%, 96%), p = .007. However, accuracy with 2 abnormal patients (*Mdn* = 94%, CI [92%, 100%]) did not differ from accuracy with 1 or 3 abnormal patients. The median percentage of responses that were underestimates of the number of abnormal patients was 1.67%, and the median percentage of responses that were overestimates of the number of abnormal patients was 0%.

Discussion

In Experiment 1 we tested whether participants could detect abnormal patient(s) and report their ordinal position when patients are represented by a sequence of earcons. Results

showed that participants could easily detect abnormal patients and report their ordinal positions with an overall median accuracy of 95%, which was reassuringly high.

Experiment 2

In Experiment 2 participants listened to earcon sequences with either 5 or 9 patients. We tested how effectively participants could identify the HR and SpO_2 levels of 1, 2, or 3 patients who had abnormal levels of HR and/or SpO_2. We wanted to assess participants' ability to extract information from earcon sequences representing typical patient load vs. the upper bound of patient load. Four hypotheses were tested:

1. Identification accuracy will decrease as the number of abnormal patients increases from 1 to 3, assuming that the decrease is due to working memory capacity limitations.
2. Identification accuracy will decrease as the sequence of patients increases from 5 to 9, assuming that the decrease is due to interference and time-based decay of information.
3. There may be a statistical interaction between sequence length and number of abnormal patients if, for example, identifying abnormal vital sign levels for 3 abnormal patients is more difficult amongst 9 than 5 patients.
4. Identification accuracy will be lower for one abnormal patient in a sequence of 5 patients, than in a sequence with only 1 patient, due to auditory interference.

Method

Power analysis. We conducted four statistical tests on our primary outcomes: the measure of identification accuracy for our four hypotheses. To maintain a familywise error rate of .05, we set the level of significance to .0125 (p = .05 / 4). Power was set to .95. We ran an a priori power analysis with G*Power, using data from a pilot study with 11 participants. We used a two-tailed test with the interaction effect size from the pilot study (partial η^2 = .292) because it was the smallest effect size obtained in the pilot study in the four tests described above. Our estimate of the correlation (r = 0.60) among repeated measures came from the results of the pilot. The power analysis indicated that we needed 34

participants to achieve statistical significance for the interaction term and, by implication, the three other statistical tests planned.

Participants. Ethics approval was granted by The University of Queensland as an amendment to 15-PSYCH04-56-AH. There were 40 participants, 31 females, 8 males and 1 undisclosed, with ages ranging from 17 to 40 years (*Md* = 20). None had done Experiment 1.

Design. We used a repeated measures design with two independent variables: number of patients in a sequence (*sequence length*) with two levels—5 and 9—and number of abnormal patients (*abnormal patients*) with three levels—1, 2 and 3. All participants also experienced a baseline condition with a sequence length of 1 and 1 abnormal patient. This condition does not fit into the crossed sequence length x abnormal patients design of the experiment. It was used to test separately whether monitoring 1 abnormal patient in a sequence of 5 patients was less accurate than in a sequence with 1 patient.

The dependent variable was participants' accuracy at identifying both the HR and SpO_2 levels for each abnormal patient in the sequence. For trials with 3 abnormal patients, we also ran an exploratory analysis of participants' accuracy at identifying the vital sign levels of the first, second, and third abnormal patient in the sequence.

Participants completed each of the three sequence length conditions twice, with the order of sequence lengths counterbalanced across participants and with no adjacent repetitions of sequence lengths. For blocks of trials with sequence lengths of 5 and 9, there were 4 trials each of 1, 2, and 3 abnormal patients, presented in a random order. There were 12 trials per block.

We logged participants' response times, but did not analyze them because response time was strongly constrained by the time participants needed to locate and press the response buttons on the screen, and therefore was not informative.

Demographic information (age, gender, and formal music training) was collected. In addition, a 12-item questionnaire probed participants' subjective ratings of the difficulty and other aspects of the task under the different conditions.

Apparatus and stimuli. The sections below describe the software and earcons used in Experiment 2.

Software. The experiment was run on a Dell desktop computer with a 24-inch screen, using a MATLAB Runtime R2012b 8.0 software application. The program introduced the earcon sounds and explained their mapping to HR and SpO_2. Participants were given 20 practice trials with feedback. Each practice trial could be repeated once if a participant's response was incorrect. Practice was followed by test trials; Figure 5 shows a screen shot during a test trial. Depending on the number of abnormal patients in the trial, response buttons for 1, 2, or 3 patients were presented.

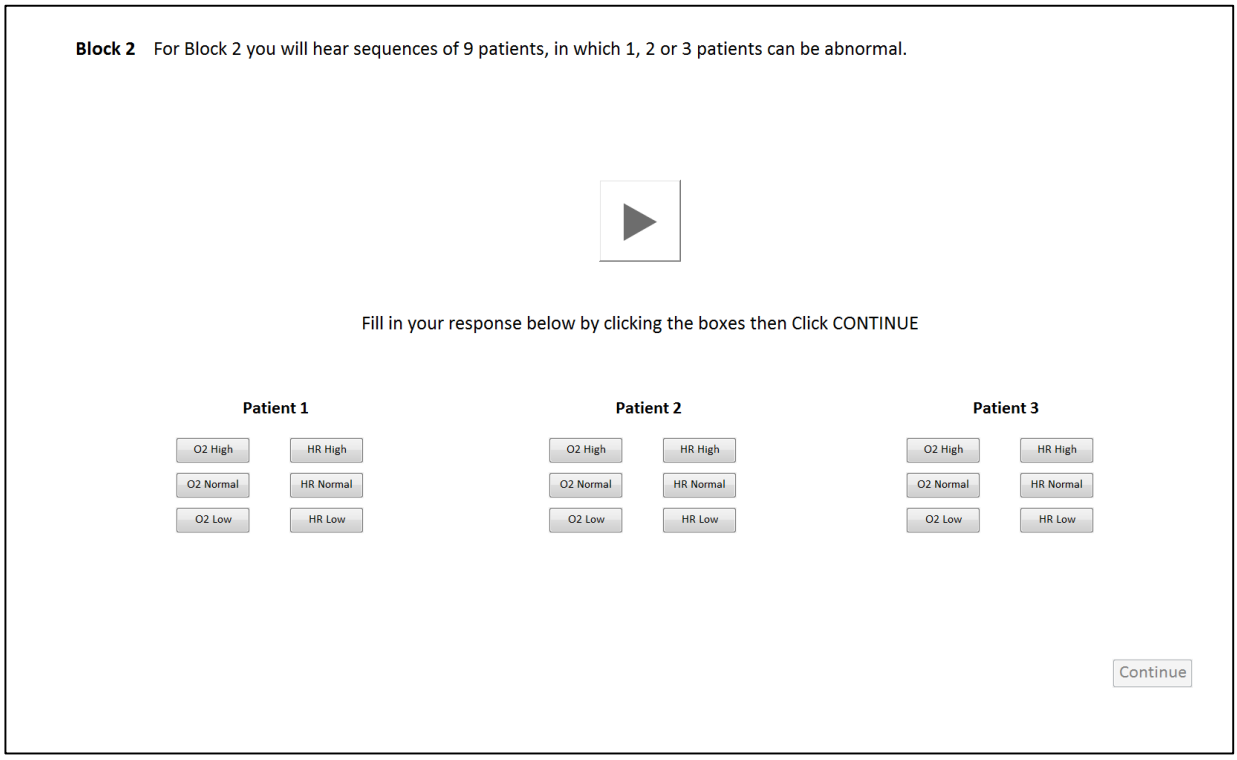

Figure 5. Interface of the software used to run Experiments 2 and 3. In the trial shown there are three response panels because there were 3 abnormal patients in the earcon sequence just experienced. The response panels appeared after the sequence of earcons had finished.

Participants completed the entire experiment using the program, and they listened to the sounds through Logitech Stereo Headset H110 headphones. The sound volume control for each computer was set to 60%, for comfortable moderate listening. Sound pressure levels could not be accurately measured from the headphones given variation in hairstyles and head garb. Screen brightness and contrast were set to 75%. Participants' responses for the practice and test trials and their questionnaire responses were saved in log files.

Earcons and multiple-patient sequences. The individual earcons used in Experiment 2 were the same as those in Experiment 1. For a sequence length of 1, there were eight possible combinations of vital signs that represented an abnormal patient (combinations of the three states for each of HR and SpO_2, minus HR Normal SpO_2 Normal). For sequence lengths of 5 and 9, we generated trials by applying the rules used in Experiment 1 for sampling 1, 2, and 3 abnormal patients and for sampling abnormal vital signs. Each sequence of earcons was independent of all previous and following sequences.

Procedure. Up to 5 participants were run at the same time in a test room with separate booths that controlled visual distraction. After a brief introduction from the experimenter, and after providing written informed consent, each participant completed a demographic questionnaire and did training and testing at a computer workstation.

Training. The mapping of vital signs to sounds was explained in the program. Participants were trained to categorize HR and SpO_2 as low, normal, or high for each abnormal patient in a sequence. Participants completed 20 practice trials, and were given feedback on the accuracy of their response after each trial. They were given two chances to select the correct answer before being shown the correct response and directed to the next trial. A sheet describing the association between the sound parameters and vital signs was displayed on a partition wall at the side of each participant's workspace throughout training and testing.

Test. Participants completed test trials that were in the same format as the practice trials. However, there was no feedback and trials were not repeated. First, participants completed three blocks containing 12 trials, one block for each sequence length condition (1, 5, or 9). A two-minute break followed, during which participants reacquainted themselves with the mappings between earcons and vital sign levels. Then participants completed a further three blocks of 12 trials, one block for each sequence length condition. Finally, participants completed a questionnaire probing their views on the ease/difficulty of the task.

Results

Identification accuracy. An inspection of residuals indicated that parametric statistics were inappropriate due to violation of normality and homogeneity of variance assumptions. Therefore, we used non-parametric tests. Medians and non-parametric 95% confidence intervals based around the median are reported. Figure 6 shows the results of each experimental condition (including the baseline); chance levels are also represented on the graph.

First, we ran a Friedman test to determine whether identification accuracy varied with the number of abnormal patients. As hypothesized, accuracy differed across the three abnormal patient conditions, χ^2 (2) = 43.13, $p < .001$, decreasing with the number of changes. We conducted pairwise comparisons on an exploratory basis, using Wilcoxon signed ranks tests—we report the results of those tests plus medians and non-parametric 95% confidence intervals. Sequences with 1 abnormal patient (*Mdn* = 97%, CI [91%, 97%]) produced significantly higher accuracy than those with 2 abnormal patients (*Mdn* = 85%, CI [75%, 91%), $p = .008$, and sequences with 2 abnormal patients produced significantly higher accuracy than those with 3 abnormal patients (*Mdn* = 67%, CI [62%, 72%]), $p = .001$. However the pairwise comparison tests were exploratory and their robustness would need to be evaluated in a subsequent confirmatory study.

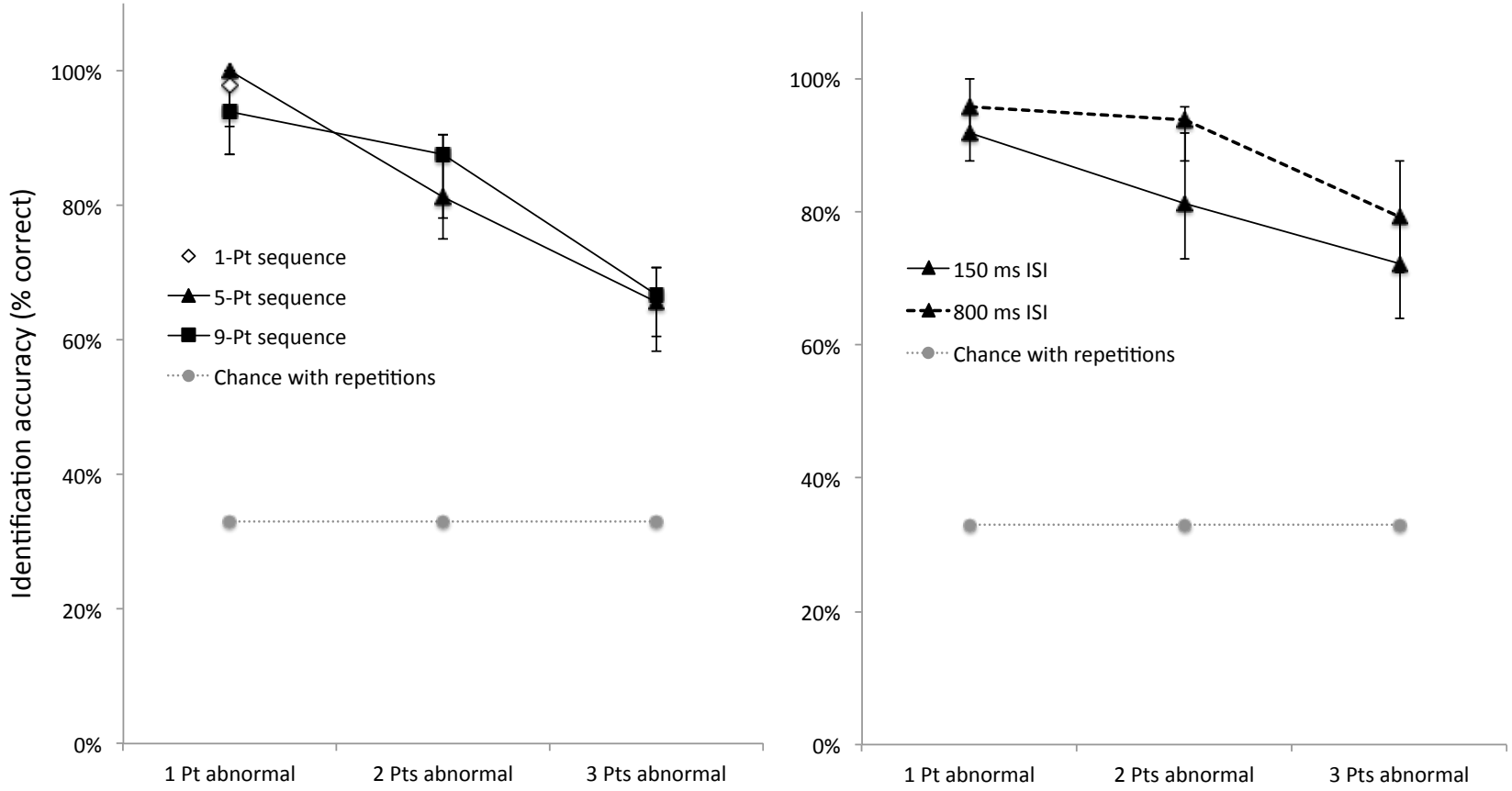

Figure 6. Identification accuracy results for Experiment 2 (left) and Experiment 3 (right), showing median percentage correct identification for each condition. In Experiment 2 all interstimulus intervals (ISI) are 150 ms. In Experiment 3, a sequence length of 5 patients only was used. Error bars are non-parametric 95% confidence intervals based on the medians. Chance levels of responding assuming potential repetitions of the same abnormality across patients in the trial are shown for comparison.

Second, contrary to our hypothesis, a Wilcoxon signed ranks test showed no significant differences in identification accuracy between sequences of 5 patients (*Mdn* = 81%, CI [74%, 86%]) and 9 patients (*Mdn* = 82%, CI [75%, 87%]), $z = .79$, $p = .428$.

Third, we tested the interaction of sequence length (5, 9) and number of patients with abnormal vital signs (1, 2, 3) using difference scores. For each participant we calculated the difference in identification accuracy between sequence lengths of 5 and 9 for each of the three levels of number of abnormal patients. A Friedman test showed that the difference in identification accuracy between the 5 and 9 sequence lengths did not differ across the number of abnormal patients, χ^2 (2) = .671, $p = .715$.

Fourth, a Wilcoxon signed ranks test determined if there were differences in how accurately participants identified 1 abnormal patient in a sequence with 1 patient (baseline) vs. a sequence of 5 patients. Contrary to our hypothesis, there was no significant difference for identifying 1 patient in a sequence of 1 patient (*Mdn* = 98%, CI [92%, 100%]) vs. a sequence of 5 patients (*Mdn* = 100%, CI [88%, 100%]), $z = -1.62$, $p = .428$.

Results of the subjective questionnaire are reported in Appendix A. Results testing the potential impact of formal music training are reported in Appendix B.

Serial position effects for conditions with three abnormal patients. We conducted a post-hoc exploratory analysis of serial position effects in how accurately participants identified the status of multiple abnormal patients. Data for 3 abnormal patients were combined for the 5 and 9 sequence lengths. A Friedman test showed significant differences in identification accuracy for the first, second, and third abnormal patients in a sequence of 3 abnormal patients, $\chi^2(2) = 7.99$, $p = .018$. Paired comparisons showed a significant difference in identification accuracy between the first abnormal patient (*Mdn* = 59%, CI [44%, 69%]) and second abnormal patient, (*Mdn* = 41%, CI [31%, 50%]), $p = .022$, suggesting a primacy effect in recall. However there were no significant differences in identification accuracy between the second and third abnormal patients, (*Mdn* = 50%, CI [38%, 56%]), $p = .721$, or

between the first and third abnormal patients, $p = .394$, suggesting there was no recency effect in recall. These findings should be tested for robustness in a fully-powered confirmatory study.

Discussion

In Experiment 2, we examined the effect of sequence length and number of abnormal patients on participants' ability to identify abnormal HR and/or SpO_2 levels. As predicted, identification accuracy decreased as the number of abnormal patients increased from 1 to 3. Accuracy dropped from 97% for 1 abnormal patient to 67% for 3 abnormal patients. These results suggest that the task challenged participants' working memory.

Unexpectedly, identification accuracy did not decrease as sequence length increased from 5 to 9 patients. The results therefore do not provide any evidence for interference or time-based decay. However, the sequence of 9 patients lasted only 2.6 seconds longer than that of 5 patients, which may not have made the sequence long enough to expose it to decay. Moreover, the additional 4 patients had normal vital signs, which may not have created any interference. Also unexpectedly, identification accuracy for 1 abnormal patient was not lower in a sequence of 5 patients than in a sequence of 1 patient. Thus, our results do not support the idea that monitoring the additional 4 normal patients adds interference, but because performance was already close to ceiling these data may not be sensitive to any difference.

Finally, the rate at which participants' accuracy at identifying abnormal patients decreased as the number of abnormal patients increased was the same for sequence lengths of 5 and 9 patients. However, our exploratory analysis suggested that when participants had to report the vital signs of three abnormal patients there was a strong primacy effect, with the first abnormal patient in sequence being more accurately reported than the second or third abnormal patient.

Overall, Experiment 2 showed that people can monitor the HR and SpO_2 of multiple patients with an auditory display with quite high accuracy, but accuracy decreases as the

number of abnormal patients increases until it is only 67% for 3 abnormal patients. Based on these results, we sought ways that accuracy could be improved, particularly for more than one abnormal patient. As discussed in the Introduction, verbal items can be resistant to decay or interference if they are maintained through rehearsal. We hypothesized that the relatively short ISI (150 ms) may not have given participants enough time to consolidate the sounds into verbal categories and rehearse them in the phonological loop. The primacy effect we found supported this interpretation.

Experiment 3

In Experiment 3 we manipulated ISI to see if a longer time between earcons would improve identification accuracy. Research has shown that a longer ISI can help participants use strategies to maintain information in working memory. For example, Ricker and Cowan (2014) found that increasing the time available for sensory and perceptual encoding made memory traces more resistant to time-based forgetting, because people can organize and execute more effective maintenance strategies such as attentional refreshing or verbal rehearsal. Redick and Lindsey (2013) supported this conclusion, suggesting that a slower presentation rate gives participants more time to rehearse information. Finally, lengthening a period of silence before an interfering stimulus enables the participant to consolidate the preceding stimulus better, reducing the effects of interference (Mercer & McKeown, 2010b). Even though lengthening the ISI would lengthen the amount of time a clinician has to attend to the sound sequence, greater accuracy might justify that disadvantage.

We compared a new ISI of 800 ms between earcons with the previous ISI of 150 ms. Sequences of 5 patients were tested for both ISIs. We chose the combination of 800 ms ISI and 5 patient earcons because it created a total time that was equivalent to a sequence of 9 patients with 150 ms intervals in Experiment 2 (see Figure 4). According to Card, Moran, and Newell's (1983) engineering model of human information processing, an 800 ms ISI would

provide time for enough cycles of the perceptual and cognitive processors to support encoding and consolidation of each earcon as it is heard. We refer to 150 ms as the 'short' ISI and 800 ms as the 'long' ISI.

We hypothesized that the longer ISI would give participants more time to consolidate the sounds into verbal categories and rehearse the categories in a phonological loop. We assumed that the extra time provided for encoding and consolidation would overcome any effects of delay or decay. Our specific hypotheses were as follows.

1. The longer ISI will significantly improve identification accuracy compared with the short ISI.
2. Identification accuracy will decrease as the number of abnormal patients increases from 1 to 3, replicating the results of Experiment 2.

Method

Power analysis. To keep a familywise Type I error rate of .05 for our two primary hypotheses, we used a significance level of .025 (p = .05 / 2) for each test. G*Power indicated that we needed a sample of N=29 to achieve statistical significance.

Participants. Ethics approval was granted by The University of Queensland as an amendment to 15-PSYC-4-56-AH. There were 31 participants, including 25 females and 6 males, with ages ranging from 17 to 46 years (*Md* = 18). None had done Experiment 1 or 2.

Design. Experiment 3 used a 2 (*ISI*: short, long) x 3 (*abnormal patients*: 1, 2, 3) repeated measures design. As in Experiment 2, the dependent variable was participants' accuracy at identifying the HR and SpO_2 levels for each abnormal patient in the sequence. The experiment had both a training and testing phase, as in Experiment 2. During testing, participants completed the two ISI conditions twice across two blocks, in one of four counterbalanced orders. The number of abnormal patients was varied randomly within each ISI condition. Trials were generated using the rules of Experiment 1 and 2 for sampling abnormalities and the locations of abnormal patients in a sequence. There were 6 trials for

each level of abnormal patients (making 18 trials per block). Subjective reactions to the task and stimuli were probed with a questionnaire.

Apparatus and stimuli. The software used for Experiment 2 was modified to fix the earcon sequences at 5 patients. In the short and long ISI conditions, earcon sequences had ISIs of 150 ms 800 ms respectively. The ISI was the time between the end of one earcon and the start of the next.

Procedure. Participants were tested in a single one-hour session in groups of one to five. All participants gave written informed consent. The procedure was the same as in Experiment 2, except for the modifications described below.

Training. The training phase was modified to include information and examples for sequence lengths of 5 only, and information and examples for both the short and long ISIs. Participants completed the same number of practice trials as in Experiment 2, with the same type of feedback. Practice trials included short and long ISIs, with 1 to 3 abnormal patients.

Testing. During testing, participants completed two blocks of trials for each ISI condition, followed by a two-minute break, before completing another two blocks of trials for each ISI condition. The number of trials per block was increased from 12 (in Experiment 2) to 18 in Experiment 3, to take advantage of additional time because we did not test sequences with 1 patient. We modified the final questionnaire to include questions about different ISIs.

Results

We report the results for the two primary outcomes, followed by secondary outcomes. Results of the subjective questionnaire are reported in Appendix A, and the potential impact of formal music training is reported in Appendix B.

Identification accuracy. A residuals analysis indicated that parametric statistics were inappropriate, so we used non-parametric tests. We report non-parametric 95% confidence intervals based around the median. The results are presented in Figure 6.

First, a Wilcoxon signed-rank test showed a significant median increase in accuracy in the long ISI condition (*Mdn* = 89%, 95% CI [85%, 95%]), compared to the short ISI condition (*Mdn* = 83%, CI [75%, 89%]), $z = 4.25$, $p < .001$. In the long ISI condition, medians for 1, 2, and 3 patients were 96%, 94% and 79% respectively.

Second, a Friedman test indicated that accuracy was significantly different across the three levels of abnormal patients, χ^2 (2) = 54.07, p < .001, decreasing as the number of abnormal patients increased.

Exploratory analyses showed that 1 abnormal patient (*Mdn* = 96%, CI [90%, 98%) was identified with significantly higher accuracy than 2 abnormal patients (*Mdn* = 89%, CI [80%, 93%]), *p* = < .001, and 2 abnormal patients were identified with significantly higher accuracy than 3 abnormal patients (*Mdn* = 76%, CI [67%, 81%]), *p* = .004.

In a further exploratory analysis we tested whether the advantage of the longer ISI was significant for each level of the number of abnormal patients. For each participant we calculated accuracy difference scores between the short and long ISI for each level of abnormal patients. A Friedman test showed that accuracy difference scores did not differ significantly across the three levels of abnormal patients, χ^2 (2) = 2.18, *p* = .336.

Serial position effects for conditions with three abnormal patients. In an exploratory investigation of potential serial position effects in memory for the status of multiple abnormal patients, data were analysed separately for the short and long ISIs when there were 3 abnormal patients.

For the short ISI, a Friedman test showed significant differences in identification accuracy for the first, second, and third abnormal patients in the sequence, χ^2(2) = 13.22, *p* = .001. Paired comparisons showed a significant difference in accuracy between the first (Mdn = 67%, CI [50%, 83%]) and second abnormal patients, (Mdn = 50%, CI [33%, 58%]), p = .006, suggesting a primacy effect in recall. There was also a significant difference in accuracy between the second and third abnormal patients, (Mdn 58%, CI [50%, 75%]) p = .006,

suggesting a recency effect in recall. However there was no significant difference in accuracy between the first and third abnormal patients, $p = 1.00$.

In contrast, for the long ISI, a Friedman test showed no significant differences in identification accuracy across the first (*Mdn* = 75%, CI [42%, 83%]), second (*Mdn* = 67%, CI [50%, 75%]), and third abnormal patients in sequence (*Mdn* 75%, CI [50%, 83%]), $\chi^2(2) = 0.496$, $p = .781$. Again, the robustness of these outcomes would need to be tested in a fully-powered confirmatory study.

Discussion

In Experiment 3 we tested if a longer ISI improved participants' accuracy at identifying abnormal HR and/or SpO_2 levels for multiple patients. As predicted, the long ISI condition improved median identification accuracy compared to the short ISI condition. Furthermore, the long ISI condition seems to have removed the primacy and recency effect seen in the short ISI condition for how effectively participants recalled the first, second, and third abnormal patients for three abnormal patients in the sequence.

These results suggest that presenting the earcons at a slower rate helps participants encode, consolidate, and maintain the items in verbal working memory. Participants' identification accuracy decreased as the number of abnormal patients increased from 1 to 3, replicating the results from Experiment 2, indicating that the task still imposed working memory challenges. However the lack of primacy and recency effects in recall with the longer ISI condition suggested that the longer ISI eased working memory challenges. The results are discussed further below.

General Discussion

We conducted three experiments to test whether participants could interpret a sequence of earcons that started with a reference tone and was followed by earcons representing HR and SpO_2 levels. Although we used earcons developed by Janata and

608 Edwards (2013) to help clinicians monitor the HR and SpO_2 levels of pre-term neonates
609 receiving supplemental oxygen, the principle of cycling intermittently through a sequence of
610 earcons could be extended to any patient population and to vital signs—and indeed to
611 multiple processes in other domains.

612 The results of Experiment 1 showed that using the earcons, participants could report
613 the number of patients with abnormal vital signs and the ordinal positions of those patients
614 with median overall accuracy of 95%. Accuracy was at 91% even for three abnormal patients.
615 The results of Experiments 2 and 3 indicated that the number of patients with one or more
616 abnormal vital signs in the sequence of earcons affects participants' accuracy in reporting
617 abnormal HR and SpO_2 levels. In Experiment 2, participants' ability to identify HR and SpO_2
618 of abnormal patients decreased from 97% to 85% to 67% as the number of abnormal patients
619 increased from 1 to 2 to 3. After encoding the first and second earcons, participants may not
620 have had time or capacity to consolidate the third earcon into verbal categories and
621 consequently its features were overwritten in working memory.

622 In Experiment 3 we examined whether participants' ability to identify vital sign levels
623 could be improved by increasing the ISI between earcons from 150 ms to 800 ms to give
624 participants more time to encode and consolidate abnormal vital signs. The long ISI condition
625 produced an overall accuracy of 89%, which was significantly higher than the overall
626 accuracy of 83% for the short ISI condition. This finding is consistent with previous research
627 demonstrating that longer ISIs can help participants encode and consolidate sounds, which in
628 turn can improve the use of maintenance strategies such as rehearsal (Mercer & McKeown,
629 2010b; Ricker & Cowan, 2014; Redick & Lindsey, 2013). With the long ISI, accuracy
630 decreased from 96% to 94% to 79% as the number of abnormal patients increased from 1 to 2
631 to 3.

632 In Experiments 2 and 3, identification accuracy decreased as the number of abnormal
633 patients increased, indicating that participants encountered capacity limitations. Previous

research estimates memory capacity to be (a) three to four verbal items or chunks when items cannot be rehearsed (Baddeley, 2012; Cowan, 2011), and (b) one to two nonverbal sounds on tasks that test simple recall or recognition (Golubock & Janata, 2013). The pattern of findings for the long ISI condition in Experiment 3 suggests that given enough time to encode, participants can reliably retain HR and SpO_2 information for two abnormal patients, making four chunks, but that information retention decreases for the third patient, who represents the potential fifth and sixth chunks of information to retain. For three abnormal patients, median identification accuracy in Experiment 2 was 67% whereas in Experiment 3 it was 72% with a short ISI, and 79% with a long ISI. Our findings are encouraging, considering that the task is more difficult than conventional recall or recognition tasks. All results are well above the probability of guessing, which we estimated at around 33%.

We expected identification accuracy in Experiment 2 to be lower for sequence lengths of 9 compared with 5, but it was not. This suggested that the additional cognitive processing required to process the four extra normal earcons in the longer sequence was not a significant source of interference. In addition, the extra 2.6 seconds over which participants had to maintain information in the longer sequence may not have contributed to the rates of decay.

Participants were equally effective at identifying 1 abnormal patient in sequence lengths of 5 as in sequence lengths of 1, but both performances were at ceiling levels. Differences may emerge under more challenging listening conditions, such as with background noise or when participants are multitasking.

In summary, the results suggest that although capacity limits are still in evidence, participants could use earcon sequences to accurately identify the number and ordinal position of abnormal patients. After 30 minutes of training, they could identify abnormal vital signs relatively accurately when there was a short interval between earcons, and more accurately when there was a longer interval.

Practical Implications

It is often argued that design of interactive systems should be based on theory (Kantowitz, 1992; Wickens, Hollands, Banbury, & Parasuraman, 2013). Our experiments suggest that memory limitations start to affect the viability of a multiple-patient auditory display when participants must identify and report the vital signs of three abnormal patients. In clinical practice, however, there may be less memory-intensive ways of using such a display when there are more than three abnormal patients.

In our experiments we only tested three levels of the two vital signs represented by each earcon—high, normal, and low. However, we envision a system of cycling earcons where more levels would be easily discriminable. This would allow more informed eyes-free monitoring of multiple patients than in current clinical settings where auditory alarms only convey that some signal has exceeded a threshold, and possibly also the type of monitor and urgency level. A system of cycling earcons could be played in an earpiece worn by the clinician, to allow mobile monitoring and decrease distraction to patients.

Experiment 1 demonstrated that participants could accurately report which patient(s) had abnormal vital signs—an ability with practical benefit in itself. Experiment 3 demonstrated that when a longer ISI is provided to encode and consolidate information in the auditory display, participants could reasonably accurately identify HR and SpO_2 levels for multiple patients. The results suggest that non-clinician listeners could quite quickly learn to extract useful information from sequences of earcons, even when the sequences were presented as independent trials. It is possible that well-motivated clinicians listening to cycles of earcons representing the changing status of their patients over time would perform even better.

If implemented in a hospital setting, the earcons in the multiple-patient display would be mapped to each patient who is being monitored. In wards, patients are typically in numbered bays, and clinicians often monitor patients who are in close proximity to each other

(for example; Beds 1 to 4). The sequence of earcons could be mapped to the numerical order of the bays, which should be familiar for clinicians, or to an arbitrary sequence of patients that makes best sense for the clinician. Even if the mapping of earcons to patients is less apparent in such a context, the auditory display would still inform listeners that a patient's condition had changed, signaling them to seek further information from conventional sources.

Limitations and Future Research

There are several limitations of the three experiments reported here that should be addressed in future research. First, the multiple-patient display is designed to play the sequence of earcons, and then to fall silent for anywhere from 15 seconds to 2 minutes before playing the next sequence. However, participants in Experiments 2 and 3 were tested solely on their ability to identify abnormal patient vital sign levels after hearing a single sequence of earcons, where each sequence was independent from all others. Future research should test participants' ability to re-orient to the multiple-patient display after the silent period, and test their ability to keep track of trends.

Second, we have not yet addressed how a cycling earcon display should behave when a patient experiences a sudden catastrophic deterioration ("crash"). If the cycling earcon display is the only monitor, and if it repeats only every minute or so, there could be a delay in getting help to the patient. A potential solution in such cases is that the next cycle could be brought forward in time, and revert to its regular cycle after that. Of course, in clinical practice a cycling earcon display would be used alongside other monitoring equipment, but it is prudent to design for all situations. In future research we will address this issue in the context of meaningful patient scenarios.

Third, participants in the present experiments had a single task: to listen to a sequence of earcons and subsequently identify abnormal patient vital signs. However, an auditory display should allow participants to complete other tasks while they maintain awareness of any deterioration (Herrmann et al., 2011; Watson & Sanderson, 2004; Woods, 1995). Future

research should add sustained cognitive or perceptual-motor load. Giving participants ongoing tasks while they listen to earcon sequences will show the kind of cognitive resources and strategies that participants use to retain items in memory under different conditions.

Fourth, there may be a complex relationship between the number of patients being monitored, the number of patients with abnormal vital signs, the rate at which the earcons are presented, and the nature of ongoing tasks. For example, lengthening the ISI beyond 800 ms might further improve identification when there is no secondary task, but it might make performance vulnerable to interference when there is a secondary task. An appropriate balance needs to be found between such considerations.

Fifth, as noted, we only tested three levels of each of the two vital signs, which resulted in 9 earcons altogether. Janata and Edwards' (2013) original set of earcons had five values for each of the two vital signs, making 25 earcons altogether. Our multiple-patient display needs to be tested with a full set of values, to see whether the present findings for sequence lengths, number of abnormal patients, and ISIs generalize to a greater number of clinical levels. This testing may result in further display design work.

Sixth, there may be auditory stimuli that are more readily learned and understood than the Janata and Edwards (2013) earcons. A key question is whether more easily learned and understood stimuli would help to overcome what appear to be fundamental auditory working memory limits. In current research, we are investigating different earcon sounds. Alternatively, there may be different learning methods such as multisensory training that improve participants' understanding and performance with the earcons (Golubock & Janata, 2013; Schlesinger, Stevenson, Shotwell, & Wallace, 2014).

Seventh, the present experiments used the pulse oximetry vital signs of HR and SpO_2. Full patient monitoring requires access to many other vital signs, relating to cardiac functioning, blood pressure, respiratory status, and so on. Further research is needed on how an eyes-free auditory display could provide further or different information for multiple patients. There has

been considerable work on sonification and earcon design for the above parameters for single patients (Loeb & Fitch, 2002; Sanderson et al., 2008, Watson & Gill, 2004; Watson & Sanderson, 2004), but not for multiple patients.

Finally, we used non-clinician participants, which may seem to be a limitation. However, as we have argued, the benefit of using clinician participants will only emerge if we use scenarios that are more clinically relevant, with dependence in patient trends over time. Our participants were tested solely on their ability to identify abnormal patient vital sign levels after hearing a sequence of earcons, where each sequence was independent from preceding and following sequences. We will test the cycling multiple-patient display with clinician participants in more clinically relevant scenarios only once we have established the most effective implementation of the display from the perspective of perception and memory.

Conclusion

Overall, the experiments indicated that a cycling earcon display might offer a viable basis for monitoring multiple processes, such as multiple patients. Participants could report which patients showed abnormal vital signs with 95% accuracy, when there were nine patients being monitored and up to three patients showing abnormal vital signs. Participants' accuracy in identifying the vital signs of up to three abnormal patients in a set of monitored patients depended on the working memory load of retaining the abnormal information, and not on the total number of patients in the monitored set. When participants were given more time to encode and consolidate information about abnormal vital signs, they could identify the HR and SpO_2 levels of three abnormal patients with 79% accuracy. Considerable further testing is needed to determine the robustness of the concept. With refinement, a cycling earcon display could potentially help clinicians monitor multiple patients in a variety of clinical contexts, and the principle could be extended to other domains where multiple processes must be monitored.

Acknowledgments

This research was supported by ARC Discovery Project DP140101822 to Sanderson, Loeb, and Liu. We thank Petr Janata for graciously providing us with the original earcons used in Janata and Edwards (2013), for the purposes of this study.

Key Points

- Clinicians often monitor the status of multiple patients, but they can be distant from physiological monitors.
- Multiple-patient monitoring through a cycling sequence of earcons, each representing a patient, may be more informative and less intrusive than alarms.
- In a laboratory study, non-clinician participants could report which patients in a sequence of 9 simulated patients had abnormal vital signs with median accuracy of 95%.
- Non-clinician participants' accuracy at identifying the vital sign levels of abnormal patients decreased as the number of simulated patients with abnormal vital signs increased from one to three, but the decrease was less pronounced when a longer time interval was used between earcons in the sequence.
- A sequence of earcons could potentially be used in applications where the status of multiple processes is monitored.

References

Aiken, L., Sloane, D., Cimiotti, J., Clarke, S., Flynn, L., Seago, J, … Smith, H. (2010). Implications of the California nurse staging mandate for other states. *Health Services Research*, *45*, 904-921.

Araújo, D., Davids, K., & Passos, P. (2007). Ecological validity, representative design, and correspondence between experimental task constraints and behavioral setting: Comment on. *Ecological Psychology*, *19*, 69-78. doi:10.1080/10407410709336951

Baddeley, A. D. (1997). *Human memory: Theory and practice*. East Sussex, England: Psychology Press.

Baddeley, A. D. (2012). Working memory: theories, models, and controversies. *Annual Review of Psychology, 63*, 1-29. doi:10.1146/annurev-psych-120710-100422

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. *The Psychology of Learning and Motivation*, *8*, 47-89. doi:10.1016/S0079-7421(08)60452-1

Blattner, M. M., Sumikawa, D. A., & Greenberg, R. M. (1989). Earcons and icons: Their structure and common design principles. *Human-Computer Interaction, 4*, 11-44. doi:10.1207/s15327051hci0401_1

Brewster, S. A., Wright, P. C., & Edwards, A. D. N. (1992). A detailed investigation into the effectiveness of earcons. In G. Kramer (Ed.), *Auditory display, sonification, audification and auditory interfaces. The Proceedings of the First International Conference on Auditory Display*, 471-498. Santa Fé, CA: Addison-Wesley.

Brewster, S. A., Wright, P. C., & Edwards, A. D. N. (1993). An evaluation of earcons for use in auditory human-computer interfaces. In S. Ashlund, K. Mullet, A. Henderson, E. Hollnagel, & T. White (Eds.), *INTERCHI'93*, 222-227. Amsterdamn, Netherlands: Addison-Wesley.

Card, S., Moran, T., & Newell, A. (1983). The psychology of human-computer interaction. Hillsdale, NJ: Lawrence Erlbaum Associates.

809 Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their
810 mutual constraints within the human information-processing system. *Psychological*
811 *Bulletin, 104*, 163-191. doi:10.1037/0033-2909.104.2.163
812 Cowan, N. (2000). The magical number 4 in short-term memory: A reconsideration of mental
813 storage capacity. *Behavioral and Brain Sciences, 24*, 87-114.
814 doi:10.1017/S0140525X01003922
815 Csapo, A., & Wersenyi, G. (2013). Overview of auditory representations in human-machine
816 interfaces. *ACM Computing Surveys, 46*(2). doi:10.1145/2543581.2543586
817 Ferris, T. K., & Sarter, N. (2011). Continuously informing vibrotactile displays in support of
818 attention management and multitasking in anesthesiology. *Human Factors, 53,* 600-611.
819 Golubock, J. L., & Janata, P. (2013). Keeping timbre in mind: Working memory for complex
820 sounds that can't be verbalized. *Journal of Experimental Psychology: Human Perception*
821 *and Performance, 39*, 399-412. doi:10.1037/a0029720
822 Herrmann, T., Hunt, A., & Neuhoff, J. G. (2011). *The Sonification Handbook*. Germany:
823 European Cooperation in Science and Technology.
824 Hinckfuss, K., Sanderson, P., Loeb, R., Liley, H., & Liu, D. (2015). Novel pulse oximetry
825 sonifications for neonatal oxygen saturation monitoring: A laboratory study. *Human*
826 *Factors: The Journal of the Human Factors and Ergonomics Society*.
827 doi:0018720815617406
828 Hubbard, T. L. (2010). Auditory imagery: Empirical findings. *Psychological Bulletin, 136*,
829 302-329. doi:10.1037/a0018436
830 Janata, P., & Edwards, W. H. (2013). A novel sonification strategy for auditory display of
831 heart rate and oxygen saturation changes in clinical settings. *Human Factors: The*
832 *Journal of Human Factors and Ergonomics Society, 55*, 356-372.
833 doi:10.1177/0018720812455433

834 Kantowitz, B. (1992). Selecting measures for human factors research. *Human Factors*, *34*,
835 387-398.
836 Lacherez, P., Seah, E., & Sanderson, P. (2007). Overlapping medical alarms are almost
837 indiscriminable. *Human Factors, 49*(4)*,* 637-645.
838 Li, D., Cowan, N., & Saults, J. S. (2013). Estimating working memory capacity for lists of
839 nonverbal sounds. *Attention, Perception, & Psychophysics, 75*, 145-160.
840 doi:10.3758/s13414-012-0383-z
841 Loeb, R. G., & Fitch, W. T. (2002). A laboratory evaluation of an auditory display designed
842 to enhance intraoperative monitoring. *Anesthesia and Analgesia, 94*, 362-368.
843 doi:10.1213/00000539-200202000-00025
844 Mercer, T., & McKeown, D. (2010a). Interference in short-term auditory memory. *The*
845 *Quarterly Journal of Experimental Psychology, 63*, 1256-1265.
846 doi:10.1080/17470211003802467
847 Mercer, T., & McKeown, D. (2010b). Updating and feature overwriting in short-term memory
848 for timbre. *Attention, Perception, & Psychophysics, 72*, 2289-2303.
849 doi:10.3758/BF03196702
850 Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition, 18*, 251-
851 269. doi:10.3758/BF03213879
852 Oberauer, K. (2009). Interference between storage and processing in working memory:
853 Feature overwriting, not similarity-based competition. *Memory & Cognition, 37*, 346-
854 357. doi:10.3758/MC.37.3.346
855 Oberauer, K., & Kliegl, R. (2006). A formal model of capacity limits in working memory.
856 *Journal of Memory and Language, 55*, 601-626. doi:10.1016/j.jml.2006.08.009
857 Oberauer, K., & Lange, E. B. (2008). Interference in verbal working memory: Distinguishing
858 similarity-based confusion, feature overwriting, and feature migration. *Journal of*
859 *Memory and Language, 58*, 730-745. doi:10.1016/j.jml.2007.09.006

860 Oberauer, K., Lange, E., & Engle, R. W. (2004). Working memory capacity and resistance to
861 interference. *Journal of Memory and Language, 51*, 80-96.
862 doi:10.1016/j.jml.2004.03.003
863 Rayo, M., & Moffatt-Bruce, S. (2015). Alarm system management: Evidence-based guidance
864 encouraging direct measurement of informativeness to improve alarm response. *BMJ*
865 *Quality and Safety, 24*, 282-286.
866 Redick, T. S., & Lindsey, D. R. B. (2013). Complex span and n-back measures of working
867 memory: A meta-analysis. *Psychonomic Bulletin & Review, 20*, 1102-1113.
868 doi:10.3758/s13423-013-0453-9
869 Ricker, T. J., & Cowan, N. (2014). Differences between presentation methods in working
870 memory procedures: A matter of working memory consolidation. *Journal of*
871 *Experimental Psychology: Learning, Memory, and Cognition, 40*, 417-428.
872 doi:10.1037/a0034301
873 Ruskin, K. J., & Dirk, H. (2015). Alarm fatigue: Impacts on patient safety. *Current Opinion*
874 *in Anesthesiology, 28*, 685–690.
875 Sanderson, P., Watson, M., Russell, W.J., Jenkins, S., Liu, D., Green, N., … Krupenia, S.
876 (2008). Advanced auditory displays and head-mounted displays: Advantages and
877 disadvantages for monitoring by the distracted anaesthesiologist. *Anesthesia and*
878 *Analgesia*, *106*(6), 1787-1797. doi:10.1213/ane.0b013e31817325cb
879 Sanderson, P., Wee, A., & Lacherez, P. (2006). Learnability and discriminability of melodic
880 medical equipment alarms. *Anaesthesia, 61,* 142-147.
881 Schlesinger, J. J., Stevenson, R. A., Shotwell, M. S., & Wallace, M. T. (2014). Improving
882 pulse oximetry pitch perception with multisensory perceptual training. *Anesthesia and*
883 *Analgesia*, *118*, 1249–1253. doi:10.1213/ANE0000000000000222.

884 Spain, R. D., & Bliss, J. P. (2008). The effect of sonification display pulse rate and reliability
885 on operator trust and perceived workload during a simulated patient monitoring task.
886 *Ergonomics, 51*, 1320-1337. doi:10.1080/00140130802120234
887 Sperling, G. (1960). The information available in brief visual presentations. *Psychological*
888 *Monographs*, *74*(11), 129.
889 Stenson, B. J., Tarnow-Mordi, W. O., Darlow, B. A., Simes, J., Juszczak, E., Askie, L., ...
890 Group, B. I. N. Z. C. (2013). Oxygen saturation and outcomes in preterm infants. *The*
891 *New England Journal of Medicine*, *368*, 2094-2104. doi:10.1056/NEJMoa1302298
892 Watson, M. (2006). Scalable earcons: Bridging the gap between intermittent and continuous
893 auditory displays. *Proceedings of the 12th International Conference on Auditory Display*
894 *(ICAD2006)*. London, UK: 20-23 June.
895 Watson, M., & Gill, T. J. (2004) Earcon for intermittent information in monitoring
896 environments. *Proceedings of the OZSCHI 2004 Conference.* Wollongong University:
897 20-23 June.
898 Watson, M., & Sanderson, P. (2004). Sonification supports eyes-free respiratory monitoring
899 and task time-sharing. *Human Factors: The Journal of Human Factors and Ergonomics*
900 *Society. 46*, 497-517. doi:10.1518/hfes.46.3.497.1636
901 Wee, A., & Sanderson, P. (2008). Are melodic medical equipment alarms easily
902 learned? *Anesthesia and Analgesia, 106*(2), 501-508.
903 Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2013). Engineering
904 psychology and human performance (4th edition). Boston, MA: Pearson.
905 Wilsch, A., & Obleser, J. (2015). What works in auditory working memory? A neural
906 oscillations perspective. *Brain Research.* Advance online publication.
907 doi:10.1016/j.brainres.2015.10.054
908 Woods, D. D. (1995). The alarm problem and directed attention in dynamic fault
909 management. *Ergonomics, 38*, 2371-2393. doi:10.1080/00140139508925274

910 Xiao, Y., Seagull, F. J., Nieves-Khouw, F., Barczak, N., & Perkins, S. (2003). Why are they
911 not responding to our alarms? Another analysis of problems with alarms. *Human Factors*
912 *and Ergonomics Society Annual Meeting Proceedings, 47*, 386-386.
913

914 **Biographies**

915 Anna Hickling received her BPsySci(Hons) from The University of Queensland in
916 2015. She is currently a research coordinator with the Cognitive Engineering Research Group
917 at The University of Queensland.

918 Birgit Brecknell is a research scientist with the Cognitive Engineering Research Group
919 at The University of Queensland. She received a BEng(Hons) in Aerospace Engineering in
920 2001, an MA in Screenwriting in 2008, and a PhD in Computer Vision in 2005, all from
921 Queensland University of Technology.

922 Penelope Sanderson is Professor of Cognitive Engineering and Human Factors in the
923 Schools of Psychology, ITEE, and Medicine at The University of Queensland, where she
924 leads the Cognitive Engineering Research Group. She received her PhD from University of
925 Toronto in 1985.

926 Robert “Butch” Loeb is Associate Professor of Anesthesiology at University of
927 Arizona and is Honorary Associate Professor in the School of Psychology at The University
928 of Queensland. He received his MD in 1983 from University of Maryland.

929

Appendix A – Questionnaire Data

We analyzed questionnaire data on an exploratory basis for both Experiment 2 and Experiment 3. Below we provide details of the questions, the results, and general conclusions.

Experiment 2

Table A1 reports the questions asked and the means and 95% confidence intervals for each question. In questions 1-5, participants rated the difficulty of the trials, their confidence in their ability to interpret the earcons, or the pleasantness of the earcon sounds. The means and 95% confidence intervals indicated that answers were generally moderate (an average value of 5 on a scale of 1 to 9). Questions 6-12 probed the difficulty of each condition of the experiment.

A repeated measures ANOVA tested for any change in participants' ratings of difficulty across the two sequence lengths (5, 9) and the three abnormal patients (1, 2, 3). Participants rated a sequence length of 9 as significantly harder than a sequence length of 5, $F(1,39) = 11.94$, $p = .001$, partial $\eta^2 = 234$. There was also a significant main effect for abnormal patients, $F(2,78) = 141.98$, $p < .001$, partial $\eta^2 = 785$ (with Greenhouse-Geisser correction). Pairwise comparisons revealed that 3 abnormal patients were rated significantly harder ($p < .001$) than 2 abnormal patients, which in turn were rated significantly harder ($p < .001$) than 1 abnormal patient ($M = 2.65$, $SE = .28$).

There was also a significant interaction between sequence length versus abnormal patients, $F(2,78) = 14.20$, $p = < .001$, partial $\eta^2 = 267$. One abnormal patient was rated as equally easy to detect for sequence lengths of 5 and 9, but 2 and 3 abnormal patients were rated as harder to detect in a sequence length of 9 than a sequence length 5.

Table A1

Experiment 2: Means and 95% Confidence Intervals for Questionnaire Items. For Questions 1, 2, 3, and 6-12, 9 is 'Extremely Hard'. For Question 4, 9 is 'Extremely Confident', for Question 5, 9 is 'Extremely Pleasant'.

Question	Mean	95% CI	
		LL	UL
Q1: Detecting an abnormal patient – easy/hard	4.13	[3.30,	4.95]
Q2: Identifying SpO_2 – easy/hard	5.00	[4.38,	5.62]
Q3: Identifying heart rate – easy/hard	4.03	[3.37,	4.68]
Q4: Confidence in identifications	5.35	[4.75,	5.95]
Q5: Pleasantness of earcon sounds	5.38	[4.75,	6.00]
Q6: 1 patient sequence, 1 abnormal patient – easy/hard	1.68	[1.36,	1.99]
Q7: 5 patient sequence, 1 abnormal patient – easy/hard	2.68	[2.09,	3.26]
Q8: 5 patient sequence, 2 abnormal patients – easy/hard	4.50	[3.97,	5.03]
Q9: 5 patient sequence, 3 abnormal patients – easy/hard	6.50	[5.97,	7.03]
Q10: 9 patient sequence, 1 abnormal patient – easy/hard	2.63	[2.03,	3.22]
Q11: 9 patient sequence, 2 abnormal patients – easy/hard	5.08	[4.56,	5.59]
Q12: 9 patient sequence, 3 abnormal patients – easy/hard	7.48	[6.99,	7.96]

In summary, ratings for general assessments of difficulty of identifying patient states, self-confidence at using the earcons and the pleasantness of the sounds all led to moderate values. Ratings for each sequence length (5, 9) crossed with each level of number of abnormal patients (1, 2, 3) led to significant main effects for each factor, indicating greater rated difficulty as sequence length and number of abnormal patients increased. There was also a significant interaction between factors. Participants rated the task harder as the number of abnormal patients increased from 1 to 3, which supported the performance results. However they also rated the task harder as sequence length increased from 5 to 9, contrary to the performance results. In addition, the questionnaire showed an interaction between sequence length and number of abnormal patients. Either the questionnaire is more sensitive to the

967 experimental manipulation than identification accuracy is, or demand characteristics of the
968 questions are influencing participants' responses.

969 **Experiment 3**

970 Table A2 reports the questions, means, and 95% confidence intervals for each
971 question. The means and 95% confidence intervals for questions 1-5 indicated that ratings
972 were generally moderate (an average value of 4.5 on a scale of 1 to 9).

973 Questions 6-12 probed the difficulty of each condition of the experiment. A 2 (ISI:
974 short, long) x 3 (abnormal patients: 1, 2, 3), a repeated measures ANOVA was used to test for
975 differences across conditions. Participants rated the sound sequences with short ISI as
976 significantly harder to work with than the sound sequences using long ISI, $F(1,30) = 17.20$, p
977 $< .001$, partial $\eta^2 = .364$. There was also a significant main effect for the number of abnormal
978 patients, $F(2,60) = 112.90$, $p < .001$, partial $\eta^2 = .790$ (with Greenhouse-Geisser correction).
979 Pairwise comparisons revealed that 3 abnormal patients were rated significantly harder to
980 work with ($p < .001$) than 2 abnormal patients, which in turn were rated significantly harder
981 ($p < .001$) than 1 abnormal patient.

982 There was also a significant interaction between sequence length and abnormal
983 patients, $F(2,60) = 5.10$, $p = < .001$, partial $\eta^2 = .167$ (with Greenhouse-Geisser correction).
984 Conditions with 1 abnormal patient were rated as similarly easy for the short and long
985 intervals, but abnormal patients of 2 and 3 were rated as harder for the short interval than for
986 the long interval.

987

Table A2

Experiment 3: Means and 95% Confidence Intervals for Questionnaire Items. All items were measured on a 9-point scale. For Questions 1-3, and 6-11, 9 is 'Extremely Hard'. For Question 4, 9 is 'Extremely Confident', for Question 5, 9 is 'Extremely Pleasant.'

Question	Mean	95% CI	
		LL	UL
Q1: Hearing an abnormal patient – easy/hard	4.29	[3.44,	5.14]
Q2: Identifying SpO_2 – easy/hard	4.32	[3.53,	5.12]
Q3: Identifying heart rate – easy/hard	4.03	[3.32,	4.74]
Q4: Confidence in identifications	5.55	[4.92,	6.18]
Q5: Pleasantness of sounds	4.45	[3.85,	5.06]
Q6: 800 ms, 1 abnormal patient – easy/hard	2.13	[1.44,	2.82]
Q7: 800 ms, 2 abnormal patients – easy/hard	3.74	[3.11,	4.38]
Q8: 800 ms, 3 abnormal patients – easy/hard	5.90	[5.14,	6.66]
Q9: 150 ms, 1 abnormal patient – easy/hard	2.61	[1.89,	3.33]
Q10: 150 ms, 2 abnormal patients – easy/hard	5.03	[4.48,	5.59]
Q11: 150 ms, 3 abnormal patients – easy/hard	7.19	[6.68,	7.71]

In summary, difficulty in identifying patient states, self-confidence at using the earcons and the pleasantness of the sounds were all rated as moderate. Ratings for each of the ISIs (150 ms, 800 ms) crossed with each level of number of abnormal patients (1, 2, 3) led to significant main effects for each factor, indicating greater rated difficulty with the ISI of 150 ms and as the number of abnormal patients increased. There was also a significant interaction: ratings of difficulty reduced more steeply as the number of abnormal patients increased for the 150 ms ISI than it did for the 800 ms ISI.

Appendix B – Impact of Music Training on Identification Accuracy

At the end of their experimental session, the participant filled out a short questionnaire that included a question asking whether they had more than 1 year of formal music training ("music trained") or not ("not music trained"). We conducted exploratory analyses of the effect of formal music training on participants' accuracy at identifying the vital signs of the abnormal patients in Experiments 2 and 3. A series of Mann-Whitney U tests were used to investigate whether music training was associated with better identification accuracy.

Experiment 2

Participants were not enrolled in the study with any consideration of their level of music training. Questionnaire answers revealed that twelve (30%) of the 40 participants had over one year of music training. Results for identification accuracy are shown in Table B1. When combining performance with 5 and 9 patients in the sequence, and with 1, 2, and 3 abnormal patients (therefore leaving aside performance for sequences with 1 patient only), participants with no music training identified the abnormal vital signs with a median accuracy of 84% whereas for participants with music training it was 91%. However there was a very large range of accuracies within each condition, leading to a failure to find a significant difference between participants with and without music training, $p = .422$. Detailed results have also been provided for each level of the number of abnormal patients, and for each combination of number of patients in the sequence and number of abnormal patients. In no case is there a significant difference across levels of music training.

Table B1

Experiment 2: Results of Mann-Whitney U tests comparing combinations of sequence length and number of abnormal patients to each level of formal music training (over 1 year of training or none). Medians and p values reported.

Condition	No Music Training	Music Training	U	Z	P
All except 1,1	84%	91%	195.5	0.812	.422
1 Abn Pt	94%	99%	214.5	1.403	.172
2 Abn Pts	82%	92%	208.0	1.182	.247
3 Abn Pts	65%	69%	199.5	0.931	.358
1,1	96%	98%	193.0	0.757	.475
5,1	94%	100%	202.0	1.086	.328
5,2	81%	89%	200.0	0.948	.358
5,3	63%	74%	230.5	1.847	.065
9,1	94%	100%	213.5	1.398	.182
9,2	84%	91%	207.5	1.171	.247
9,3	66%	69%	172.5	0.133	.896

Experiment 3

Participants were enrolled in the study in a manner that would equalize the number of participants with and without formal music training; 16 (52%) of the 31 participants had a year or more of formal music training. Results for identification accuracy are shown in Table B2.When combining results across all conditions, participants with no music training identified the abnormal vital signs with a median accuracy of 87% whereas for participants with music training it was 88%, which was not a significant difference, p = .821. There were no significant differences in identification accuracy for participants with and without formal music training, even though the median accuracy for participants with formal music training was always slightly higher than for those without.

Table B2

Experiment 3: Results of Mann-Whitney U tests comparing combinations of sequence length and number of abnormal patients to each level of formal music training (over 1 year of training or none). Medians and p values reported.

Condition	No Music Training	Music Training	U	Z	P
All conditions	87%	88%	125.0	0.198	.861
150 ISI	83%	84%	119.0	- 0.040	.984
800 ISI	88%	91%	135.0	0.593	.572
1 Abn Pt	96%	95%	123.5	0.140	.892
2 Abn Pts	89%	89%	127.0	0.277	.800
3 Abn Pts	76%	76%	128.5	0.336	.740
150,1	92%	94%	122.0	0.081	.953
150,2	81%	82%	125.0	0.198	.861
150,3	72%	73%	115.5	- 0.178	.861
800,1	96%	100%	123.5	0.148	.892
800,2	92%	95%	127.0	0.279	.800
800,3	75%	82%	138.0	0.712	.495

Discussion

In both Experiments 2 and 3, there was no statistically significant association between participants' music training and their ability to identify the vital signs of patients. The question we asked participants about their music training is strongly associated with participants' ability to learn and interpret auditory displays in previous studies (Hinckfuss et al., 2015; Lacherez, Seah, & Sanderson, 2007; Sanderson, Wee, & Lacherez, 2006; Wee & Sanderson, 2008). Given that no such association exists in Experiment 2 or 3, it seems that participants' ability to identify the status of multiple patients from an auditory display is not dependent on music training. Our analysis of the task, taken together with the literature, suggests that participants' performance is more strongly dependent on working memory.