Searching for meaning in sound:

Learning and interpreting alarm signals in visual environments

Sinè McDougall[1], Judy Edworthy[2], Deili Sinimeri[1], Jamie Goodliffe[3],

Daniel Bradley[2], James Foster[2]

[1]Bournemouth University, Fern Barrow, Poole, BH12 5BB, UK.

[2]University of Plymouth, Drake Circus, Plymouth, Devon, PL4 8AA, UK.

[3]University of Nottingham, University Park, Nottingham, NG7 2RD, UK.

Author note: We gratefully acknowledge the support and advice of Roger Schvaneveldt, University of Arizona, Charlie Frowd, University of Central Lancashire, Peter Thomas, Bournemouth University and Martin Curry, Human Factors Advantage.

Corresponding author: J. Edworthy, University of Plymouth, Drake Circus, Plymouth, Devon, PL4 8AA, UK. ; email: J.Edworthy@plymouth.ac.uk; Tel: +441752 584818.

Running head: Learning and interpreting alarm signals

Abstract

Given the ease with which the diverse array of environmental sounds can be understood, the difficulties encountered in using auditory alarm signals on medical devices are surprising. In two experiments, with non-clinical participants, alarm sets which relied on similarities to environmental sounds (concrete alarms, such as a heartbeat sound to indicate 'check cardiovascular function') were compared to alarms using abstract tones to represent functions on medical devices. The extent to which alarms were acoustically diverse was also examined: alarm sets were either acoustically different or acoustically similar within each set. In Experiment 1 concrete alarm sets, which were also acoustically different, were learned more quickly than abstract alarms which were acoustically similar. Importantly, the abstract similar alarms were devised using guidelines from the current global medical device standard (IEC 60601-1-8, 2012). Experiment 2 replicated these findings. In addition, eye tracking data showed that participants were most likely to fixate first on the correct medical devices in an operating theatre scene when presented with concrete acoustically different alarms using real world sounds. A new set of alarms which are related to environmental sounds and differ acoustically have therefore been proposed as a replacement for the current medical device standard.

Key words: Auditory alarms, eye tracking, semantic networks, IEC 60601-1-8, alarm signals

Public significance statement: Current guidance given in international standards suggests using ‘families’ of related tones to represent functions on medical devices. However, the tones are hard to distinguish from one another and their meanings are unclear. This research showed that alarm signals which were related to the real world (e.g. a heartbeat sound to indicate ‘check cardiovascular function’) rather than tones were much more easily learned especially when they could be easily discriminated from one another acoustically. Alarm signals like these direct attention more effectively to appropriate medical equipment and are set to replace the current international standard alarms in late 2019/early 2020.

The consequences of ineffective alarms have been well documented in a variety of contexts (e.g. Bliss & Acton, 2003; Drew et al., 2014; Sendelbach & Funk, 2013; Stanton & Edworthy, 1998). Over the last few years particular concern has been expressed with respect to alarm systems on medical devices, with significant steps being taken in an attempt to reduce alarm-related problems. For example, the number of deaths associated with alarm-related events led to a national summit in the United States in 2011 and the Joint Commission issued a 'Sentinel Event Alert' in 2013 with strategic recommendations for a 'frequent and persistent' problem. Since then, a raft of work has addressed the problem of over-alarming in order to reduce 'alarm fatigue' (see Cvach, 2012; Welch, 2011; Whalen et al., 2014).

Organizations concerned with medical device safety such as the Association for the Advancement of Medical Instrumentation (AAMI, US) have been centrally involved in addressing the issue of over-alarming by providing comprehensive guidance on how to reduce the number of false alarms, how to set clinical parameter limits appropriately for different groups of patients, and other measures which can be performed in the clinical work situation. AAMI also provides databases on important practical issues such as clinical parameter settings for patients with different demographics. The problems associated with alarm fatigue have distracted attention from a deeper consideration of the influence of the properties of the alarm signals themselves (Kristensen, Edworthy & Özcan, 2016), though there is little doubt that the actual alarm signals contribute to the problem and are incorporated into most definitions of alarm fatigue.

Alarm signals tend to be either abstract, tonal sequences or more simple beeps and buzzes which do not lend themselves to easy learning and identifiability. One of the reasons that the clinical soundscape is dominated by tonal alarm signals is a standard governing the safety of medical devices globally, IEC 60601-1-8. This is a lengthy document covering topics such as the use of intelligent alarms, appropriate triggering of alarms, and the reduction of false alarms. In an attempt to enhance safety, it also presents a set of alarm signals for eight specific clinical functions. Within this standard, this set of sounds is known

as the ‘reserved’ alarm signals. While medical device manufacturers are not mandatorily required to use these sounds, they are obliged to demonstrate that any alternative sounds are at least as effective as the reserved sounds. The influence of this standard is such that, even when manufacturers choose not to use the reserved sounds, they tend to use similar tonal substitutes. Given the capacity of the loudspeakers on many medical devices to produce much richer sounds, the adherence to such a narrow design remit seems unnecessarily restricting, and does not make the best use of a human auditory system that has evolved to understand the sonic world around us.

Recognizing sounds in our environment, from a frog croaking to the reassuring whirr of our computer when we switch it on, occurs effortlessly (Marcell, Malatonos, Leahy & Comeaux, 2007; Marcell, Borrella, Green, Kerr & Rogers, 2000; Shafiro, 2008). The sound landscape, or soundscape, created by groups of environmental sounds informs our understanding as events unfold. For example, ‘night-time camping’ sounds might include frogs croaking, crickets chirping, a tent zip, yawning and a mosquito buzzing. This sort of soundscape is much easier for listeners to parse than the typical soundscape of a clinical environment. Whereas the sounds in a natural environment are acoustically varied and indicate the status of the objects and events making those sounds, the alarm signals in a clinical environment are neither. Indeed, there is a considerable weight of evidence showing that concrete alarms, using real world metaphors, are better understood and recalled than abstract sounds (Belz, Robinson & Casali, 1999; Bonebright & Nees, 2007; Bussemakers & de Haan, 2000; Edworthy et al., 2017; Fagerlonn & Alm, 2010; Graham, 1999; Isherwood & McKeown, 2017; Leung, Smith, Parker & Martin, 1997; McKeown & Isherwood, 2007; McKeown, Isherwood & Conway, 2010; Stevens, Brennan, Petocz & Howell, 2009). This is not least because the prior mental frameworks or schemas associated with known sounds can be used to facilitate understanding and learning of new material (e.g. Bartlett, 1932; Bransford & Johnson, 1972; Kumaran, Summerfield, Hassabis & Maguire, 2009; Mandler, 1984; McClelland, McNaughton & O’Reilly, 1995; Moffat, Siakaluk, Sihu, Pexam, 2015; Schank & Abelson, 1977; Tse et al., 2007; Schwanenflugel, Harnishfeger & Stowe, 1988).

We would expect alarm signals which are both signs in a semiotic sense, and close auditory metaphors for an event, to be easier to learn than abstract tones where there is no link between the alarm signal and its meaning. However, environmental sounds are on average more acoustically diverse than tonal sounds, which might be an additional factor influencing the ease with which environmental sounds can be learned.

In a clinical setting, tonal alarm signals are likely to present two major sources of confusion for the listener. The first is that they are abstract in nature, meaning that it takes time for learners to develop an association between the alarm signal and its function – a connection which is made much more quickly when concrete alarm signals are used because they capitalize on our existing knowledge of environmental sounds. The second source of confusion is the degree of acoustic variation or variability in the alarm signal set. For example, all of the sounds in the current reserved IEC 60601-1-8 set have the same rhythmic pattern and lie in the same pitch range, significantly restricting the acoustic variability of the whole set of alarm signals making differentiation difficult. Abstract tonal alarms by their very nature are likely less acoustically varied than more concrete everyday sounds, making the two factors of acoustic variability and degree of concreteness naturally confounded. The interaction of these two naturally confounded factors has never been explored in the literature.

In Experiment 1 participants were presented with one of four sets of alarms varying in their concreteness (concrete vs abstract) and their acoustic variability (different/similar) to be found within each sound set. Concrete alarm signals used plausible metaphors for their intended meaning (e.g. the rattling of a small pill box to indicate drug administration rather than the sound of a dog barking). Abstract alarms, by their very nature, employ arbitrary relationships between the signal sounds and their meanings and participants were therefore expected to be more confused about the possible meanings of the alarm signals (Familant & Detweiler, 1993; Isherwood & McKeown, 2017; Petocz, Keller & Stevens, 2008; Stephan, Smith, Martin, Parker & McAnally, 2006). It was also expected that this confusion would be compounded when sounds within the set were harder to discriminate acoustically.

Conveniently, the current IEC reserved set of alarms represent one of the two extreme conditions in this 2 x 2 design.

In Experiment 2 participants were presented with the same sets of alarm signals, and were also presented with an operating theatre scene and asked to map the alarm signals that they heard to the relevant medical devices sounding the alarm (see Figure 2). This was done while their eye movements were tracked in the expectation that as participants learned the alarm-equipment mappings they would be able to locate the relevant medical devices more effectively (Bellenkes, Wickens & Kramer, 1997; Gegenfurtner, Lehtinen & Saljo, 2011; Sheridan, 1970; Schulz et al., 2014; Schriver, Morrow, Wickens & Talleur, 2008).

Experiment 1: Alarm signal Distinguishability and Learning

Difficulties with the reserved set of alarm signals in the current international medical standard are well documented, though they were well designed given what was known at the time (c.f. Block, 2008; Block, Rouse, Hakala & Thompson, 2000). Studies with both non-clinical (Sanderson, Wee & Lacherez, 2006; Williams & Beatty, 2005) and clinical participants (Lacherez, Seah & Sanderson, 2007) have shown that these alarm signals are difficult to learn even after repeated exposures. While other research has shown that clinical staff with musical training learn faster than those without (Sanderson et al., 2006; Wee & Sanderson, 2008), this only serves to confirm the difficulty associated with distinguishing between similar alarm signals, particularly since alarm signals need to be understood easily by staff with different roles and levels of experience.

IEC 60601-1-8 specifies seven alarm risk categories and one general alarm category (Kerr, 1985; Kerr & Hayes, 1983; see Appendix 1d). A key feature of the standard is that it specifies the acoustic and structural elements of the alarm signals that should sound when one of these eight categories of event occurs. The high priority alarm signals specified for each of these categories were employed in this experiment, constituting the ‘abstract similar’ alarm set. This set was compared to three other sets of eight alarm signals designed by the second author: concrete different, concrete similar, and abstract different alarm signal sets (see Appendix 1a-c respectively). On the basis of previous research, it was expected that

concrete alarms, consisting of relevant environmental sounds, would be learned more quickly since participants would be able to draw on their existing world knowledge in order to make inferences about their meaning (e.g. Isherwood & McKeown, 2017; Petocz et al., 2008; Stephan et al., 2006). In addition, when alarm signals within sets could be easily distinguished acoustically, it was expected that they would be more accurately identified than when they were more acoustically similar.

Participants were presented with the alarm signals repeatedly over a series of 15 blocks of experimental trials to mimic learning over a series of repeated exposures to the alarms. Where participants were unable to correctly identify the function when first presented with each alarm signal, they were given an additional two attempts before being given feedback about the correct sound, again designed to resemble real-life learning. Participants' learning of the alarm signal-meaning[1] relationships was assessed after completion of 1, 8, and 15 blocks of trials when they were asked to rate the perceived relatedness of all possible alarm-function relationships. This included *incorrect* as well as *correct* alarm-meaning pairs in the expectation that, when they had been learned, relatedness ratings for correct alarm-meaning pairs would be high while relatedness ratings for incorrect pairs would be low, i.e. the differentiation between correct and incorrect pairs was expected to increase showing that sound-function relationships were less confused with one another.

Method

Participants

Sixty-four non-clinical participants (47 female) were recruited from Plymouth University (M(age)= 24; range= 18-55). All participants reported normal or corrected-to-normal vision and that they had no hearing problems. Given that previous research suggests that clinical practitioners and university students typically do not differ in their ability to learn clinical alarm signals (c.f. Lacherez et al., 2007, with Sanderson et al., 2006, and Williams & Beatty, 2005), student participants were recruited on the assumption that if alarm-meaning

relationships can be learned by naïve undergraduates with no healthcare experience then they will also be accessible to healthcare professionals with varying roles and levels of experience. This study was approved by the Ethics Panel of Plymouth University and all participants gave informed consent prior to taking part.

The sample size in this experiment was comparable to those used in similar previous research. A study carried out by Isherwood & McKeown (2017) is particularly pertinent. Effect sizes were reported using partial eta squared; effect sizes were deemed large if ηp^2>.41, medium > .18, and small >.08 (Fritz, Morris & Richler, 2012). Moderate effect sizes were reported for concrete vs abstract alarm signal comparisons, easy vs difficult sound-meaning mappings, and learning across blocks of trials with a total sample size of 24, where concreteness and sound-meaning mappings were between-subjects effects (ηp^2=0.25, 0.27, and 0.37 respectively; Isherwood & McKeown, 2017). Relatedness ratings have been less commonly reported and the most relevant research did not report effect sizes. Statistically significant findings have been reported for directly related, related, and unrelated sound-meaning pairs and visual icon meaning pairs with total sample sizes of 60-63 participants, with 20 and 21 participants in each of three stimulus type groups (Stephan et al., 2006; McDougall, Curry & de Bruijn, 2001). On this basis, it was expected that the sample size in Experiment 1 should have sufficient power to detect moderate effect sizes.

Materials and Apparatus

Four sets of eight alarm signals were developed, each designed to represent the eight high-priority alarms signals in the current IEC standard; two sets were concrete, using real world metaphors; two sets were abstract, using tones to represent meaning. For each of these pairs of sets (one concrete set and one abstract set), one set of alarm signals had significant acoustic variability across the set of eight signals, while in the other set there was much less acoustic variability (the acoustically similar sets). See Appendices 1a-d for details of the acoustic and temporal features of each set of alarm signals. The degree of acoustic similarity and variability within the four alarm signal sets was examined by carrying out a

short pilot study with a small number of non-clinical participants. This pilot confirmed that the 'similar' sets were more similar to one another as a whole than the 'different' sets for both the concrete and the abstract alarms.

As far as was possible, the alarm signals were the same loudness and length. It was not possible to make the signals exactly the same length, particularly in relation to the concrete alarm signals, because the sounds used represented actual events and/or objects which meant that they were of different lengths. Some were repeated (some repeated more than once, depending on their natural length) so each alarm signal was curtailed at a point where a cycle of the sound had been completed. The average length of a sound was approximately two seconds. Equality of loudness was also difficult to achieve because of way the energy was spread over the different concrete alarm signals. However, as far as possible the Root Mean Square loudness of each sound was equated across sounds.

Because the alarm signals in the concrete different alarm set are broadly similar to those intended to be added to an update of IEC 60601-1-8, the General alarm signal used in this set was an abstract sound, as this is the sound likely to appear in the update of the standard. The reason for this is that the general alarm has a special status: it has no specific meaning or referent, making the use of a sound metaphor difficult. However, the tones used were typical of the general alarm sounds heard in both clinical and non-clinical settings on a day-to-day basis (e.g. a typical audible alarm on a medical device, a cellphone, or a computer).

Participants were presented with alarm signals using a Viglen DQ67SW computer with a Realtek High Definition Audio 24-bit 48000 Hz sound card. Participants listened to alarm signals via Behringer HPM 1000 headphones.

Procedure

Participants were randomly allocated to one of the four experimental groups; 17 participants were presented with concrete different alarm signals, 14 with concrete similar, 17 with abstract different and 16 with abstract similar sounds. They were told that they would be presented with a series of alarm signals and that they would be asked what those

alarm signals meant. Participants were presented with 15 blocks of trials, with eight trials in each, with each signal from the presentation set being presented once in random order. In the first - practice - block of trials, participants had opportunity to familiarize themselves with the signals and the experimental procedure, thereafter they were presented with 15 blocks of experimental trials. The following procedure applied for each experimental trial:-

(i) A fixation cross appeared at the center of the computer screen for 1s.

(ii) A blank screen was shown for 2s during which one alarm signal from the presentation set was played through the speakers.

(iii) A screen appeared showing all of the possible eight functions, or meanings, that the alarm signal represented with a click-box beside each function. The eight alarms and their meanings were displayed in two columns with a click-box beside each in the following fixed order: general alarm, oxygenation, temperature and power supply in the left hand column and drug administration, perfusion, ventilation and cardiovascular in right hand column. Participants were asked to click as quickly as possible on the click-box of the meaning they thought matched the signal. This screen showed until the mouse click or for a maximum of 10s. A 'Try again' message appeared on the screen for 1.5s when participants selected the wrong meaning. Participants were given two further attempts. If unsuccessful on three occasions, the correct meaning was highlighted using a red box on the screen. It would have been possible to ask all participants simply to keep making attempts until they mouse-clicked on the correct meaning, however, for those in more difficult experimental conditions this was likely to be de-motivating. Giving participants three attempts made it possible to assess how confused participants were while giving them the opportunity to learn signal-meaning associations.

(iv) There was a 500ms inter-stimulus interval between trials during which a blank screen appeared.

Participants were asked to rate the perceived relatedness of alarm signal-meaning pairs after completing 1, 8, and 15 experimental blocks of trials using a Likert scale

(1=completely unrelated; 10=very closely related). All possible alarm signal-meaning pairs were presented for rating, a total of 64 in all on each of the three occasions. As in the experimental trials, a cross appeared in the center of the screen for 1s, followed by a screen with one of the possible meanings of the sound, again at the center of the screen with the sound being presented simultaneously with the presentation. Participants had up to 10s to press 1-10 on a standard keyboard. There was a 1s inter-stimulus interval.

Design

The extent to which sounds were (a) concrete and (b) similar to one another was varied orthogonally in a 2 alarm concreteness (concrete vs abstract) x 2 alarm similarity (similar vs different) between-subjects design. Dependent variables were as follows:-

(i) *No. of accurate responses*. The number of correct meanings selected on participants' first attempt for each block of eight experimental trials. This provided an index of participants' learning.

(ii) *Number of extra attempts*. The number of incorrect second and third attempts made in each block of eight trials (maximum 16). This provided an index of the extent to which participants were uncertain of the meaning of the alarm signals.

(iii) *Pairwise relatedness ratings*. Perceived relatedness ratings (1=completely unrelated; 10=very closely related) of all pairwise alarm signal-meaning combinations after blocks 1, 8, and 15. This made it possible to examine the extent to which alarm signals were (a) perceived as being closely related to the correct (i.e. designated) meaning and (b) perceived as being distantly related to the incorrect meanings.

Results and Discussion

No. of Accurate Responses and No. of Extra Attempts

Table 1 summarizes participants' accuracy of responding. The summed frequency with which participants were accurate in each of eight trials in each block was calculated as a ratio over the possible times a correct response could have been made, e.g. in the concrete similar condition in Block 1 58 correct out of a possible 112 (8 items x 14 participants) was scored, and this is expressed as a percentage, 51.79%. Those presented

with abstract similar alarm signals had the lowest number of accurate responses, those presented with concrete similar and abstract different sounds fared considerably better, and concrete different alarm signals had high response accuracy from the outset.

The effects of alarm concreteness (concrete vs abstract), alarm similarity (similar vs different) over blocks of trials (1-15) on response accuracy was examined using Generalized Estimating Equation (GEE) analysis utilizing a negative binomial distribution with a log link function (i.e. the natural log of the dependent variable was modelled) and an autoregressive correlation matrix with the offset set to eight (the maximum number correct in each block of trials). This type of analysis is appropriate for count data which may not be normally distributed (Garson, 2013; Hardin & Hilbe, 2012). Furthermore, independence is not assumed between repeated measures obtained longitudinally (i.e. over 15 blocks of trials where performance in one block of trials is likely to affect the next; Liang & Zeger, 1986; McCullagh & Nelder, 1989). For these analyses the Wald Chi-Square statistic provided a measure of the significance or otherwise of experimental effects. Reference categories for the exponentiated coefficient were, concrete alarms, dissimilar alarms and the first block of trials, see (i)-(iii) respectively. The GEE analysis showed that several of the expected effects were statistically significant:

(i) Concrete alarms were more accurately identified than abstract alarms, $\chi^2(1)=64.08$, $p<.001$, *95% CI* = 0.19 – 0.67, *Exp(B)* = 1.46,

(ii) Dissimilar alarms were more accurately identified than similar alarms, $\chi^2(1)=40.89$, $p<.001$, , *95% CI* = 0.18 – 0.67, *Exp(B)* = 1.46,

(iii) The number of accurate responses increased as participants learned alarm-meaning associations across blocks of trials, $\chi^2(14)=159.70$, $p<.001$ [2].

Alarm concreteness interacted significantly with blocks of trials, $\chi^2(14)=74.77$, $p<.001$, as did alarm similarity, $\chi^2(14)=43.25$, $p<.001$ and there was a 3-way interaction between these three factors, $\chi^2(14)=29.49$, $p<.001$. As can be seen from Table 1, the 3-way interaction

Table 1. Percentage of times the correct meaning was selected given an alarm signal in each block of 15 experimental trials given concrete different, concrete similar, abstract different and abstract similar alarms. Percentages are derived from the frequency with which correct meanings were chosen given the number of trials in which each type of alarm signal was presented.

	Percentage of times the correct meaning was selected given an alarm signal(i.e. on first attempt)				Percentage of second and third attempts made where the correct meaning was selected given an alarm signal			
Blocks of Trials	Concrete Different (n=17)	Concrete Similar (n=14)	Abstract Different (n=17)	Abstract Similar (n=16)	Concrete Different (n=17)	Concrete Similar (n=14)	Abstract Different (n=17)	Abstract Similar (n=16)
1	95.14	51.79	49.31	26.47	2.21	38.84	41.91	63.67
2	95.14	58.93	49.31	33.82	2.94	31.70	43.01	53.91
3	93.06	66.07	48.61	35.29	1.84	30.36	40.81	53.91
4	95.83	67.86	56.94	33.09	0.37	21.88	34.19	55.86
5	99.31	74.11	65.97	43.38	0.74	19.64	29.41	47.66
6	98.61	79.46	63.19	45.59	1.47	14.29	27.57	48.83
7	97.22	76.79	71.53	44.12	1.10	17.41	23.90	46.09
8	97.92	73.21	72.92	46.32	0.00	17.41	19.49	42.97
9	97.92	82.14	74.31	40.44	1.10	15.18	22.43	46.88
10	97.92	76.79	77.08	50.00	0.00	16.96	19.12	39.45
11	100.00	79.46	72.22	53.68	0.37	8.93	25.37	39.06
12	99.31	81.25	80.56	50.00	0.37	12.50	16.91	36.72
13	99.31	79.46	79.86	56.62	1.10	12.50	16.18	32.03
14	97.92	83.04	81.25	56.62	0.37	8.93	15.81	31.25
15	99.31	83.93	83.33	54.41	0.00	8.93	15.44	33.20

* Responses on the first, second and third attempts do not sum to 100 because participants still made erroneous responses after three attempts.

appears to be the result of differing rates of learning across trials between conditions. Accuracy for concrete different alarms is uniformly high: in all other conditions learning occurs over time but this appears to be greater in the concrete similar and abstract different conditions when compared to the abstract similar condition, where performance is poorest.

Participants were able to make two additional attempts at a correct response in an experimental trial if their first attempt was unsuccessful; a maximum of 16 additional attempts was therefore possible across a block of eight experimental trials. In practice, participants rarely used all 16 extra attempts and made a number of correct responses on their first additional attempt. Table 1 shows the number of second and third attempts made for each type of alarm signal; this is highest for abstract similar sounds with lower numbers of extra attempts being made for concrete similar and abstract different alarm signals. This data was subjected to the same GEE analysis as for the number of accurate responses with the offset set to 16 and with concrete alarms, dissimilar alarms and block 1 acting as the reference categories for the exponentiated coefficient. Again, all the main effects were significant:

(i) Alarm concreteness, $\chi^2(1)=114.88$, $p<.001$, *95% CI* = -32.18 - -28.11, *Exp(B)* = 8.09, with concrete alarm signals requiring fewer extra attempts than abstract signals

(ii) Alarm similarity, $\chi^2(1)=73.78$, $p<.00$, *95% CI* = .10 – 1.43, *Exp(B)* = 2.15, with dissimilar alarms requiring fewer extra attempts than similar alarms

(iii) Blocks of trials, $\chi^2(14)=216.400$, $p<.001$, with extra attempts reducing as blocks of trials progressed.

Alarm concreteness interacted significantly with blocks of trials, $\chi^2(14)=49.14$, $p<.001$, as did alarm similarity, $\chi^2(14)=36.84$, $p=.001$. The interaction between concreteness and similarity was also significant, $\chi^2(14)=35.49$, $p<.001$ but the 3-way interaction between concreteness, similarity and blocks of trials did not reach significance, $\chi^2(14)=18.75$, $p=.07$.

As noted at the bottom of Table 1, participants could, on any given trial, continue to select the wrong alarms even after 3 attempts and so responses for first, second and third

attempts do not sum to 100%. The mean percentage persistently incorrect even after 3 attempts for each condition was 1.47%, 7.35%, 5.47% and 10.58% for the concrete different, concrete similar, abstract different, abstract similar conditions respectively.

Alarm-Equipment Relatedness Ratings

Participants were asked to rate how closely they perceived the relatedness of possible alarm signal-meaning pairings using a 1-10 scale (1=not associated at all, 10=very closely associated) after completing 1, 8 and 15 blocks of trials (see Figure 1).

A 4-way mixed analysis of variance with correctness (correct vs incorrect) and blocks of trials (1 vs 8 vs 15) as within-subjects factors and concreteness (concrete vs abstract) and alarm signal similarity (different vs similar) as between-subjects factors was carried out to examine relatedness ratings further. Data were normally distributed in each condition (all ps>.05 using Shapiro-Wilks testing), however, Mauchly's tests of sphericity were significant, $\chi^2(2)=49.83$, $p<.001$ and $\chi^2(2)=51.64$, $p<.001$ for the effect of blocks of trials and its interaction with correctness respectively so the Greenhouse-Geisser correction was used for these effects.

Participants' relatedness ratings were higher for correct than incorrect alarm-meaning pairs, M(correct) =4.43, SD = 1.19, M(incorrect) = 2.35, SD = 0.73, see Figure 1) and increased across blocks of trials, M(Block 1) = 3.22, SD = 0.76, M(Block 5) = 3.40, SD = 0.70, M(Block 9) = 3.55, SD = 0.68. Figure 1 shows that as expected ratings for correct and incorrect items diverged as participants learned, becoming higher for correct and lower for incorrect pairs producing a significant Correctness x Block interaction (see Table 2). This divergence was more marked when alarms signals were concrete and differed within the set resulting in a significant 3-way Correctness x Concreteness x Similarity interaction. The 4-way interaction between all the fixed factors appears to arise because ratings in the concrete different condition (Figure 1(a)) do not diverge over blocks of trials but instead are well differentiated from outset).

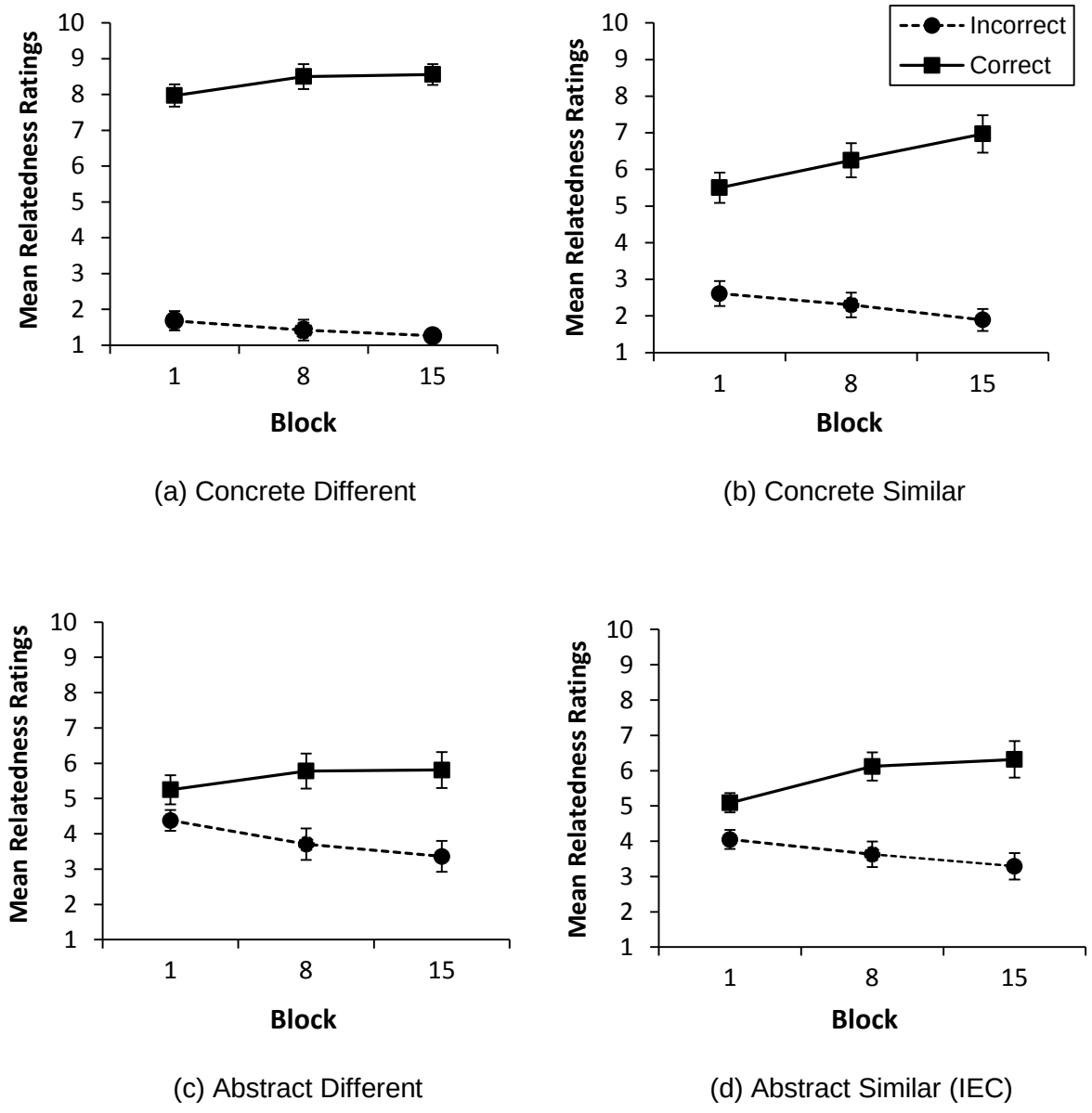

Figure 1. Relatedness ratings for correct and incorrect alarm signal-meaning combinations completed after learning trials in blocks 1, 8, and 15. Error bars show the standard error of the mean.

Table 2. Relatedness ratings of alarm signal-meaning pairs: Summary of significant main effects and interactions from 4-way analyses of variance examining the effects of correctness (correct vs incorrect alarm-meaning relationships), blocks of trials (1, 8, 15), alarm signal concreteness (concrete vs abstract), and alarm signal similarity (different vs similar).

	Df	F	p	η^2
Main effects*				
Correctness	1,63	163.84	<.001	.722
Blocks of trials	1.29, 81.17	190.56	<.001	.752
Interactions				
Correctness x Block	1,63	136.51	.001	.684
Correctness x Concreteness	1,63	4.42	.04	.022
Correctness x Concreteness x Similarity	1.29, 81.17	7.35	.009	.105
Correctness x Block x Concreteness x Similarity	1.28, 80.50	6.93	.006	.099

*Where effects are not significant they have not been included in this table.

Use of alarm-equipment relatedness ratings and confusions to determine alarm signal and set usability. An identical 4-way analysis of variance was carried out on the relatedness ratings data by-items, rather than by-participants, and revealed a very similar pattern of findings (for the sake of brevity, it is not reported here). Table 3 shows how the efficacy of the alarm signals in newly designed sets could be considered simply and effectively. The assumption is that the meaning of the alarm signal is unclear when the difference between correct and incorrect relatedness ratings is small. This lack of clarity is also manifest in performance *inaccuracy*, the percentage of occasions on which the meaning chosen given the alarm signal is incorrect. The extent of the confusion is also apparent in the number of other meanings chosen by participants for a given item. The number of meanings confused is lowest for concrete different and highest for abstract similar alarms. Table 3, although descriptive in nature, presents convergent data which makes it possible to target weak or strong alarms within a set, e.g. the perfusion alarm signal in the concrete different set appears to be most easily confused with other meanings, while the general alarm signal in the abstract similar set appears to be effective relative to others in the set because it is tonally distinct. While this is not a focus in this paper, such an approach might be usefully developed for more general use in alarm signal design and development.

Table 3. Mean difference between relatedness ratings for correct and incorrect alarm signal-meaning pairs obtained after learning blocks 1, 8 and 15, percentage of times incorrect meanings were assigned given all possible attempts, and the number of other meanings selected for each alarm signal.

	Block 1			Block 8			Block 15		
Alarm signal	Difference correct/ incorrect ratings	% Incorrect	No. of other meanings selected	Difference correct/ incorrect ratings	% Incorrect	No. of other meanings selected	Difference correct/ incorrect ratings	% Incorrect	No. of other meanings selected
				Concrete Different					
General	3.37	15.00	3	3.70	0.00	0	4.83	0.00	0
Cardiovascular	2.17	5.26	1	2.11	0.00	0	3.66	0.00	0
Drug Admin	2.77	10.00	2	3.63	0.00	0	4.47	0.00	0
Oxygenation	2.89	14.29	3	3.44	0.00	0	3.50	0.00	0
Perfusion	4.16	26.09	5	4.79	10.00	2	5.09	0.00	0
Power Supply	2.30	15.00	3	3.47	5.26	1	3.46	5.26	0
Temperature	1.94	5.26	1	4.23	0.00	0	4.47	0.00	1
Ventilation	2.86	14.29	2	2.76	10.00	2	3.56	0.00	0
Alarm signal Set Mean	*2.81*			*3.52*			*4.13*		
				Concrete Similar					
General	0.56	56.00	6	2.70	42.11	5	2.50	31.58	5
Cardiovascular	-0.14	20.00	3	1.16	27.78	5	2.14	18.75	3
Drug Admin	1.14	47.83	5	2.40	27.78	4	3.63	0.00	0
Oxygenation	0.61	62.50	5	1.63	18.75	3	2.57	18.75	3
Perfusion	1.63	65.52	7	2.27	23.53	3	3.30	7.14	1
Power Supply	0.10	78.13	6	1.37	52.38	5	1.24	33.33	4

Temperature	1.61	38.89	4	2.14	35.29	5	2.74	20.00	3
Ventilation	2.24	23.53	2	2.13	27.78	3	3.60	0.00	0
Alarm signal Set Mean	*0.97*			*1.98*		*4.12*			*2.37*
				Abstract Different					
General	0.76	31.82	5	2.03	31.82	4	3.04	16.67	3
Cardiovascular	0.50	74.29	7	1.45	34.78	5	1.16	31.82	4
Drug Admin	2.70	45.83	6	4.27	24.38	4	4.14	16.67	3
Oxygenation	2.22	56.62	7	3.59	40.00	5	3.60	26.32	4
Perfusion	-0.49	77.78	7	2.22	51.72	7	2.58	45.83	5
Power Supply	-0.43	50.00	7	-0.71	41.67	6	1.31	37.50	4
Temperature	0.71	59.26	6	2.68	25.00	3	3.07	25.00	4
Ventilation	2.14	72.97	6	2.12	55.56	6	1.87	41.67	5
Alarm signal Set Mean	*1.01*			*2.21*			*2.60*		
				Abstract Similar					
General	4.78	43.00	7	6.74	0.00	0	5.62	0.00	0
Cardiovascular	-0.24	80.49	6	1.28	65.92	6	1.45	60.71	6
Drug Admin	-0.79	75.68	7	0.75	74.86	6	1.87	63.64	6
Oxygenation	0.37	66.67	6	1.98	57.69	6	3.27	38.09	5
Perfusion	0.63	75.68	7	3.59	52.00	7	3.86	54.17	7
Power Supply	-0.02	88.37	7	0.61	64.52	7	2.57	50.00	7
Temperature	2.64	74.19	7	3.58	52.00	6	3.17	40.91	4
Ventilation	0.91	82.05	6	1.47	62.50	5	2.38	52.00	4
Alarm signal Set Mean	*1.04*			*2.50*			*3.02*		

Experiment 2: Learning and distinguishing alarm signals in a visual environment

In healthcare settings alarm signals are used to direct staff attention to clinically critical patient information on medical equipment. However, few studies have examined the way in which alarm signals direct visual attention or examined the changes in eye movements as alarm signal-meaning associations are learned (but see Dehais et al., 2014; Kodappully, Srinivasan & Srinivasan, 2016; Stevenson, Schlesinger & Wallace, 2013). Experiment 2 sought to replicate and extend the findings of Experiment 1 using the same sets of alarm signals, while also examining the efficacy with which participants' visual attention was directed to appropriate equipment. Participants were shown a static operating theatre scene and the efficacy of their eye scan paths was tracked as they selected piece of medical equipment associated with the alarms they heard. It was expected that the number of fixations participants required prior to fixating on the correct piece of equipment would reduce as participants learned alarm signal-equipment associations (e.g. Al-Moteri, Symmons, Plummer & Cooper, 2017; Bellenkes, Wickens & Kramer, 1997; Gegenchambers furtner et al., 2011; Schulz et al., 2014; Schriver, Morrow, Wickens & Talleur, 2008). As in Experiment 1, the ease with which alarm signal-equipment relationships would be learned would depend on the concreteness and acoustic discriminability of the alarm signals. While the use of the static scene cannot be equated with a live operating theatre situation, it provides visual sense and context that would not be present in a list of options (as in Experiment 1) and gave participants the opportunity to search a visual scene so providing greater face validity.

Method

Participants

Forty non-clinical participants (28 female) were recruited from Bournemouth University (M(age) = 25; range 18-60). All reported no known hearing problems and had either normal or corrected-to-normal vision. This study was approved by the Ethics Committee of Bournemouth University and all participants gave informed consent prior to taking part.

Materials & Apparatus

The alarm signal sets used were identical to those in Experiment 1. The experiment was conducted using an Eyelink 1000+ with 75cm distance between the computer screen and headrest linked to a Hewlett Packard Compaq Elite 8300 SFF computer with a 23-inch HP EliteDisplay E231 screen. The eye tracker sampled the position of the right eye with a frequency of 1 KHz. Experimental trials were controlled using Experiment Builder software (SR Research Ltd., 2015): this software is a visual experiment creation tool designed for use with the Eyelink 1000+.

Figure 2 shows the image of the operating theatre presented to participants with the pieces of equipment associated with each alarm signal, each representing a specific function related to the alarm signals. On the image, relevant interest areas are superimposed, outlined in black, and the meaning of the associated alarm signals is shown. The image used is from a hospital in the United Kingdom. The locations of the alarms were selected on the basis that they were both plausible (developed through discussions with relevant clinicians) and represented a good spread of alarm locations across the screen although in real clinical situations the locations of the specific alarms may be somewhat different. The image was adjusted using Adobe Photoshop CS6 so that it was possible to visually identify one plausible piece of equipment for each alarm signal and ensure that there was enough separation in space between pieces of equipment to create distinct interest areas. An infusion pump was added to the image in order to represent ‘check drug administration’. Importantly, the use of a static image meant that participants were able to use a headrest, increasing the accuracy and reliability of the calibration of the eye tracker.

The fixation index is the number of fixations made from the beginning of the trial prior to fixating on the correct interest area (for reviews of potential eye tracking measures see Al-Moteri et al., 2017; Asan & Yang, 2015). It acted as a measure of the extent to which participants were confused, moving from one piece of equipment to the next, prior to localizing attention on the piece of equipment (i.e., the correct interest area). Other measures of eye tracking such as number and length of fixations were not used since they were likely to be difficult to interpret: participants might fixate more frequently and for longer

as a result of learning alarm signal-equipment relationships but equally might fixate less on correct areas of interest (AOIs) once the appropriate visual cue had been viewed. Similarly measures of transitions and movements back to previous AOIs might also be regarded as a candidate eye movement measure since participants having difficulty learning alarm signal-equipment pairings might tend to move more frequently from one AOI to the next and back as they searched for an appropriate piece of equipment. However, other researchers have noted that those with greater expertise may adopt different strategies, either focusing on many points quickly, or carrying out a limited search to specific points within a scene (Tiersma, Peters, Mooij & Fleuren, 2003), making these measures difficult to interpret.

Procedure

Participants were assigned to one of the four alarm signal groups in strict serial order. This method of allocation to groups ensured equal numbers in each participant group. Systematic bias in allocation to groups was extremely unlikely because the serial order allocation was strictly applied and participants indicated their interest in taking part via an electronic participation system where appointment times were allocated automatically. Participants were seated in front of the computer screen with their chin placed on a head rest. Nine-point calibration was used to ensure reliable measurement of eye movements. Participants were asked to imagine that they were in an operating theatre and that they would hear alarm signals coming from the in-theatre equipment. In order to ensure 'patient safety', they would need to click with the mouse as quickly and as accurately as possible to indicate which piece of equipment should be checked given the alarm signal.

Familiarization with alarm signals. Prior to commencing experimental trials, participants were given the opportunity to familiarize themselves with the alarm signals. Then each piece of equipment was highlighted, outlined in red, for 12s with the meaning of each alarm displayed in a label beneath the relevant area. The relevant alarm signal was presented after 1s, 5s and 9s during the 12-second period. Participants were instructed to listen carefully so that they could click on the appropriate equipment when they heard the alarm signals later in the experiment.

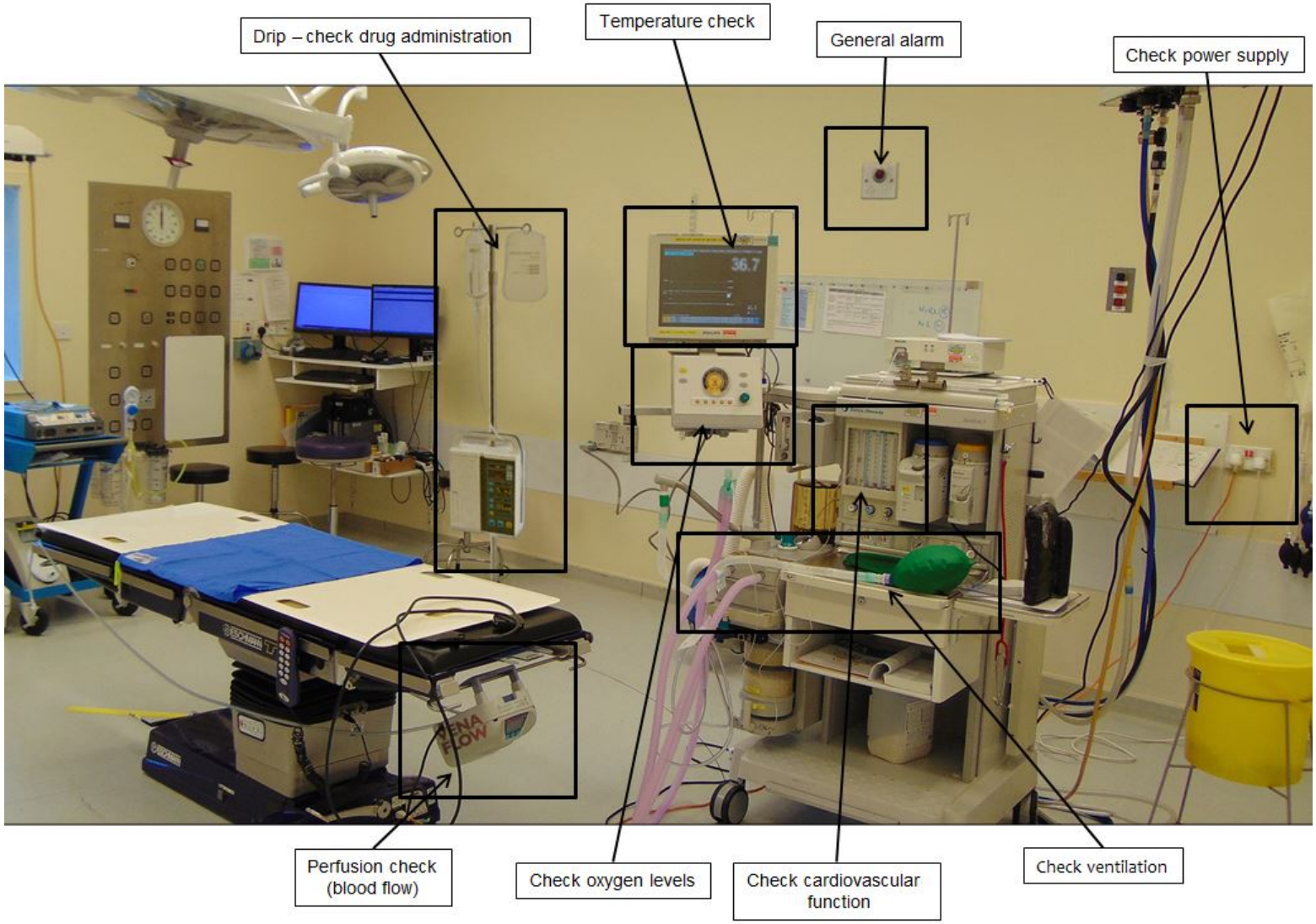

Figure 2. Picture of operating theatre used in Experiment 2 showing eye tracker interest areas and alarm meanings associated with each piece of equipment

Experimental trials. Each trial commenced with an alarm signal. Participants then used the computer mouse to indicate the equipment they thought corresponded to the alarm signal they had heard. If participants clicked on the correct piece of equipment, the equipment was outlined in white for 1s before moving on to the next trial. If participants clicked on the wrong equipment the outline did not appear and the words 'Try again' appeared at the top of the screen for 1.5s. After three attempts, the equipment was outlined in white for 1.5s before moving on to the next trial. There was a 1s inter-stimulus interval between trials.

Participants were presented with nine blocks of eight experimental trials in which each alarm signal was presented once in random order. Nine blocks of trials were presented instead of 15 blocks because learning was observed in Experiment 1 after nine blocks of trials and because pilot testing indicated that those presented with the more difficult to learn alarm sets found 15 blocks of trials arduous. As in Experiment 1, ratings were obtained of all possible alarm signal-referent pairings with the referents in Experiment 2 being the pieces of equipment depicted in the operating theatre scene. Meaning pairings were obtained, this time after one, five, and nine blocks of trials.

Design

A 2 alarm concreteness (concrete vs abstract) x 2 alarm similarity (similar vs different) x 9 blocks of trials (1-9) with alarm concreteness and similarity as between-subjects factors and blocks of trials as a within-subjects factor. Dependent variables were the same as for Experiment 1 with the addition of fixation data from eye tracking as follows:-

(i) *No. of accurate responses*. Selection of the appropriate piece of operating theatre equipment given the alarm signal on participants' first attempt.

(ii) *No. of extra attempts*. The number of incorrect second and third attempts made in each block of 8 trials (maximum 16).

(iii) *Pairwise relatedness ratings*. A seven-button box was used to collect relatedness ratings so the Likert scale ranged from 1=very closely related to 7=completely unrelated rather than 1-10 as in Experiment 1.

(iv) *The fixation index*. The number of fixations made from the beginning of the trial prior to fixating on the correct interest area. The fixation index was obtained for (a) participants' first attempt and (b) participants second and third attempts.

Results and Discussion

No. of Accurate Responses and No. of Extra Attempts

Table 4 summarizes the accuracy of participants' responses. Values in the table are the frequency with which correct meanings were selected, calculated as a percentage of the number of times the alarm signals were presented (e.g. for concrete different alarms in block 1 this ratio was 66/80 and for abstract similar alarms 24/80, producing percentages of 82.50 and 30.00 respectively). Table 4 shows that the group presented with concrete different alarm signals identified alarm signal-equipment mappings more accurately from the outset, those in the concrete similar and abstract different groups reached similar levels of accuracy only after nine blocks of trials, while those in the abstract similar condition continued to make a large number of errors throughout.

The GEE analysis used to examine the accuracy data was identical to Experiment 1 except that there were nine levels in the blocks of trials within-subjects factor. In this instance, abstract alarms, similar alarms and block 1 acted as reference categories for the exponentiated coefficient. The GEE model revealed that all three fixed factors were significant:

(i) Alarm concreteness, $\chi^2(1) = 23.43$, $p < .001$ with concrete alarms superior to abstract alarms, *95% CI* = .-.37 - -.03, *Exp(B)* = 0.82 .

(ii) Alarm similarity, $\chi^2(1) = 18.05$, $p < .001$ with different alarms superior to similar alarms, *95% CI* = -.32 – 0.01, *Exp(B)* = 0.86.

(iii) Blocks of experimental trials, $\chi^2(1) = 107.58$, $p < .001$ [1] with learning increasing across trials.

There was also a significant 3-way interaction between the effects of alarm concreteness, alarm similarity, and blocks of trials, $\chi^2(1) = 22.35$, $p = .004$, resulting from higher response

Table 4. Percentage of times the correct alarm meaning was selected given an alarm signal in each block of nine experimental trials given concrete different, concrete similar, abstract different and abstract similar alarms. Percentages are derived from the frequency with which correct equipment was chosen given the number of trials in which each type of alarm signal was presented.

	Percentage of times the correct meaning was selected given an alarm signal (i.e. on first attempt)				Percentage of second and third attempts made where the correct meaning was selected given an alarm signal			
Blocks of Trials	Concrete Different (n=10)	Concrete Similar (n=10)	Abstract Different (n=10)	Abstract Similar (n=10)	Concrete Different (n=10)	Concrete Similar (n=10)	Abstract Different (n=10)	Abstract Similar (n=10)
1	82.50	45.00	47.50	30.00	13.13	45.63	43.13	60.00*
2	80.00	38.75	37.50	30.00	13.75	51.25	49.38	60.00
3	88.75	53.75	47.50	51.25	7.50	40.00	40.00	37.50
4	93.75	58.75	52.50	45.00	5.00	33.13	36.25	48.75
5	93.75	68.75	65.00	46.25	4.38	26.88	28.13	46.25
6	96.25	63.75	62.50	47.50	2.50	28.13	27.50	43.13
7	90.00	67.50	75.00	45.00	5.63	21.88	20.00	40.00
8	97.25	76.25	62.50	50.00	1.88	17.50	26.25	38.75
9	96.25	82.50	78.75	45.63	3.13	12.50	17.50	32.50

* Responses on the first, second and third attempts do not sum to 100 because participants were still making erroneous responses after three attempts.

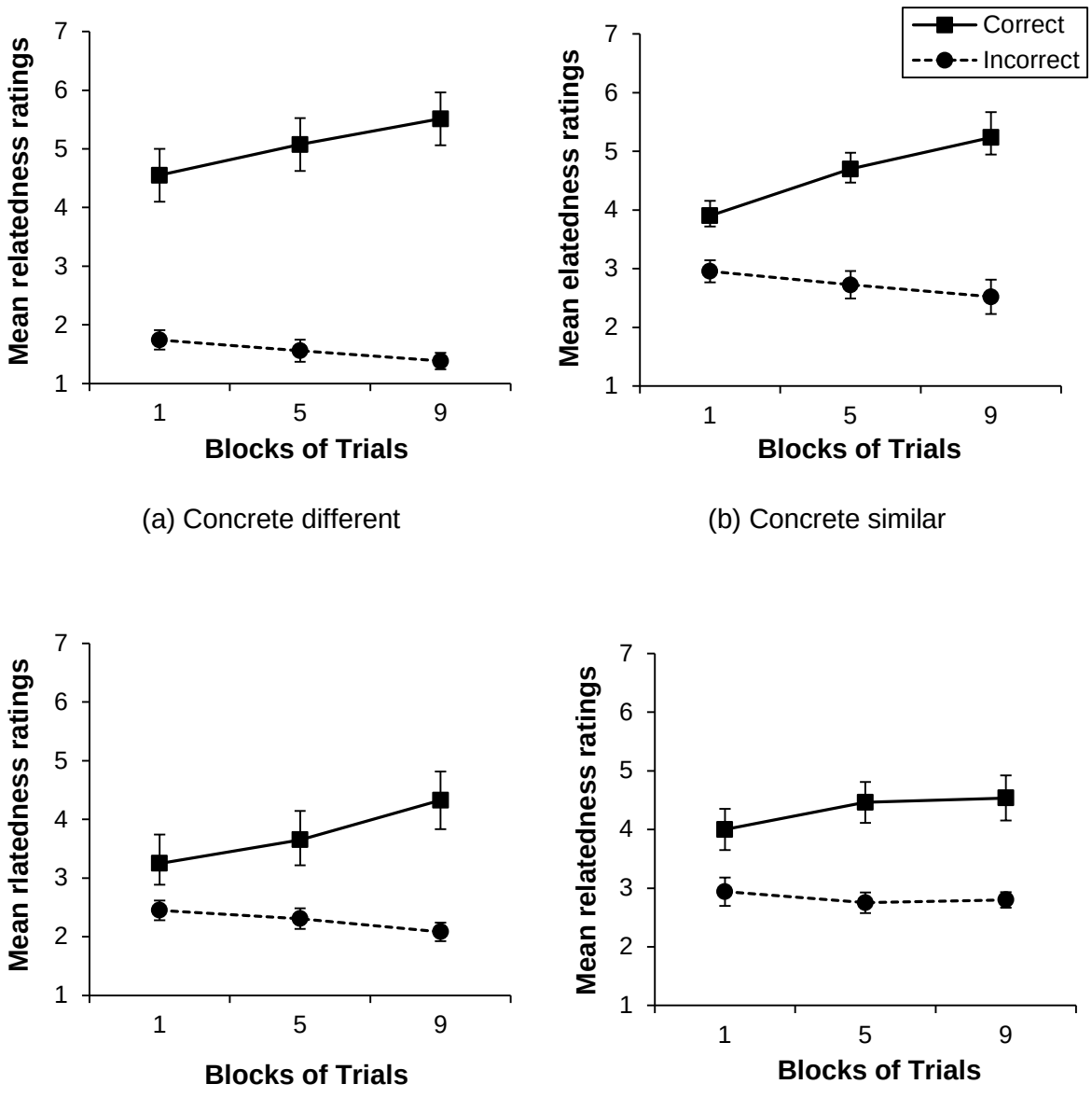

Figure 3. Relatedness ratings for correct and incorrect alarm signal-equipment combinations completed after learning trial blocks 1, 5, and 9. Error bars show the standard error of the mean.

accuracy for concrete different alarm signals in comparison to other alarm signal types with these differences reducing over blocks of trials. No other interactions were significant.

An identical GEE analysis was used to examine the percentage of second and third attempts made by participants with the offset set to 16 since counts of extra attempts were out of a possible maximum of 16. Abstract alarms, similar alarms and the first block of trials acted as the reference categories for the exponentiated coefficient. As expected, the main effects were significant:

(i) Alarm concreteness, $\chi^2(1) = 14.58$, $p < .001$, *95% CI* = -.21 - 3.66, *Exp(B)* = 5.60, with more additional attempts made for abstract alarms;

(ii) Alarm similarity, $\chi^2(1) = 12.70$, $p < .001$, *95% CI* = -.68 – 3.45, *Exp(B)* = 4.00, with more additional attempts for similar items;

(iii) Blocks of experimental trials, $\chi^2(1) = 26.27$, $p < .001$, with attempts reducing over blocks.

There was also a significant 3-way interaction between concreteness, similarity and blocks of trials, $\chi^2(1) = 28.02$, $p < .001$, mirroring the effects seen in response accuracy. In this analysis two of the 2-way interactions were also significant; concreteness x similarity, $\chi^2(1) = 5.15$, $p = .023$, and concreteness x block, $\chi^2(1) = 19.29$, $p = .013$.

As noted at the bottom of Table 4, participants could select the wrong alarms even after 3 attempts and so responses for 1st, 2nd and 3rd attempts do not sum to 100%. The mean percentage persistently incorrect after 3 attempts for each condition was 2.73%, 7.57%, 9.23% and 11.39% for the concrete different, concrete similar, abstract different, abstract similar conditions respectively.

Alarm Signal-Equipment Relatedness Ratings

Figure 3 shows the mean relatedness ratings for correct and incorrect alarm signal-equipment pairs obtained after learning blocks 1, 5 and 9. A 4-way mixed analysis of variance with correctness (correct vs incorrect) and blocks of trials (1 vs 5 vs 9) as within-subjects factors and concreteness (concrete vs abstract) and sound similarity (different vs

similar) as between-subjects factors was carried out to examine relatedness ratings further. Data were normally distributed in each condition (all ps > .05 using Shapiro-Wilks testing) and Mauchly's tests of sphericity were not significant.

Relatedness ratings for correct pairings were higher than for incorrect ratings, M(correct) =4.43, SD = 1.19, M(incorrect) = 2.35, SD = 0.73, ratings increased across blocks of trials, M(Block 1) = 3.22, SD = 0.76, M(Block 5) = 3.40, SD = 0.70, M(Block 9) = 3.55, SD = 0.68, and were higher for alarms in sets where they sounded different, M = 3.63, SD = 0.56, and lower when alarms sounded similar, M = 3.16, SD = 0.67 (see Table 5). Figure 3 shows a similar pattern to Experiment 1. Ratings for correct and incorrect items became increasingly different as participants learned across blocks of trials (see Correctness x Block interaction, Table 5). Correct and incorrect ratings differed more when items where concrete than when they were abstract (see Correctness x Concreteness interaction; c.f. Figure 3(a) and (b) with (c) and (d)). A 3-way Correctness x Concreteness x Similarity interaction resulted from the fact that while correct and incorrect ratings were more differentiated for concrete alarms, this was particularly true for those where alarms also sounded different from one another (c.f. Figure 3(a) and (b)). The 4-way interaction was not significant.

Table 6 reveals the extent to which participants understood, or were confused, about the piece of equipment associated with each alarm signal. Participants presented with concrete different alarm signals were able to differentiate correct and incorrect alarm-equipment relationships most effectively (see Difference correct/incorrect ratings), were less likely to be inaccurate (% Incorrect) and were less likely to select a variety of alternative meanings (No. of other meanings selected) in comparison to other alarm sets. As in Experiment 1, this data makes it possible to identify weaker alarms within the set, such as the perfusion alarm signal in concrete different alarm signal set. By the same token, it is evident that the general alarm signal in the abstract similar alarm signal set, is particularly effective because it is less similar than other items within the set by virtue of having a fixed, rather than a variable, pitch pattern (see Appendix 1).

Table 5. Relatedness ratings of alarm-equipment pairs: Summary of significant main effects and interactions from 4-way analyses of variance examining the effects of correctness (correct vs incorrect alarm-meaning relationships), blocks of trials (1, 5, 9), alarm signal concreteness (concrete vs abstract), and alarm signal similarity (different vs similar).

	Df	F	p	η^2
Main effects*				
Correctness	1,36	106.50	<.001	.747
Blocks of trials	1,72	8.09	.001	.184
Similarity	1,36	5.62	.023	.135
Interactions				
Correctness x Block	1,72	25.20	<.001	.412
Correctness x Concreteness	1,36	8.83	.005	.197
Correctness x Concreteness x Similarity	1,72	4.16	.049	.103

*Where effects are not significant they have not been included in this table.

Table 6. Mean difference between relatedness ratings for correct and incorrect alarm signal-meaning pairs obtained after learning blocks 1, 5 and 9, percentage of times incorrect meanings were assigned given all possible attempts, and the number of other meanings selected for each alarm signal.

	Block 1			Block 5			Block 9		
Alarm Signal	Difference correct/ incorrect ratings	% Incorrect	No. of other meanings selected	Difference correct/ incorrect ratings	% Incorrect	No. of other meanings selected	Difference correct/ incorrect ratings	% Incorrect	No. of other meanings selected
				Concrete Different					
General	3.37	0.00	0	3.70	0.00	0	4.83	0.00	0
Cardiovascular	2.17	62.50	5	2.11	16.67	1	3.66	0.00	0
Drug Admin	2.77	0.00	0	3.63	0.00	1	4.47	0.00	0
Oxygenation	2.89	37.50	2	3.44	0.00	0	3.50	25.00	3
Perfusion	4.16	9.09	1	4.79	0.00	0	5.09	0.00	0
Power Supply	2.30	30.77	3	3.47	16.67	2	3.46	0.00	0
Temperature	1.94	42.86	4	4.23	9.09	1	4.47	25.00	3
Ventilation	2.86	30.77	3	2.76	16.67	2	3.56	9.09	1
Alarm Signal Set Mean	*2.81*			*3.52*			*4.13*		
				Concrete Similar					
General	6.10	61.11	6	2.70	35.71	4	2.50	40.00	5
Cardiovascular	3.60	78.57	6	1.16	68.42	5	2.14	42.86	4
Drug Admin	3.90	43.75	5	2.40	23.08	3	3.63	25.00	3
Oxygenation	3.50	73.91	6	1.63	94.12	5	2.57	28.57	3
Perfusion	4.20	50.00	5	2.27	43.75	4	3.30	9.09	1
Power Supply	3.60	78.26	6	1.37	52.94	5	1.24	9.09	1

Temperature	3.70	68.42	7	2.14	9.09	1	32.74	70.00	3
Ventilation	3.40	40.00	4	2.13	50.00	5	3.60	0.00	0
Alarm Signal Set Mean	*4.00*			*1.98*			*6.46*		
				Abstract Different					
General	0.46	57.14	6	2.83	43.75	5	3.17	35.71	5
Cardiovascular	0.43	80.64	7	1.44	30.77	4	2.71	16.67	2
Drug Admin	2.69	25.00	3	2.94	28.57	4	4.17	28.57	4
Oxygenation	0.94	73.91	6	0.16	94.12	6	1.17	28.57	3
Perfusion	1.74	37.50	4	2.27	30.77	4	3.44	0.00	0
Power Supply	0.66	50.00	4	1.64	46.67	6	3.09	23.08	2
Temperature	1.14	55.56	5	1.07	63.16	4	2.24	30.77	3
Ventilation	-0.40	72.73	7	0.90	52.04	5	2.29	55.56	6
Alarm Signal Set Mean	*0.96*			*1.66*			*2.79*		
				Abstract Similar					
General	3.93	46.67	5	5.17	0.00	0	4.90	9.09	1
Cardiovascular	0.69	85.29	6	0.56	75.00	5	1.67	60.00	4
Drug Admin	0.41	68.18	6	0.90	71.43	6	0.77	33.33	2
Oxygenation	0.56	80.77	6	0.79	100.00	4	1.31	37.50	4
Perfusion	1.36	55.56	5	2.54	50.00	5	2.34	9.09	1
Power Supply	0.51	83.33	7	1.80	33.33	3	0.73	43.75	6
Temperature	0.76	66.67	6	1.37	73.91	6	2.04	57.14	5
Ventilation	0.27	61.90	7	0.56	68.18	6	0.13	59.09	6
Alarm Signal Set Mean	*1.06*			*4.37*			*1.74*		

Number of Fixations Prior to Fixating on Correct Interest Area (Fixation Index)

The fixation index was obtained for (a) participants' first attempt and (b) participants' second and third attempts (see Figure 4). Because participants rarely required second or third attempts when given different concrete alarm signals, the data available was insufficient for reliable analysis of the first fixation index (see Figure 4(a)). This meant that it was not possible to consider concreteness and alarm signal similarity as factors in an omnibus analysis. Instead, a Generalized Estimated Equation was used to conduct a 3-way analysis with alarm type (concrete similar vs abstract different vs abstract similar) as a between-subjects factor and blocks of trials (1-9) and number of attempts (first vs second vs third attempts) as within-subjects factors. A subsidiary GEE analysis was carried out to consider the effect of blocks of trials (1-9) on the fixation index when participants were making their first attempt in the concrete different condition. This type of analysis was employed because of the nature of the data: fixation count data was obtained repeatedly over a series of time points (blocks of trials). Prior to the analysis two extreme values were replaced using Winsorization with the 95th percentile value used for replacement (Chambers, Kokic, Smith & Cruddas, 2001; Winer, 1971). The data were transformed using a square root transformation to remove positive skew (Winer, 1971): a linear distribution was assumed, combined with an autoregressive correlation matrix since scores in blocks of trials could not be assumed to be independent.

While there was no overall difference in the number of fixations across conditions, $\chi^2(2)=2.64$, $p=.267$ as hypothesized the number of fixations was lower when participants identified the correct alarm signal-equipment mapping on their first attempt rather than on their second or third attempt, $\chi^2(1)=134.50$, $p<.001$, *95% CI* = -1.25 – -0.37, *Exp(B)* = 0.47. The main effect of blocks of trials was also significant suggesting that participants were making fewer fixations as they learned alarm-equipment pairings over blocks of trials, $\chi^2(8)=89.54$, $p<.001$. The reduction in the number of fixations was largely the result of fewer fixations being required on second and third attempts (see Figure 4) and this resulted in a

significant attempts by block interaction, $\chi^2(8)=22.96$, $p<.001$, and a 3-way attempts x alarm type x blocks interaction, $\chi^2(16)=45.72$, $p<.001$. Comparison of the line graphs for each condition in Figure 4(b-d) suggests that participants presented with concrete similar alarms are still very uncertain in initial blocks of trials (1-4) making considerably more fixations prior to attending to the appropriate equipment; in later trials participants appear to become much more certain, making fewer fixation attempts. If participants knew the correct piece of equipment to check when the alarm signaled, it took on average between 5-10 fixations to arrive at the correct interest area. Where participants were still taking 2-3 attempts to do so, the number of fixations was greater. In the concrete similar condition, the number of fixations reduced as they learned alarm signal-equipment pairings across experimental trials with the number of fixations being comparable for first and later attempts in later blocks of trials. However, differences still remain even in later blocks of trials for the abstract different and abstract similar conditions.

To summarize, the effects of alarm concreteness and similarity on participants' learning of alarm signal-equipment mappings were comparable to those observed in Experiment 1. The way in which alarms could direct, or misdirect, visual attention was also evident. Where alarm signals were difficult to distinguish from one another and the equipment was also in close proximity in the operating theatre scene (oxygen, temperature, cardiovascular and ventilation checks), then most confusion resulted. This was particularly evident in early learning trials for those using concrete similar alarms and across all learning trials for those using abstract similar alarms. Where participants were uncertain about which medical device the alarm signal related to, the number of fixations made prior to fixating on the correct piece of equipment was significantly higher (see Figure 4). This suggests that their scan paths were significantly more complicated prior to alighting on the appropriate interest area. Again, those using concrete different alarms appeared to have a particular advantage, while those using abstract similar alarms found learning most difficult.

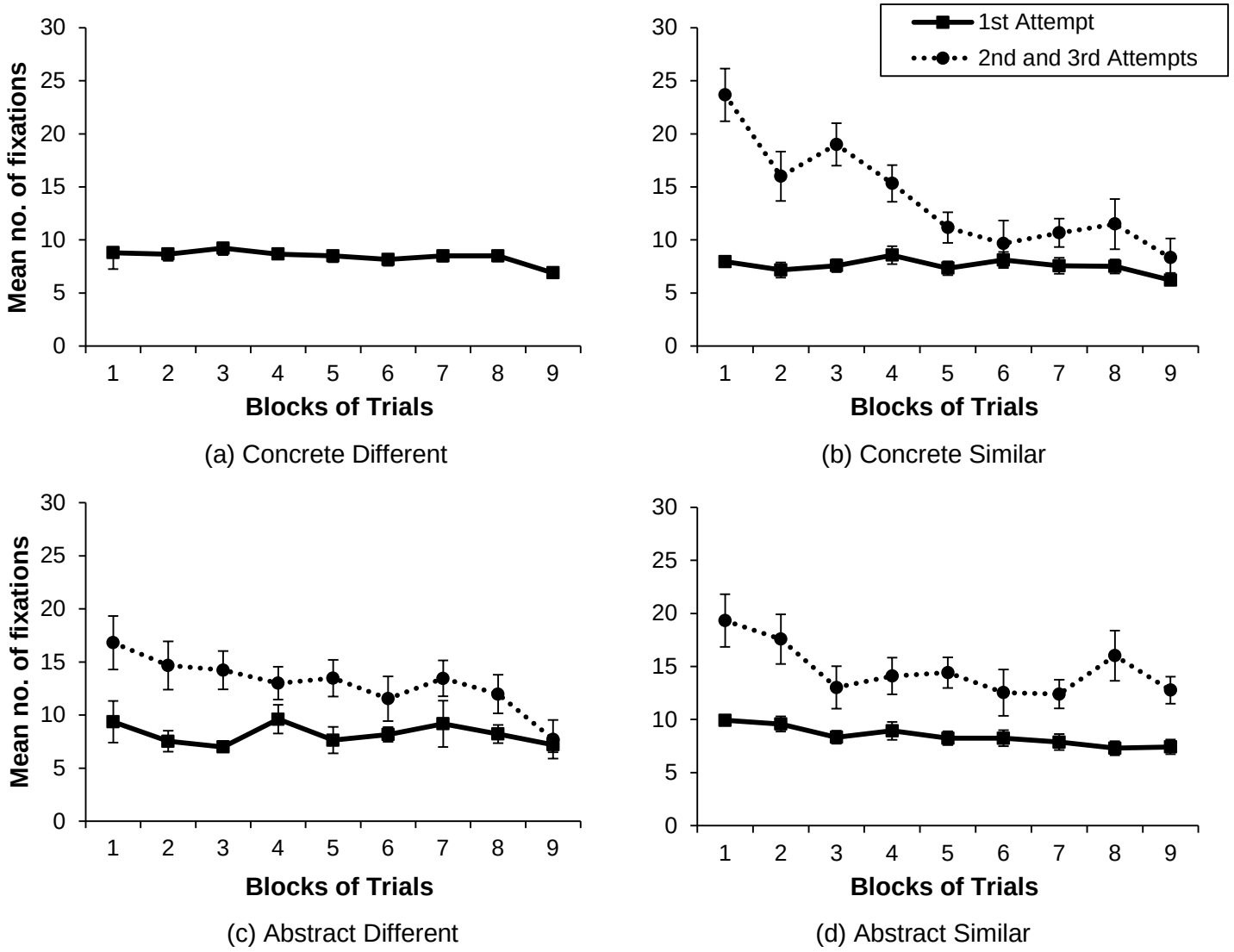

Figure 4. Number of fixations made prior to fixating on the correct interest area in each experimental condition. The number of fixations required when participants were correct on their first attempt or correct on their second or third attempt is shown except in the Concrete Different condition because so few extra attempts were made.

General Discussion

The alarm signals specified in the current edition of an international medical device safety standard, IEC-60601-1-8, suffer from problems which are common to other clinical alarm signals currently in use. The meaning of the alarm signals are difficult to learn and the alarm signals are easily confused (Lacherez et al., 2007; Sanderson et al., 2006; Williams & Beatty, 2005). The experiments reported in this paper compared the alarm signals from the current international standard with alternative alarm sets that differed in the extent to which the sounds used related to environmental sounds (e.g. the sound of a heartbeat to indicate 'check cardiovascular function') or were made up of a series of simple tones (concrete vs abstract alarms) and the extent to which they sounded similar to others in the alarm set (similar vs different alarm signals). The aim was to examine the extent to which these factors determined participants' ability to learn alarm signal-referent associations and to build an effective cognitive framework of the sounds in the alarm signal set allowing them to attend to medical devices appropriately.

As hypothesized, in both experiments concrete alarm signals were learned more quickly than abstract alarm signals and similar-sounding alarms took longer to learn than those that were dissimilar and easily discriminable. When alarms in a set were both concrete and dissimilar, this appeared to confer a particular advantage. In this set, the meaning of many alarms were correctly ascertained from the outset (see Figures 1(a) and 3(a)) with few confusions and fewer fixations were required before fixating on the appropriate equipment (see Figure 4). Conversely, those learning the alarm signals designed using the existing international standard (the abstract similar alarm signals) appeared to suffer a double setback in their learning. With these, participants took longer to learn them and were very confused about their meaning. It is possible to argue that these differences may gradually disappear as users gain expertise with the alarm set but differences remained after 16 blocks of learning trials (see

Table 1 and Figure 1) and the evidence to date suggests that difficulties in identifying and interpreting the alarms remain (Sanderson et al., 2006; Williams & Beatty, 2005).

Several studies have demonstrated considerable advantages for concrete sounds over abstract tones as alarm signals (Belz, Robinson & Casali, 1999; Bonebright & Nees, 2007; Bussemakers & de Haan, 2000; Edworthy et al., 2017; Fagerlonn & Alm, 2010; Graham, 1999; Isherwood & McKeown, 2017; Leung, Smith, Parker & Martin, 1997; McKeown & Isherwood, 2007; McKeown, Isherwood & Conway, 2010; Stevens et al., 2009; note that concrete alarm signals are sometimes referred to as auditory icons in these studies). However, it is clear from the studies presented here that the advantage of concrete sounds over abstract tones is not just a function of the easier mapping of signal-referent relationships found for concrete alarms, but also of the degree of acoustic variability within the set of sounds. This is an important finding and one which has both experimental and practical application. In experimentation, our finding suggests that in any future comparisons of the learnability of contrasting alarm signal sets, care should be taken to control for, or systematically manipulate, the amount of acoustic variability within sets so that the effects of this variability is either eliminated or known. From a practical viewpoint, it is important to understand the cumulative nature of these effects when designing new alarm sets. A set of concrete alarm signals which all sound fairly similar to one another will *not* perform better than a set of abstract signals with considerable acoustic variation. By the same token, the learnability of a set of abstract alarms can be improved by increasing the acoustic variability of the set.

Why are alarm concreteness and acoustic similarity so important in determining the interpretation and learning of alarm signal sets? Obtaining ratings of the relatedness of all possible alarm signal-meaning pairs, both correct and incorrect, helps to explain these findings. When participants were able to use their prior knowledge of sounds they had encountered previously (because the sounds are real-world, everyday sounds) they were able to make effective semantic mappings between the alarm signal and its meaning (c.f. Figure 1(b) and 1(d)

and Figure 3(b) and (d); Frank, Rudy, Levy & O'Reilly, 2005; Greene, Spellman, Dusek, Eichenbaum & Levy, 2001; Smith & Squire, 2005; Kumaran, 2013). This mapping using prior knowledge in order to facilitate understanding is well known in the literature on recall and prose comprehension (Bransford & Johnson, 1972; Kumaran et al., 2009; Mandler, 1984; McClelland et al., 1995; Schank & Abelson, 1977; Tse et al., 2007) and there is also existing evidence that this is the case for auditory alarms and information sounds (Gaver, 1989; Isherwood & McKeown, 2017; Stephan et al., 2006). This ability to draw on rich mental models of the real world in order to intuit possible meanings in new stimuli is one that has been utilized widely in designing visual icons for some considerable time (Smith, Irby, Kimball, Verplank & Harslem, 1982) and relies not merely on naming recognizing objects visually or auditorily, but on drawing upon the rich real-world understanding, as well as the axioms and heuristics we naturally have relating to those things. There is also growing evidence from the neuroscientific literature that environmental sounds are interpreted utilizing brain areas designed to solicit initial sound identification and interpretation in a way that is much less likely to be possible with the collections of tones used in abstract alarm signals (Griffiths & Warren, 2002; Lewis et al., 2004; Sharda & Singh, 2012; Tomasino et al., 2015).

Efficacy of semantic mapping from alarm signal to referent was also aided when the alarm signals sounded different to one another in the set and could be easily distinguished (c.f. Figure 1(a) and 1(c)). Given the musicality, or musical training, needed to distinguish alarm signals in sets employing the current alarm standard (Sanderson et al., 2006; Wee & Sanderson, 2008), it is not surprising that if sounds cannot be distinguished from one another initially then they are unlikely to result in effective mapping between the sound and its meaning. This is particularly likely to be compounded in everyday environments where the alarm signal is simply one of many others that must be interpreted. What is remarkable is that so little attention has been given to date to considering the multiple mappings created *across and between* sounds in alarm signal sets to allow users to create more effective mental models.

Although research examining the contingency of eye movements visually in hospital theatres and other settings is growing rapidly (e.g., Kogkas, Darzi & Mylonas, 2017; Kodappully et al., 2016; Koh, Park, Wickens, Ong & Chia, 2011; Marquard et al.., 2011; Schulz-Stubner, Jungk, Kunitz & Rossaint, 2002), little attention has been given to the factors affecting how attention might best be directed to the equipment associated with alarm signals (but see Stevenson et al., 2013). What little evidence there is suggests that participants' mental framework of the alarm set combined with their mental model of the context in which the alarms are used guide eye tracking and predict performance when using alarms (Burtscher, Kolbe, Wacker & Manser, 2011). In Experiment 2 the efficacy of participants' scan paths was examined as they learned the acoustic-visual mappings between alarm signals and medical devices. The way in which alarm signals could direct, or misdirect, visual attention was also evident. When alarm signals were difficult to distinguish from one another and the equipment was also in close *visual proximity* in the operating theatre scene (oxygen, temperature, cardiovascular and ventilation checks), then most confusion of meaning resulted (see Tables 3 and 6), and the number of fixation prior to fixating on the appropriate piece of equipment increased. This suggests that mental frameworks, if effective, were driving participants' scanning behavior. While a similar result might have been obtained for eye-movement data using the screen set-up in Experiment 1 (where we might have looked at the scanning patterns in relation to the words on the screen rather than the equipment used in Experiment 2) the finding that the data was conflated as a function of the physical proximity of the items of equipment provides an additional point of interest in our data. In Experiment 1 the words were evenly spaced across the screen so would not have been conducive to such a finding.

One of the key findings in this study is that performance across the two studies was very similar, despite the somewhat different paradigms. Of particular note is that in Experiment 2 participants were only given the names of the functions of the alarm signals at the beginning, and at the end of the fifth and ninth trial. This reduces the likelihood that the results of

Experiment 1 and other experiments in this program (e.g. Edworthy et al., 2017) are the result of purely semantic mapping between the auditory icons and the words which represent them (for example the cardiovascular alarm signal to the word ‘cardiovascular’). It shows that alarm signals which are easier to learn and distinguish make it easier to locate objects in 2D space. The additional cues available in these richly harmonic sounds also mean that they are easier to locate in 3D space (Edworthy et al., 2017).

Practical Application

Previous studies have demonstrated that the alarm signals being developed for the update of IEC 60601-1-8 are easier to learn and to localize when compared with most other alarm signals (for learning) and less harmonically rich alarm signals (Edworthy et al, 2017; Edworthy, Reid, Peel, Lock, Newbury, Foster, & Farrington, 2018). The study extends and reinforces the superiority of these alarms in two ways:

(i) By demonstrating that their efficacy is a consequence of both the concreteness of the individual alarm signals and the acoustic variability of the set as demonstrated by a factorial study, demonstrating that both factors are important and contribute to the known better performance of auditory icons under these circumstances.

(ii) Eye-movement data correlate with ease of learning, showing that alarm signals which are easier to learn also direct eye movements more directly in 2D space. While the experiment did not include localizability in 3D space, we already know that the proposed update alarm signals are easier to localize in 3D space because they are harmonically rich. Combining these two findings implies that the auditory icons are easier to localize for two different, likely additive, reasons: the sounds themselves direct the ears more accurately, and the easy learnability of the sound makes it easier to locate the sound through prior knowledge of its meaning.

Petocz et al. (2008), on the basis of their analysis of auditory warning and alarm-meaning relations, recommended a procedure for designing and developing auditory warning systems

(see Figure 2, p.175). Briefly, they outline the need to specify the auditory and visual characteristics of the operating system environment, to identify all warning alarms/messages that are required and order them in terms of importance. Once this has been done a series of potential alarms can be designed or selected for each message with the best being selected from each of these possibilities by gathering data from potential users of the system regarding the associations they make to the alarms select along with relatedness ratings of alarm-meaning (signal-referent strength).

Part of the process of developing and testing the proposed new alarms is to generate performance metrics on learnabiity, localizability, audibility and other metrics which might be useful for those charged with developing or testing their own audible alarm systems, both for understanding what might be achievable in terms of alarm signal design and also to make comparison possible between proprietary alarms and those developed as part of this project. These metrics are based on the findings of the series of studies and publications which have been part of this project (Edworthy et al 2017: Edworthy et al, 2018; McNeer, Bodzin Horn, Bennett, Edworthy & Dudaryk, 2018).

In all likelihood, no alarm signal set will be fixed and unchanging, there will always be change and development and therefore scope for new confusions to arise between alarm signals. We therefore recommend the following as a practical way of assessing the efficacy of individual alarms within a set and the extent and nature of confusions arising within a set. Alarm signal designers should ensure that:-

(a) Sounds are easily discriminable from one another. Careful consideration should be given to how similarity and difference are achieved within a set. There will be some situations which require alarm signals to sound similar, and others where they should sound different, and this will influence the learnability of the whole set.

(b) Sounds utilize real world or other existing metaphors known to the users (i.e. are concrete) to allow the inference of meaning.

Once a potential set of alarms has been designed, they should be evaluated by:

(c) Examining how accurately potential users can identify *all* the potential alarm signal-meaning pairs within the set and asking users to rate their relatedness.

(d) The data collected can be used to identify hard-to-identify alarm signals and the extent and nature of confusion between items in the set (see Tables 3 and 6).

This procedure is not dissimilar to the comprehensibility testing used to establish international standards visual symbols and icons (see IEC 11581-10: 2010).

Limitations and Future Directions

Although it is clear that both clinical and non-clinical participants have difficulty learning alarms designed using the guidelines from the existing standard (c.f. Lacherez et al., 2007, with Sanderson et al., 2006, and Williams & Beatty, 2005), it would be useful to replicate our findings using clinical populations. In a similar vein, Experiment 2 would ideally have been conducted in an operating theatre with both trainee and experienced anesthetists since they have primary responsibility for monitoring and dealing with alarms in theatre (along subsequent eye tracking of those with other theatre roles such as nurses and surgeons). Such an experiment would include more careful consideration of the alarm signal-equipment mappings to ensure that they were accurate (rather than simply plausible as was the case here) as well as consideration of the likely timeline during which alarms might be most likely for different surgical procedures. Aside from the ethical issues of conducting such a study in the UK, the mobile eye tracking systems currently in use are only now gaining sufficiently reliable resolution to examine fixations on items which are in close proximity (i.e., for the alarm signals associated with checks on temperature, oxygen, cardiovascular function and ventilation in Figure 2). This was important given that spatial proximity and well as acoustic similarity appeared to be important in determining confusability between alarms. In addition, the interplay between spatial proximity and acoustic similarity is only suggested by our data and needs further investigation.

Our aim was not to examine alarm fatigue but there is clearly further work to be done in examining the use of novel and distinguishable alarm signals in context. There is a need to create appropriate 'sound landscapes', or soundscapes, that fit well with their context of use and with users' existing mental models. We would argue that alarm signal sets may be best thought of as 'sound events … sequences of closely grouped and temporally related environmental sounds that tell a story' (Marcell et al., 2007, p.561). Creating appropriate soundscapes using groups of, albeit, artificially created sounds will effectively inform users as events unfold. As everyday users of language and environmental sound it should be possible to create sets of sounds which may be as superficially disparate as the sounds of frogs croaking, a tent zipping, and yawning associated with night-time camping but, by being relevant, the sounds will allow us to understand the nature of the whole sound event and act upon it 'achieving the greatest possible cognitive effect for the smallest possible processing effort' (Sperber & Wilson, 1986).

Future work should consider theoretical development and testing which would bring together these cognitive and psycholinguistic approaches with others dealing with learning and attention allocation in human factors. The SEEV model may be a good candidate to create effective cross-talk between these two domains (Wickens, Hollands, Banbury & Parasuraman, 2016). Although it is currently used primarily as a model of dynamic visual attention, it relies on probabilistic use of mental models in a way not dissimilar to that envisaged by Sperber & Wilson (1986). In the SEEV model user expectancy is seen as an "accurate mental model of the statistical properties of the environment, acquired through experience" (Koh, Park, Wickens, Ong & Chai, 2011, p.235) and mental models also drive the perceived value of an event in a given situation. The present research suggests that effort involved in learning alarms of this nature is considerably less for concrete different alarm signals than for abstract similar alarm signals and that learning of appropriate semantic mappings builds rapidly as effective mental models are formed to enable interpretation of alarms. Furthermore, participants' fixation patterns in Experiment 2 appeared to be determined by the mental model they were able to

build; those learning concrete different alarm signals quickly developed more effective fixation patterns, allocating visual attention more effectively. Research is needed to examine the way in which mental models may change dynamically in complex contexts with multiple competing demands and this is where the SEEV model may prove particularly useful as a theoretical framework for future research.

Footnote

1. ‘Meaning’ is used here rather than ‘function’ to denote the importance of learning meaning in context rather than rote learning of alarm signal-function rote associations.

2. For the sake of brevity, only the omnibus statistics for the factor are given and the contrasts between all blocks are not listed.

References

Al-Moteri, M.O., Symmons, M., Plummer, V. & Cooper, S. (2017). Eye tracking to investigate cue processing in medical decision making: A scoping review. *Computers in Human Behavior, 66,* 52-66.

Asan, O. & Yang, Y. (2015). Using eye trackers for usability evaluation of health information technology: A systematic literature review. J*ournal of Medical Internet: Human Factors*, 2, e5.

Bartlett, F.C. (1932). *Remembering: A Study in Experimental and Social Psychology*. Cambridge, UK: Cambridge University Press.

Bellenkes, A. H., Wickens, C. D., & Kramer, A. F. (1997). Visual scanning and pilot expertise: The role of attentional flexibility and mental model development. *Aviation, Space and Environmental Medicine, 48*, 569–579.

Belz, S. M., Robinson, G. S., & Casali, J. G. (1999). A new class of auditory warning signals for complex systems: Auditory icons. *Human Factors, 41*, 608-618.

Bliss, J.P. & Acton, S.A. (2003). Alarm mistrust in automobiles: How collision alarm reliability affects driving. *Applied Ergonomics, 34*, 499-509.

Block, F.E. (2008). “For if the trumpet give an uncertain sound, who shall prepare himself to the battle?”(I Corinthians 14: 8, KJV). *Anesthesia & Analgesia, 106*, 357-359.

Block, F.E., Rouse, J.D., Hakala, M., Thompson, C.L. (2000). A proposed new set of alarm sounds which satisfy standards and rationale to encode source information. *Journal of Clinical Monitoring and Computing, 16*, 541– 546.

Bonebright, T.L. & Nees, M.A. (2007). Memory for auditory icons and earcons with localization cues. ICAD-421, *Proceedings of the International Conference on Auditory Display*, Montréal, Canada.

Bransford, J.D., & Johnson, M.K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and*

Verbal Behavior, 11, 717-726.

Burtscher, M.J., Kolbe, M., Wacker, J. & Manser, T. (2011). Interactions of team mental models and monitoring behaviors predict team performance in simulated anesthesia inductions. *Journal of Experimental Psychology: Applied, 17*, 257-269.

Bussemakers, M.A. & de Haan, A. (2000). When it sounds like a duck and it looks like a dog . Auditory icons vs earcons in multimedia environments. *Proceedings of the International Conference on Auditory Display, 184-189*.

Chambers, R., Kokic, P., Smith, P. & Cruddas, M. (2001). Winsorization for identifying and treating outliers in business surveys. Proceedings of the 2nd International Conference on Establishment Surveys (ICES III), pp. 717-726. American Statistics Association, Buffalo, NY, June 17-21.

Cvach, M. (2012). Monitor alarm fatigue: an integrative review. *Biomedical Instrumentation & Technology, 46*, 268-277.

Dehais, F.R., Causse, M., Vachon, F., Régis, N., Menant, E. & Tremblay, S. (2014). Failure to detect critical auditory alerts in the cockpit: Evidence for inattentional deafness. *Human Factors, 56*, 631-644.

Drew, B. J., Harris, P., Zègre-Hemsey, J. K., Mammone, T., Schindler, D., Salas-Boni, R. & Hu, X. (2014). Insights into the problem of alarm fatigue with physiologic monitor devices: a comprehensive observational study of consecutive intensive care unit patients. *PloSOne, 9*, e110274.

Edworthy J, Reid S, McDougall S, Edworthy J, Hall S, Bennett D, Khan J & Pye E. (2017). The Recognizability and Localizability of Auditory Alarms: Setting Global Medical Device Standards, *Human Factors,* 59, 1108-1127.

Edworthy, J., Reid, S., Peel, K., Lock, S., Williams, J., Newbury, C., Foster, J & Farrington, M. (2018). The impact of workload on the ability to localize audible alarms. *Applied Ergonomics*, *72*, 88-93.

Fagerlonn, J. & Alm, H. (2010). Auditory signs to support traffic awareness. *Institute of Engineering and Technology, 4*, 262-269.

Familant, M.F. & Detweiler, M.C. (1993). Iconic reference: Evolving perspectives and an organizing framework. *International Journal of Man-Machine Studies, 39*, 705-728.

Frank, M.J., Rudy, J.W., Levy, W.B., & O'Reilly, R.C. (2005). When logic fails: Implicit transitive inference in humans. *Memory & Cognition, 33*, 742-750.

Fritz, C.O., Morris, P.E. & Richler, J.J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, 141, 2-18.

Garson, G.D. (2013). *Generalized Linear Models and Generalized Estimating Equations.* Statistical Associates Blue Book Series. Asheboro, NC: Statistical Publishing Associates.

Gaver, W.W. (1989). The Sonic Finder: An interface using auditory icons. *Human-Computer Interaction, 4*, 67-94.

Gegenfurtner, A., Lehtinen, E. & Saljo, R. (2011). Expertise differences in comprehension of visualization: A meta-analysis of eye-tracing research in professional domains. *Educational Psychology Review, 23*, 523-552.

Graham, R. (1999). Use of auditory icons as emergency warnings: Evaluation within a vehicle collision avoidance application. *Ergonomics, 42*, 1233-1248.

Greene, A.J., Spellman, B.A., Dusek, J.A., Eichenbaum, H.B. & Levy, W.B. (2001). Relational learning with and without awareness: Transitive inference using nonverbal stimuli in humans. Memory & Cognition, 29, 893-902.

Griffiths, T.D. & Warren, J.D. (2002). The planum temporale as a computational hub. *Trends in Neuroscience, 25*, 348-353.

Hardin, J.W. & Hilbe, J.M. (2012). *Generalized estimating equations*. 2nd edition. Boca

Raton, FL: Chapman & Hall/CRC.

International Electrotechnical Commission. (2012). *IEC 60601-1-8: Medical electrical equipment–General requirements, tests and guidance for alarm systems in medical electrical equipment and medical electrical systems*. Geneva, Switzerland: IEC.

Isherwood, S.J. & McKeown, D. (2017). Semantic congruency of auditory warnings. *Ergonomics, 60*, 1014-1023.

Joint Commission (2013). Medical Device Alarm Safety in Hospitals. *Sentinel Event Alert, Issue 5*, April 8, 2013. Accessed from: https://www.jointcommission.org/assets/1/18/SEA50_alarms_4_5_13_FINAL1.PDF.

Kerr, J.H. (1985). Warning devices. *British Journal of Anaesthesia, 57*, 696 –708.

Kerr, J.H. & Hayes, B. (1983). An “alarming” situation in the intensive therapy unit. *Intensive Care Medicine, 9*, 103-104.

Kodappully, M., Srinivasan, B. & Srinivasan, R. (2016). Towards predicting human error: Eye gaze analysis for identification of cognitive steps performance by control room operators. Journal of Loss Prevention in the Process Industries, 42, 35-46.

Kogkas, A.A., Darzi, A. & Mylonas, G.P. (2017). Gaze-contingent perceptually enabled interactions in the operating theatre. *International Journal of Computer Assisted Radiology & Surgery, 12*, 1131-1140.

Koh, R.Y.I., Park T., Wickens, C.D., Ong, L.T. & Chia, S.N. (2011). Differences in attentional strategies by novice and experienced operating theatre scrub nurses. *Journal of Experimental Psychology: Applied, 17*, 233-246.

Kristensen, M.S., Edworthy, J. & Özcan, E. (2016). Alarm fatigue in the ward, *Sound Effects, 6*, 89-104.

Kumaran, D. (2013). Schema-driven facilitation of new hierarchy learning and in the transitive inference paradigm. *Learning & Memory, 20*, 388-394.

Kumaran, D., Summerfield, J.J., Hassabis, D. & Maguire, F.A. (2009). Tracking the emergence of conceptual knowledge during human decision making. *Neuron, 63*, 889-901.

Lacherez, P., Seah, E. L., & Sanderson, P. (2007). Overlapping melodic alarms are almost indiscriminable. *Human Factors, 49*, 637-645.

Leung, Y. K., Smith, S., Parker, S., & Martin, R. (1997). Learning and retention of auditory warnings. *Proceedings of the International Community for Auditory Display 1997*. http://www.icad.org/Proceedings/1997/LeungSmith1997.pdf

Lewis, J.W., Wightman, F.L., Brefcynski, J.A., Phinney, R.E, Binder, J.R. & DeYoe, E.A. (2004). Human brain regions involved in recognizing environmental sounds. *Cerebral Cortex, 14*, 1008-1021.

Liang, K.Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*, 13-22.

Mandler, G. (1984). *Mind and Body: Psychology of Emotion and Stress*. New York: Norton.

Marcell, M., Malatanos, M., Leary, C. & Comeaux, C. (2007). Identifying, rating, and remembering environmental sound events. *Behavioral Research Methods, 39*, 561-569.

Marcell, M., Borella, D., Green, M., Kerr, E. & Rogers, S. (2000). Confrontation naming of environmental sounds. *Journal of Clinical and Experimental Neuropsychology, 22*, 830-864.

Marquard, J.L., Henneman, P.L., He, Z. Jo, J., Fisher, D.L. & Henneman, E.A. (2011). *Journal of Experimental Psychology: Applied, 17*, 247-256.

McClelland, J.L., McNaughton, B.L. & O'Reilly, R.C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failure of connectionist models of learning and memory. *Psychology Review, 102*, 419-457.

McCullagh, P. & Nelder, J.A. (1989). *Generalized linear models.* Second edition. London: Chapter & Hall/CRC.

McDougall, S., Curry, M. & de Bruijn, O. (2001). The effects of visual information on users' mental models: An evaluation of pathfinder analysis as a measure of icon usability. *International Journal of Cognitive Ergonomics, 5*, 59-84.

McKeown, D. & Isherwood, S. (2007). Mapping candidate within-vehicle auditory displays to their references. *Human Factors, 49*, 417-428.

McKeown, D., Isherwood, S. & Conway, G. (2010). Auditory displays as occasion setters. *Human Factors, 52*, 54-62.

McNeer, R.R., Horn, D.B., Bennet, C.L. , Edworthy, J.P. & Dudaryk, R. (2018). Auditory icon alarms are more accurately and quickly identified than current standard melodic alarms in a simulated clinical setting. Anesthesiology, 129, 58-66.

Moffat, M., Siakaluk, P.D., Sidhu, D.M., Pexman, P.M. (2015). Situated conceptualization and semantic processing: Effects of emotional experience and context availability in semantic categorization and naming tasks. Psychonomic Bulletin & Review, 22, 408-419.

Petocz, A., Keller, P. E., & Stevens, C. J. (2008). Auditory warnings, signal-referent relations, and natural indicators: re-thinking theory and application. *Journal of Experimental Psychology: Applied, 14*, 165-178.

Sanderson, P. M., Wee, A., & Lacherez, P. (2006). Learnability and discriminability of melodic medical equipment alarms. *Anaesthesia, 61,* 142-147.

Schank, R.C. & Abelson, R.P. (1977). *Scripts, plans, goals, and understanding.* Hillsdale, N.J.: Lawrence Erlbaum.

Schriver, A.T., Morrow, D.G., Wickens, C.D. & Talleur, D.A. (2008). Expertise differences in attentional strategies related to pilot decision making. *Human Factors, 50*, 864-878.

Schulz-Stubner, S., Jungk, A., Kunitz, O. & Rossaint, R. (2002). Analysis of the

anesthesiologist's vigilance with an eye-tracking device. A pilot study for evaluation of the method under the conditions of a modern operating theatre. *Anaesthesist, 51*, 180-186.

Schulz, C.M., Schneider, E., Kohbecher, S., Hapfelmeier, A., Heuser, F., Wagner, K.J., Kochs, E.F. & Schneider, G. (2014). The influence of anaesthetics' experience on workload, performance and visual attention during simulated critical incidents. *Journal of Clinical Monitoring & Computing, 28*, 475-480.

Schwanenflugel, P.J., Harnishfeger, K.K. & Stowe, R.W. (1988). Context availability and lexical decision for abstract and concrete words. *Journal of Memory and Language, 27*, 499-520.

Sendelbach, S. & Funk, M. (2013). Alarm fatigue: A patient safety concern. *AACN Advanced Critical Care, 24*, 387-388.

Shafiro, V. (2008). Development of a large-item environmental sound test and the effect and the effects of short-term training with spectrally-degraded stimuli. *Ear and Hearing, 29*, 775-790.

Sharda, M. & Singh, N.C. (2012). Auditory perception of natural sound categories – an fMRI study. *Neuroscience, 214*, 49-56.

Sheridan, T.B. (1970). On how often the supervisor should sample. *IEEE Transactions on Systems Science and Cybernetics, 6*, 140-145.

Smith, C. & Squire, J.R. (2005). Declarative memory, awareness and transitive inference. *Journal of Neuroscience, 25*, 10138-10146.

Smith, D.C., Irby, C., Kimball, R., Verplank, B. & Harslme, E. (1982). Designing the Star User Interface. *BYTE*, April 1982, 242-282.

Sperber, D. & Wilson, D. (1986). Relevance: Communication and Cognition. Cambridge: Blackwell.

SR Research Ltd. (2015). SR Research Experiment Builder User Manual (Version

1.10.1630). Mississauga, Canada: SR Research Ltd.

Stanton, N. & Edworthy, J. (Eds., 1998). *Human factors in auditory warnings.* USA: Gower Technical.

Steelman-Allen, K., McCarley, J.S., Wickens, C.D., Sebok, A. & Bzostek, J. (2009). N-SEEV: A computational model of attention and noticing. *Human Factors and Ergonomics Society Annual Meeting Proceedings, 53*, 774-778.

Stephan, K. L., Smith, S. E., Martin, R. L., Parker, S. P., & McAnally, K. I. (2006). Learning and retention of associations between auditory icons and denotative referents: Implications for the design of auditory warnings. *Human Factors, 48*, 288-299.

Stevens, C., Brennan, D. & Petocz, A. & Howell, C. (2009). Designing informative warning signals: Effects of indicator type, modality, and task domain on recognition speed and accuracy. *Advances in Cognitive Psychology, 5*, 42-48.

Stevenson, R. A., Schlesinger, J. J., & Wallace, M. T. (2013). Effects of divided attention and operating room noise on perception of pulse oximeter pitch changes: A laboratory study. *The Journal of the American Society of Anesthesiologists, 118*, 376-381.

Tiersma, S.E., Peters, A.A.W., Mooij, H.A. & Fleuren, G.J. (2003). Visualising scanning patterns of pathologists in the grading of cervical intrapithelial neoplasia. *Journal of Clinical Pathology*, 56, 677-680.

Tomasino, B., Canderan, C., Marin, D., Maieron, M., Gremese, M., D'Agostini, S., Fabbro, F., & Skrap, M. (2015). Identifying environmental sounds: A multimodal mapping study. *Frontiers in Human Neuroscience, 9*, Article 567.

Tse, D., Langston, R.F., Kakeyama, M., Bethus, I., Spooner, P.A., Wood, E.R., Witter, M.P., & Morris, R.G. (2007). Schemas and memory consolidation. *Science, 316*, 76-82.

Wee, A.N., & Sanderson, P.M. (2008). Are melodic medical equipment alarms easily

learned? *Anesthesia & Analgesia, 106*, 501-508.

Welch, J. (2011). An evidence-based approach to reduce nuisance alarms and alarm fatigue. *Biomedical Instrumentation & Technology, 45*, 46-52.

Whalen, D. A., Covelle, P. M., Piepenbrink, J. C., Villanova, K. L., Cuneo, C. L., & Awtry, E. H. (2014). Novel approach to cardiac alarm management on telemetry units. *Journal of Cardiovascular Nursing, 29*, E13-E22.

Wickens, C.D., Goh, J., Helleberg, J., Horrey, W.J. & Talleur, D.A. (2003). Attentional models of multitask pilot performance using advanced display technology. *Human Factors, 45*, 360-380.

Wickens, C.D., Hollands, J.G., Banbury, S. & Parasuraman, R. (2016). *Engineering Psychology and Human Performance.* New York, Routledge.

Wickens, C.D. & McCarley, J.S. (2008). *Applied Attention Theory*. Boca Raton, FL: CRC Press.

Winer, B.J. (1971). *Statistical Principles in Experimental Design.* Tokyo: McGraw-Hill.

Williams, S. & Beatty, P.C.W. (2005). Measuring the performance of audible alarms for anaesthesia. *Physiological Measurement, 26*, 571–581.

Appendix 1: Characteristics of alarm signals in each set

Alarm Functions[2]	Nature of the Alarm Signals	
	Concrete Alarm Sets	
	(a) Concrete Different Alarms	(b) Concrete Similar Alarms
General	Very fast version of the general sound indicated in Appendix 1(d), the current general IEC alarm. The temporal spacing between the third and the fourth pulse is proportionately longer	Rushing water sound, similar to white noise
Power down	A fan slowing down from full action to off, dropping an octave in pitch	Two pulses of a sink plunger
Cardiovascular	'Lub-dup' effect generated by three two-pulse units of a fist knocking on a flexible, hard surface	Two pulses of pounding on water, akin to an object hitting water
Perfusion	Two bursts of bubbling water	Short bubbling sounds
Drug Administration	Rattling of a small pill box	Continual water dripping in a cave
Oxygenation	Four pulses of an aerosol on-off action	One single ocean wave, with a shaped onset and offset (so louder in the middle)
Ventilation	Single, slowed down breathing out action	Wind blowing through a tunnel, increasing in intensity

Temperature	Frying on a hob	The sound of a fire burning

	Abstract Alarm Sets	
	(c) Abstract Different Alarms	**(d) Abstract Similar Alarms**
General	7-pulse unit with a distinctive rhythm and pitch pattern. The middle three pulses played in quicker succession that the first two and last two. Rise in pitch in the middle of the burst	A burst of three regularly spaced pulses (each pulse ranging between 100ms – 300ms), followed by a burst of two regularly spaced pulses in the following pattern and all at the same pitch: c c c – c c
Power down	Five short pulses of a car horn sound, with a raised pitch on the fourth pulse	A burst of three regularly spaced pulses (each pulse ranging between 100ms – 300ms), followed by a burst of two regularly spaced pulses in the following pattern, with the first and fourth pulses an octave higher: C c c – C c
Cardiovascular	Frequency-modulated tone with pitch sweep (rising in pitch by a tone from start to end)	A burst of three regularly spaced pulses (each pulse ranging between 100ms – 300ms), followed by a burst of two regularly spaced pulses in the following pattern: c e g – g C
Perfusion	High pitched, very short pulse repeated twenty-eight times	A burst of three regularly spaced pulses (each pulse ranging between 100ms – 300ms), followed by a burst of two regularly spaced pulses in the following pattern: c f# c – c f#
Drug Administration	Low-pitched buzzer	A burst of three regularly spaced pulses (each pulse ranging between 100ms – 300ms), followed by a burst of two regularly spaced pulses in the following pattern: C d g – C d
Oxygenation	Single-pitched continuous tone	A burst of three regularly spaced pulses (each pulse ranging between 100ms – 300ms), followed by a burst of two regularly spaced pulses in the following pattern: C b a – g f
Ventilation	Three bursts of a low-pitched double-pulse unit	A burst of three regularly spaced pulses (each pulse ranging

		between 100ms – 300ms), followed by a burst of two regularly spaced pulses in the following pattern: c a f – a f
Temperature	Two pulses of a major triad chord	A burst of three regularly spaced pulses (each pulse ranging between 100ms – 300ms), followed by a burst of two regularly spaced pulses in the following pattern: C d e – f g