

On: 01 August 2013, At: 05:55

Publisher: Taylor & Francis

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Ergonomics

Publication details, including instructions for authors and subscription information:

<http://www.tandfonline.com/loi/terg20>

Learning medical alarms whilst performing other tasks

Judy Edworthy^a, Christina Meredith^b, Elizabeth Hellier^a & David Rose^a

^a School of Psychology, University of Plymouth, Drake Circus, Plymouth, PL4 8AA, UK

^b High Wycombe, Buckinghamshire, UK

Published online: 30 Jul 2013.

To cite this article: Ergonomics (2013): Learning medical alarms whilst performing other tasks, Ergonomics, DOI: 10.1080/00140139.2013.819448

To link to this article: <http://dx.doi.org/10.1080/00140139.2013.819448>

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at <http://www.tandfonline.com/page/terms-and-conditions>

Learning medical alarms whilst performing other tasks

Judy Edworthy^{a*}, Christina Meredith^b, Elizabeth Hellier^a and David Rose^a

^a*School of Psychology, University of Plymouth, Drake Circus, Plymouth PL4 8AA, UK;* ^b*High Wycombe, Buckinghamshire, UK*

(Received 19 March 2012; final version received 19 June 2013)

Two studies are reported which first observe, and then attempt to replicate, the cognitive demands of intensive care unit (ICU) activity whilst concurrently learning audible alarms. The first study, an observational study in an ICU ward, showed that the alarms are very frequent and co-occur with some activities more than others. The three most frequently observed activities observed in the ICU were drugs (calculation, preparation and administration), patient observation and talking. The cognitive demands of these activities were simulated in a second, laboratory-based experiment in which alarms were learned. The results showed that performance in the alarm task generally improved as participants were exposed to more repetitions of those alarms, but that performance decrements were observed in the secondary tasks, particularly when there were two or three of them. Some confusions between the alarms persisted to the end of the study despite prolonged exposure to the alarms, confusions which were likely caused by both acoustic and verbal labelling similarities.

Practitioner Summary: The cognitive demands of working in an ICU were observed and simulated whilst alarms were learned. Alarms should generally avoid sharing similar rhythmic (and other) characteristics. The simulation task described here could be used for testing alarm learning without requiring a clinical environment.

Keywords: alarms; task sharing; learning

Introduction

Significant research effort has been focused on the ability of clinical practitioners to recognise the meanings of the audible alarms to which they are typically exposed whilst working. The main areas of research here have focused on existing recognition problems *in situ*, the ease or difficulty experienced learning alarms which have been mandated through medical device standards and the effect of alarm design principles on the learnability and recognisability of the resulting alarms. Few studies have looked at people's ability to learn and recognise alarms either in real clinical environments or in laboratory settings where the work demands of such an environment are simulated. In this paper, we explore people's learning of alarms both as a single task and whilst performing tasks which mimic the cognitive demands of working in an intensive care unit (ICU).

Early evidence concerning clinicians' understanding of the meaning of alarms (Loeb et al. 1992; Momtahan, Tansley, and Hetu 1993) indicated that operating room alarms were badly recognised by those who need to work with them on a day-to-day basis. Nurses and doctors working in an operating room and recovery room recognised the meaning of fewer than half of the alarms used in those areas (Momtahan, Tansley, and Hetu 1993). In addition, the urgency of the alarm was poorly matched to the urgency of the situation which the alarms were designed to signal.

In terms of people's ability to learn mandated alarms, recent evidence suggests that these can be difficult to learn and retain. The alarms which accompany International Electrotechnical Commission (IEC) 60601-1-8 are a case in point. IEC (2006) standard 60601 is concerned with medical electrical equipment, and so has wide scope and impact. Parts 1–8 are concerned with general requirements for basic safety and essential performance, including the alarms used with that equipment. The set of alarms currently supporting IEC 60601-1-8 have been shown to be difficult to learn. For example, studies have shown that even after considerable exposure to the alarms associated with IEC 60601-1-8, recognition is low, with very few people able to discriminate between all of the alarms specified in the set (Lacherez, Seah, and Sanderson 2006; Sanderson, Wee, and Lacherez 2006; Williams and Beatty 2005). A study of nurses (Wee and Sanderson 2008) demonstrated that even after two training sessions in the meaning of the alarms, only one nurse (out of 22) managed to identify all of the alarms correctly. Nurses with a level of musical training performed better at learning the meanings of the alarms than did those without musical training and were also able to distinguish better between high- and medium-priority versions of the alarms.

Significant research effort has also been put into understanding the acoustic and other design features of alarms which render them difficult or easy to learn. In more general terms, considerable research evidence suggests that 'auditory icons',

*Corresponding author. Email: jedworthy@plymouth.ac.uk

which are typically (though not always) real-world sounds, are easier to learn than more abstract sounds (Belz, Robinson, and Casali 1999; Graham 1999; Keller and Stevens 2004; Leung et al. 1997; Perry et al. 2007; Ulfvengren 2003). The main reason for this appears to be that the relationship between the alarm sound and its referent (the situation/hazard it is signalling) is closer than is typically found for abstract alarms, making it easier for participants to recall the function once they have heard the alarm. Most studies show that speech is easy to learn and that abstract alarms (both traditional abstract alarms such as buzzers and bells and more current abstract alarms such as beeps and other tones) are much more difficult to learn.

Despite the knowledge that 'auditory icons' and sounds which bear an understandable relationship to their referents are easier to learn, alarms used in the medical world are typically abstract in nature and continue to be used. If abstract tones are used, and continue to be used (which is likely), then it is important that their design is optimised. Here, there are research data which demonstrate that there are particular acoustical features of alarms which will tend to cause confusion between them (and should, therefore, be avoided). One of the most important of these is that alarms with similar temporal patterns will tend to be confused with one another even when they differ along other dimensions (Edworthy and Meredith 1997; Patterson 1982). This includes continuous tones which differ considerably in pitch (Edworthy and Meredith 1997). Edworthy et al. (2011) have demonstrated how these two features can have a significant effect on confusions between a set of alarms, and how adjustments to those features can improve both the discriminability and learnability of that same set. This study also demonstrates the benefit of using one or two auditory icons within the set, for ease of learnability. The alarms specified in IEC 60601-1-8 share the same temporal pattern, and they are abstract in nature (they are melodies); these two factors (probably) largely account for the difficulty in learning them. Recent research has suggested that changes to the amplitude envelope of these alarms may improve their efficacy (Gillard and Shultz 2012).

Of course, when alarms are signalled in the real medical environment, there are many other factors in operation which may not necessitate high levels of learning, or may mediate that learning. If the main purpose of an alarm is to attract attention, then a signal which does nothing more than attract the clinician to the relevant piece of equipment may be all that is required. Here, the main requirement is that the alarm should be audible and localisable. Also, there are many other cues available to the clinician when actually dealing with a patient. For example, there will be visual cues both from the apparatus and the patient. Depending on the procedures being carried out, there will be a greater expectation that some pieces of equipment will be more likely to signal than others. There will be demographic details of the patient (such as age, general state of health, weight and so on) which will have indicated *a priori* the kinds of problems which might occur during medical procedures. There will be other professional and personal (possibly emotional) responses from the clinician when hearing alarms which might not arise in a non-clinical environment. These factors will influence both the extent to which the meanings of alarms need to be learnt and the degree to which the context of the alarms helps or hinders the process of learning the meanings of those alarms in an actual clinical context. However, studies which have made direct observations of alarms in the clinical context do not directly address the issue as to whether clinicians immediately recognise the meanings of alarms.

For example, Bitan et al. (2004) carried out an observational study of the relationship between alarms and the activities of nurses in a neonatal ICU, and demonstrated that though responses to alarms did not appear to be immediate, the probability of particular care activities was higher when some of the alarms sounded than others. Their analysis shows that the likelihood of a response increased up to 1 min after the alarm sounded (the longest time-lag analysed) and that actions were more likely if an alarm had sounded than if it had not. Bitan et al. interpret this as nurses' behaviour at any one point being placed on a continuum from purely respondent behaviour (reacting to the alarm) to purely self-initiated behaviour. The point on that continuum at any time is dependent on the current value of the system information. The speed with which clinicians respond to alarms is typically slow, and indeed introducing a delay into the sounding of alarms has been demonstrated to be helpful in reducing the noise pollution and distraction caused by unnecessary alarms (Gorges, Markewitz, and Westenskow 2009).

Seagull and Sanderson (2001) carried out an in-depth study of the way anaesthetists use alarms across a range of procedures and produced a number of findings which are useful for the future development of alarm systems. For example, they showed that the frequency of occurrence and type of alarm produced vary across different phases of medical procedures, responses to alarms differ across phases and that responses to the same alarm differ in different contexts. Both Bitan et al.'s (2004) and Seagull and Sanderson's (2001) studies demonstrate how alarms are used in the work environment and illustrate the complex relationship between alarms and related activities. These data, though extensive, do not necessarily address the question as to whether the clinicians knew, *a priori*, the meanings of the alarms to the extent that they would have been able to name them in the absence of other cues. In addition, these studies are concerned with responses to those alarms in a complex work environment where the whole workflow in relation to alarms is under investigation. Factors such as the likelihood and relevance of alarm occurrence in the procedure being carried out, localisation of the alarm and other more general issues are of key importance in these studies.

Another issue which impedes progress in understanding the ease or difficulty with which alarms are learnt is the extremely high false alarm rate which often occurs in the real work environment. Many studies highlight the fact that the vast majority of alarms which occur in clinical environments are false alarms, caused either because of artefacts (bad sensors, patients moving or being moved) or conservative settings of alarm limits, meaning that most alarms do not signify clinically significant events – events that are clinically adverse and require intervention (Chambrin et al. 1999; Imhoff and Kuhls 2006; Keller et al. 2011; Schmid et al. 2011; Siebig et al. 2010; Talley et al. 2011). The central problem here is that of alarm fatigue: alarms sound so often and are so frequently false that medical staff become immune and habituated to those alarms with the consequence that patient safety is compromised. The US Food and Drug Administration received 566 reports of patient deaths being related to alarms on monitoring devices between 2005 and 2008 (FDA 2011); the President of the Association for the Advancement of Medical Instrumentation (AAMI) has challenged the health-care community by issuing the mission that ‘By 2017, no patient will be harmed by adverse alarm events’ (Logan 2011, 1); and the Emergency Care Research Institute (ECRI 2010) has indicated that alarm hazards are number 1 in its top 10 hazards for 2012. All of this adds to the problem of developing a paradigm which looks directly at the ease or difficulty of learning alarms in a realistic (at least in cognitive terms) setting. Recent reviews (Borowski et al. 2011; Edworthy 2012) illustrate how the many problems associated with medical alarms are interwoven and still remain unresolved. New theoretical frameworks continue to be developed in this area (Raymer, Bergstrom, and Nyce 2012) and new, potentially important components of responses to auditory alarms continue to be revealed (Sakai et al. 2011).

Thus, we cannot necessarily look to observational studies of actual behaviour related to alarms in addressing the issue of how well or badly those alarms are learned or recognised. One approach which has hitherto not been taken is the one presented in this paper: Here, we first observe typical activities which take place in an ICU, then attempt to mimic those activities at a cognitive level whilst requiring participants to recognise and name alarms as a primary task. Thus, whilst we are not concerned with actual workflow around alarms in the real environment, we are concerned with people’s ability to learn and retain the meanings of abstract alarms when they are performing other tasks similar to those which would be carried out in a real environment.

Our approach was first to determine how alarms co-occur with other activities in the hospital ward, and then to simulate the cognitive demands of those tasks whilst participants were exposed to alarms. We began by carrying out an observational study of the incidence and types of alarms which signal in a typical ICU, and the types and number of tasks which co-occur with those alarms. We then developed a multiple task laboratory study which allows us to look at people’s ability to learn alarms whilst carrying out other activities mimicking those of the ICU environment.

Study 1

Observational studies focusing on alarms demonstrate that alarms sound very frequently and for long periods of time, and that they are usually not clinically relevant (Talley et al. 2011), in that the alarm has not signified a change in the patient’s status that requires clinical intervention. Some studies in this area have been concerned with the direct relationship between alarms and activities, and so have required techniques which allow the mapping of activities back to the alarms. For example, Seagull and Sanderson (2001) adopted a set of procedures which included process tracing, which required the keeping of an activity log; field notes and discussions with staff; and post-operative review and collection of other records related to the procedures observed. Bitan et al. (2004) observed nurses’ responses to neonates in an ICU by taking detailed, real-time measurements of the triggering of alarms from the neonates’ monitors and recorded whether or not nurses then attended the patient. However, these studies do not tell us directly whether or not clinicians were able to recognise the alarms when heard. As this is our focus, we adopted a different strategy.

In this first study, we were interested in the activities which occur and the alarms which sound in the ICU. Our aim was to chart both the activities and the alarms. Our intention was to build a picture of the types of activities which occur in the ICU and to chart the degree to which activities co-occur, in order to provide a basis for developing laboratory tasks which simulate those activities whilst learning alarms. Similarly, our aim in charting the alarms was to get a sense of how often alarms sounded, what type of alarms they were and the acoustic nature of those alarms. Thus, we were looking at activities and alarms as two discrete entities, in order to develop a paradigm for our second study.

Method

Preliminary work

Two multi-bed intensive care wards at Derriford Hospital, Plymouth, UK, were selected for the observations. These were visited at various times of the day by two of the authors (CM and DR) for a total of 16 h. The routines of nurses were recorded informally on a notepad and ranged from relatively quiet periods to periods of fairly intense activity. The nature of

the study was discussed with medical staff. The experimenters attended a training meeting to explain the study and to get feedback from staff. Other employees beyond the medical staff were spoken to and observed (e.g. receptionists, porters and physiotherapists) because they were sometimes either involved directly in the activities of the ICU or had insights into the kinds of activities occurring there. Nurses were asked questions about their work routines, and they were observed, so that an outline of day-to-day routine care could be ascertained. This helped to clarify the categories of behaviour that should be included in the checklists. The experimenters also learned the alarm sounds, as it was necessary to note specific alarm occurrence during the observation task.

On the basis of the procedures above, a 16-category checklist was developed. This checklist contained the behaviours that were directly observed, as well as those that were felt likely to occur based on discussion of ward activities and earlier drafts of the checklist itself. These categories include teaching, staff talking, talking to the patient, talking to visitors, staff handover, observation (including both formal and informal), drugs (calculation, preparation and administration, hereafter called 'drugs'), physiotherapy, doctor's visit, admission, discharge, X-ray, ECG, answering the telephone, responding to alarms and a category for all other events not already included (other).

A nine-category checklist for alarms and other frequently occurring sounds was also developed on the basis of the preliminary work. Here, the alarms and other sounds (particularly, pagers and telephones which can occur frequently, but are not strictly alarm sounds) considered and informally observed to be the most frequently sounding were listed separately, and those which occurred less frequently were assigned to an 'other' category. The alarms and other sounds listed were ventilator, syringe pump, pulse oximeter, infusion pump, ECG monitor, humidifier, telephone, pager and a single category for other alarms and auditory signals. The checklist and the procedure were tested during an informal 'run' to check that the observation task was doable and that the checklists were comprehensive.

Observations

All observations took place in a four-bedded ICU in a hospital in Plymouth, UK. An observation session of 30-min duration was devised. This 30-min period involved a 1-min lead-in, followed by fourteen 16-category checks on activity at 2-min intervals, followed by a 1-min finish time. The 2-min intervals were timed with a stopwatch. At every 2-min marker, both observers would check on all activities being carried out at each of the beds. The observers noted both the bed number (1–4) and the activity (or activities) taking place at that bed. For any activities which took more than 2 min, only the first occurrence of that activity was recorded. The observers also noted occurrences of alarms using the alarms checklist, as they occurred, noting each alarm and the time it occurred. Only the first occurrence of the alarm was noted, thus if an alarm was still sounding during a second or third 2-min interval, its continued occurrence was not further noted. If the alarm ceased and began again in a later 2-min period, this was counted as a new occurrence of the alarm.

In total, 5 h of observations were carried out in this way, over a period of several days. The observations were undertaken between 13.00 and 14.30 h in the daytime as the informal stage had suggested that this was a relatively busy time on the unit.

Results and discussion

Inter-observer reliability

The degree of agreement between the observers was calculated both via ϕ coefficients for individual categories of events (Siegal and Castellan 1988), which looks at individual present/not present recordings of activities for each occurrence, and overall percentage agreement on the number of occurrences of each event, a measure also often used and for which 80% overall agreement is deemed acceptable (Odam and Ogawa 1992). There were no occurrences of X-ray, ECG or other activities not on the checklist ('Other' on the checklist) during the observation sessions. Inter-observer reliability and percentage agreement measures for each of the activities observed ranged from 0.96 to 0.59 (ϕ) and between 99% and 87% (percentage agreement), and were as follows: teaching, 0.80 (94%); staff talking, 0.63 (87%); talking to the patient, 0.84 (94%); staff handover, 0.59 (96%); talking to visitors, 0.71 (96%); observation, 0.77 (89%); drugs, 0.90 (96%); physiotherapy, 0.95 (98%); doctor's visit, 0.78 (96%); admission, 0.96 (99%); discharge, 0.78 (97%); answering the telephone, 0.96 (99%) and responding to alarms, 0.90 (96%). Thus, all of the categories were above the generally accepted percentage level of 80%, though the ϕ coefficients were lower for some of the talking categories. The difference between the two measures suggests that the observers did note the occurrences of each of the events accurately overall, but sometimes did not agree about the 2-min slot in which that activity began. This would affect ϕ coefficients but not percentage accuracy measures. This is particularly true for the talking categories.

Inter-observer reliabilities and percentage agreement for the individual alarms were as follows: ventilator, 1 (100%); pulse oximeter, 0.86 (98%); syringe pump, 0.95 (99%); infusion pump, 0.88 (99%); ECG monitor, 1 (100%); humidifier, 1

(100%) and telephone, 0.96 (99%). No occurrences of pager or other alarms were observed during these measurements. Thus inter-observer reliability in terms of which alarms sounded was very high according to both measures. One would expect this, given that it is much easier to indicate the start of an alarm than to make a judgement about, for example, which kind of talking activity is occurring within a specific 2-min slot.

Frequency of activities and alarms

The observation data were collated in order to indicate the frequency of each of the activities and of the alarms sounding. Figure 1 shows the number of times each of the activities was recorded over the study. The top activities were observation (177 observations), staff talking (117 observations), responding to alarms (112 observations), drugs (106 observations) and talking to the patient (99 observations). These activities account for a large majority of the total activity (65% of all activities when taken together), and on many occasions two or more of these activities co-occurred.

Thus the data demonstrate that the four key activities in which staff were engaging when the ward was busy were: talking (the two categories of staff talking and talking to the patient account for 23% of all activity), drugs, observation and responding to alarms. In addition to the additional occurrence of talking to visitors, talking necessarily accompanied many of the other less frequently observed activities such as doctors' visits, answering the telephone and staff handover.

Figure 2 shows the total number of times each alarm occurred as a proportion of the total number of occurrences of all alarms recorded during the observations. Only six of the alarms in the ICU sounded during our observations (the telephone sounded as well, but here we present medical alarms specifically), and they typically sounded fairly often. Ventilators, pulse oximeters, syringe pumps and infusion pumps sounded quite regularly during the observations, and ECG monitors and humidifiers less so. Because an alarm, once on, could sound for an indefinite period, the amount of time that any of the alarms was on is unknown.

Concurrent tasks

Each 2-min observation was broken down by activities and alarms in order to see which alarms co-occurred with which, and with how many, activities. Table 1 shows the number of times each of the alarms co-occurred with each of the activities.

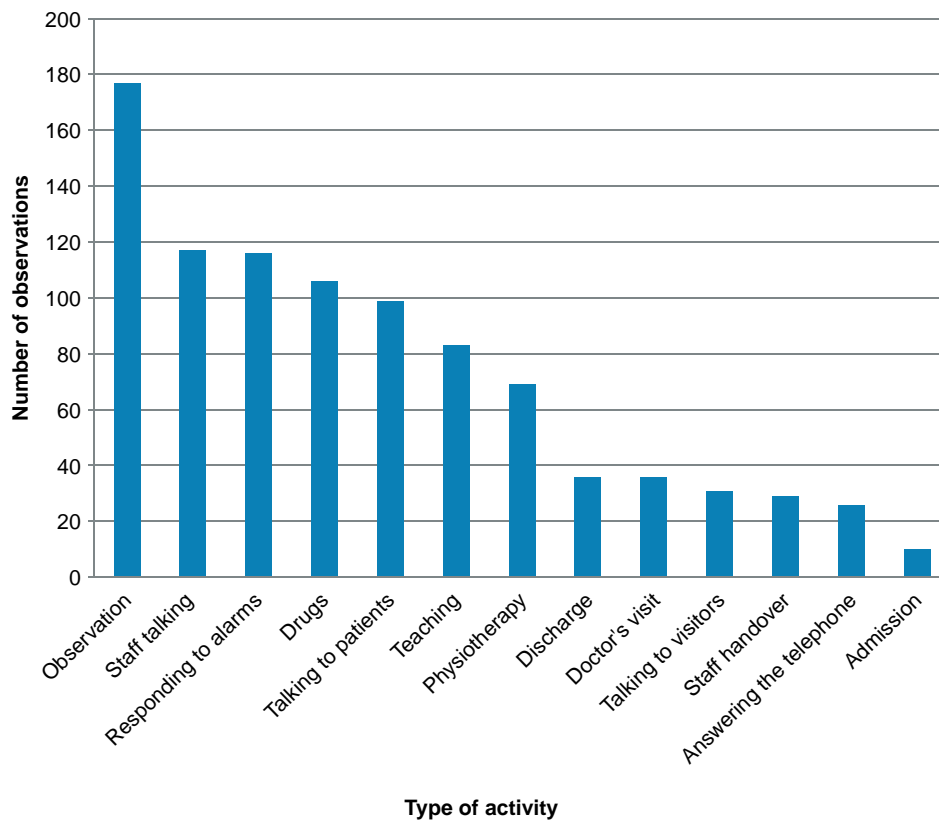
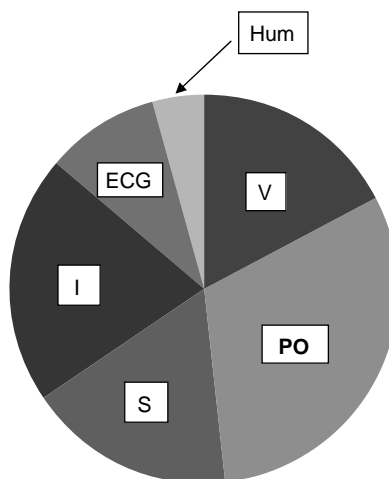


Figure 1. Number of occurrences of each of the activities observed, Study 1.



V–Ventilator (20 occurrences)
 PO–Pulse Oximeter (36 occurrences)
 S–Syringe Pump (20 occurrences)
 I–Infusion Pump (24 occurrences)
 ECG–ECG monitor (11 occurrences)
 Hum–Humidifier (5 occurrences)

Figure 2. Number of occurrences of each of the alarms observed, Study 1.

On some occasions the alarms were associated with more than one activity, which is why the total number of alarms indicated in Table 1 is higher than that shown in Figure 2; they have been counted every time they co-occurred with an activity, meaning that they have in some instances been double-, triple- or even quadruple-counted. Of the 112 occurrences of the alarms, 24 were associated with a single activity, 67 were associated with two activities, 17 were associated with three activities and 4 were associated with four activities.

The main purpose of Study 1 was to obtain data to allow some understanding of the kinds of activities which take place in the ICU, and how they co-occur. The next stage was to develop tasks which mirror those observed activities. In the study reported next, we also increased the workload in steps in order to assess the added burden of these extra tasks on people's ability to recognise and continue to learn alarms.

Table 1. Number of times each of the alarms co-occurred with each of the observed activities, Study 1.

Activity	V	PO	S	I	ECG	Hum	Total
None	0	0	0	0	1	0	1
Teaching	10	0	2	7	1	0	20
Staff talking	9	8	6	8	4	1	36
Talking to the patient	0	12	5	5	1	4	27
Staff handover	0	1	4	0	4	0	9
Talking to visitors	0	1	3	2	3	0	9
Observation	2	26	10	8	3	2	51
Drugs	4	2	16	19	3	0	44
Physiotherapy	14	0	1	7	0	5	27
Doctor's visit	0	1	3	3	1	0	8
Admission	0	0	0	0	0	0	0
Discharge	0	7	0	0	0	0	7
Answering the telephone	0	0	1	0	1	0	2
Total	39	58	51	59	22	12	241

Note: V, ventilator; PO, pulse oximeter; S, syringe pump; I, infusion pump; ECG, ECG monitor; Hum, humidifier.

Study 2

In this study, learning and naming alarms was presented as the primary task throughout, and secondary tasks were progressively introduced. These secondary tasks were designed to mirror, at a cognitive level, the main tasks we observed in Study 1. The main concurrent tasks we observed were 'observation', 'drugs' and talking (of various types). Alarm occurrences could be accompanied by one, two or all three of these activities. Our aim therefore was to broadly simulate the tasks we observed in the ICU with laboratory-based tasks for non-clinical participants so that they had both face validity and validity at a more conceptual and theoretical level. Wickens' 4D multiple resources model was used to provide a check that the tasks we developed made similar perceptual and cognitive demands on our participants to those that the ICU tasks we observed made on the clinical staff (Wickens 1984, 2002; Wickens and Liu 1988; Wickens, Sandry, and Vidulich 1983). Wickens' 4D multiple resources model (e.g. Wickens 2002) is typically used as a model for predicting the extent to which concurrent tasks will interfere with one another on the basis of the degree to which they share similar stages, modalities, visual channels and modes. However, the model can be used to make comparisons between the ICU activities we observed, and the tasks we developed in the laboratory.

At the heart of Wickens' 4D multiple resources model are four categorical elements, each divided into two discrete levels. These four categories underpin the extent to which time-sharing between tasks will be possible – the more those task share similar elements, the more difficult they will be to perform at the same time. Thus, it was important that our laboratory tasks embodied the same elements as the activities we observed in the ICU, so as to avoid confounding our simulation with inappropriate time-sharing problems.

The four dimensions in Wickens' model are (1) stages – here, the two stages are mental activities (perception and cognition) and responding, (2) modalities – here, the basic distinction is between auditory and visual modalities, (3) visual channels – this is a distinction between two types of visual processing, between focal and ambient vision, for which there is good evidence that separate cognitive structures are used; and (4) processing codes – this is the distinction between spatial and symbolic (usually verbal) codes. Each of our laboratory tasks simulated important mental aspects of the activities we observed in the ICU.

The primary task throughout was listening to alarms and then naming them. The secondary tasks we used were an observation task, a mental arithmetic task and a talking task. In the first instance, listening to alarms uses the same stage (mental activity rather than response), modality (auditory rather than visual) and processing code (verbal rather than spatial), which would be used when alarms are listened to in the ICU. The visual channel is barely engaged in this task. Our observation task uses a similar subgroup of the dimensions that observing patients would in the ICU; namely the same stage (again mental activity rather than response), modality (largely visual rather than auditory) and processing code (spatial rather than verbal, though there was a verbal element to our laboratory observation task). Our laboratory observation task also probably had a greater focus on focal visual processing than the ICU observations we were mimicking, a result of the smaller laboratory setting in which we carried out the task. Our mental arithmetic task, intended to mimic the activity of 'drugs', focused on the first of the three drugs activities, namely calculation, which was largely mental in nature. Here, the stage is again mental rather than response based, the modality is visual, the visual channel is focal and the code is verbal. This mirrors the sub-components of the multiple resource model that would be involved in carrying out drugs calculations in the ICU. Finally, we used a talking task in our laboratory, simulating the talking that frequently took place in the ICU, which by definition must use the same subset of the multiple resource model components as the talking that took place in the ICU.

We did not attempt to mirror responses in as exact a fashion, for two key reasons. The first was that whilst it was important to engage the same sub-components of the multiple resources model for our laboratory tasks as for the main observed activities in order to properly simulate both the activities and their interactions at a cognitive and perceptual level (one of the two stages in the model), the kinds of responses that could follow in the ICU were varied and sometimes hard to discern (and was also not the main focus of Study 1). There was no single response for each of the main activities we observed, but potentially several, making it difficult to mimic any meaningful response in our laboratory tasks. The other reason for not mirroring responses in as direct a way is that it is clear from multiple resource model research that the resources used for perceptual and cognitive activities are functionally separate from those involved in the selection and execution of responses.

Participants started the study by learning a set of alarms. They were then required to identify (and continue to learn) those alarms when being subject to one, two and then three secondary tasks. Secondary tasks were introduced gradually as the experiment progressed. Accordingly, workload increased gradually as the experiment progressed. This technique is similar to that of Bliss and Dunn (2000), who wished to recreate the cognitive demands of a real-work situation in which the workload demands typically increase progressively.

For each participant, the experiment proceeded as follows: in Phase 1, the participants carried out a learning task in which they were presented with 12 alarms and were required to begin to learn them; the alarms were each presented 12 times. Each participant then took part in three further phases, in the same order. In all phases, they were required to continue to learn the alarms, the learning and identification of which was presented as the primary task. In Phase 2, participants continued to learn the 12 alarms whilst being presented with a mental arithmetic task followed or preceded by an observation task. In Phase 3, they were required to simultaneously carry out both a mental arithmetic task and an observation task whilst continuing to learn the alarms. In Phase 4, they carried out a simultaneous mental arithmetic/observation/talking task whilst continuing to learn the alarms. Data were collected for both the learning task and the secondary task(s) except in the case of the talking task, which was simply the participants engaging in a discussion with the experimenter, whilst carrying out the other tasks.

The alarms used in the study are described in Table 2. The alarms tested are typical of those found in an ICU, in that many of them are high pitched and have tonal repeating patterns of similar types. Acoustically, high pitch is of little benefit, can cause unnecessary annoyance and can be difficult to localise. None of the alarms tested here had any obvious relationship between the hazards that they signalled. Many of them share similar rhythmic patterns, and would be difficult to differentiate on the basis of verbal descriptions of the sounds alone. One of the key features known to cause confusion between alarms is the sharing of similar temporal patterns (Patterson 1982; Edworthy and Meredith 1997). On this basis, we anticipated confusion between some of the alarms in this set. For example, two of the alarms (HM and R) were continuous tones and can be thought of as the most acoustically and perceptually simple alarms. Although they are considerably different in pitch (HM is much lower in pitch than R), previous evidence (Edworthy and Meredith 1997) would suggest that these alarms would be readily confused. At the next level of complexity, some of the alarms were single-pitched sounds played intermittently (P and V). There was also a group of sounds which consisted of a two-tone pattern, repeated (A, Hum, O and PM). Here, the absolute pitches and the nature of the difference in pitch between the tones were variable; thus, though the listener had some information which might have allowed discrimination between them, in general these alarms, and certainly the verbal descriptions one might use for them, were otherwise similar. For one of them (Hum), the repetition was of a double, rather than a single, pulse, whereas it was a single pulse for the other three alarms of this type. There were two alarms which had a rather distinct temporal progression (B and F). Both of these began with a longer, lower pitched tone followed by much higher, faster tones. For one of these (B) the higher repeating pattern was two-tone, and for the other (F) it was a single tone. The final two alarms (S and P) were each unique. S was a complex rhythmic pattern of short and long tones at a very high pitch and P had a clear rhythm and pitch pattern (three pulses at 280 Hz followed by one at 220 Hz). P thus sounded more like a defined melody than the other alarms, and was similar to the opening four notes of Beethoven's Fifth Symphony. This offered the potential for learning via mnemonic. P was also lower in pitch than many of the others.

Method

Materials

Alarms. A set of 12 alarms were used in this study. Five of the six most frequently occurring alarms in Study 1 [ventilator 'V', pulse oximeter (here called pulse meter, 'PM'), infusion pump 'I', syringe 'S' and humidifier 'Hum'] were included in

Table 2. Description of each of the 12 alarms tested, Study 2.

Alarm identifier	Name	Description
A	Aircall	Two-tone pattern: 0.5 s at 500 Hz + 1 s at 250 Hz (plus 0.2 s gap); repeating
B	Bleep	Initial tone 0.7 s at 280 Hz followed by repeating 2.6 kHz/2 kHz pattern for 4 s; repeating
HM	Heart monitor	1 kHz continuous tone
F	Fire	Initial tone of 1 s at 260 Hz followed by burst of short (0.25 s) tones at 2.6 kHz; repeating
Hum	Humidifier	Two pulses of 0.15 s each at 1 kHz followed by two (0.15 s each) at 930 Hz; repeating
I	Infusion	Three pulses for 0.15 s each at 280 Hz followed by one for 0.15 s at 220 Hz; repeating; 'Beethoven's Fifth'
O	Oxygen	One pulse for 0.5 s at 260 Hz followed by one for 0.5 s at 275 Hz; repeating
P	Perfusion	0.5 s tone at 1760 Hz followed by interval of 1 s; repeating
PM	Pulse meter	One tone for 0.25 s at 400 Hz followed by one tone for 0.25 s at 310 Hz; repeating
S	Syringe	Rhythmically complex pattern of tones lasting 1 and 0.1 s at 3 kHz; repeating
V	Ventilator	Tone for 0.35 s at 2.2 kHz followed by pause of 0.35 s; repeating
R	Respiratory	Continuous tone at 2.35 kHz

this set, along with seven other alarms and other alarm-type sounds recorded from other areas of the hospital during preliminary visits. We did not include the ECG alarm (the final alarm of the six most frequently occurring alarms in Study 1) in this set as this typically requires no action (it functions typically as a monitoring sound). Table 2 shows the names of the alarms used throughout, as well as a description of their main acoustic features. These were presented for learning throughout the four phases of the study. In all cases, a relatively simple name was used for the alarm, which meant that sometimes the name given to the alarm was the name of its function (e.g. perfusion 'P' and oxygen 'O'); in some cases, this was the name of the equipment (e.g. ventilator 'V' and heart monitor 'HM'), and in some cases the common term for a sound was used (e.g. bleep 'B'). We changed the name of the pulse oximeter to pulse meter for this study, as it was thought that the participants would understand the nature of a meter, but not an oximeter.

Secondary tasks. Three secondary tasks were used in various combinations – a mental arithmetic task, an observation task and a talking task. The mental arithmetic task used was a 'brackets' task. The task consists of A4 sheets containing 18 calculations per sheet. For each calculation, participants were presented with two calculations and were required to tick the one which resulted in the larger number. For example, a single item in this test might consist of:

$$\frac{42}{7}, \quad (\text{A})$$

$$27 - (6 \times 3). \quad (\text{B})$$

Participants were required to calculate both A and B in their heads and were then to select which of the two gave the higher answer, which in this case is B. The tasks were done on paper. The order of the calculations was counterbalanced across participants by presenting the sheets in different sequences.

The observation task was a vigilance task presented on a second computer, and was an 'F and E' task (Fischler et al. 1980). Starting with a clean screen, a letter F would appear either once every 0.9 s (the 'fast' mode) or every 3 s (the 'slow' mode), in both cases for 30 s. These would gradually fill up the screen. At one point in the sequence, a letter 'E' would appear and the participant's task was to press the spacebar on the computer as soon after they had detected the 'E' as possible.

The talking task (Phase 4 only) consisted of the experimenter engaging the participants in conversations as to how to reach local landmarks. They were also engaged in conversations about the courses they were currently doing. Participants were made aware that talking was an extra, explicit task in the experiment. The topics of conversation were kept the same throughout, although the detailed content varied. No formal measure of performance on this task was taken, as the purpose of this task was to simulate talking in the ICU.

Participants

Twenty participants (12 female and 8 male) participated in the study. They were recruited from the undergraduate participant pool in the School of Psychology, University of Plymouth. The ages ranged from 18 to 37 years. None reported any hearing problems and none had any medical experience.

Procedure

Phase 1. The experiment took place in a university laboratory containing a computer controlling the alarm task. A set of 12 alarms was learned using a paired-associate learning paradigm similar to that used by Edworthy et al. (2011). First, each of the 12 sounds was presented to the participant via a computer (in a different order for each participant) and the names of each of the alarms were told to the participant by the experimenter. The names given were those listed in Table 2. The individual names of the alarms were also presented to participants as cards placed in front of them. Participants were also given a brief description as to how each of the alarms might function in the ICU (e.g. which kinds of equipment they might be associated with, or which physiological functions they might be associated with), but this was done only once, briefly, at the start of the study. Once the 12 alarms had been presented, each of them was presented to each participant (in a different random order) and the participants were asked to name each alarm. If they were incorrect, they were told the correct answer and were also shown the card with the name. This was repeated 11 times (12 alarms, 12 times each) and their responses were recorded. The pace of this phase was determined by the speed with which the participants could respond and the time taken to give the feedback, but typically took approximately 45 min. The participants were then asked to return 1 week later in order to participate in Phase 2.

Phase 2. Each participant returned a week later for Phase 2. The experimental set-up used was a pair of computers, one concerned with the alarm task and the other with the observation task. The mental arithmetic task was set out on paper between the two computers. First, the participants were reminded of the names and sounds of all the 12 alarms previously tested by showing them the names on the computer and by the experimenter telling them the names as the alarms were played. The experiment then began. The participant took part in two conditions: one alongside a mental arithmetic task and the other alongside an observation task. In both conditions, they were presented with each of the 12 alarms four times each, and were asked to name them whilst carrying out the current secondary task. The order of the two conditions was counterbalanced across participants. The primary task was to name each alarm as it was played by the computer. The alarms were presented at approximately 30-s intervals, with each of the 12 alarms being presented four times per participant per secondary task, block-randomised across each of the two secondary tasks. Participants were told that the alarms would be played at random intervals by the computer and that they were required to respond as quickly and as accurately as possible by clicking a mouse pointer on the name of the alarm they thought had been played. They were told that their priority was to respond to the alarm sounds. In the mental arithmetic task, they were instructed to work through as many of the tasks as possible on a continuous basis after having the instructions explained to them. In the observation task, participants were presented alternately with the 'slow' and the 'fast' versions of the F and E task on a continuous basis. They were told,

You will see a number of 'F's appearing on the screen one after another, filling up the screen. At some point the letter 'E' will appear instead of an 'F'. As soon as you see the E press the spacebar as quickly as you can.

Each new task appeared either when the participant pressed a key to indicate that an 'E' had appeared, or was timed out. Each of the two sessions lasted approximately 25 min. Participants were asked to return the following week for Phase 3.

Phase 3. The participants were first reminded of the names and sounds associated with each of the 12 alarms, as described above. They were then told that responding to the alarms was the primary task and that the alarms would sound randomly across the experiment, and they were required to respond by clicking the mouse pointer on the correct answer. They were then reminded of both of the secondary tasks (mental arithmetic and observation) and were told that in this phase they were required to carry out both secondary tasks, along with the primary task, simultaneously. The phase took approximately 25 min. Each alarm was presented four times (at approximately 30-s intervals) in a block-randomised order across the phase. Participants were asked to return the following week for Phase 4.

Phase 4. This was conducted a week after Phase 3, and consisted of the identical procedure described in Phase 3, except that throughout this phase the experimenter engaged in conversation with the participants on such topics as how to get to landmarks in the city, and matters relating to the course they were studying.

At the end of each phase, participants were asked to rate the perceived workload of the task on a 1–100 scale.

Results and discussion

The data for Phases 1–4 are presented separately, as comparisons between the initial and later stages are not necessarily meaningful at a statistical comparison level. Thus, the comparisons between Phase 1 and the other phases are narrative rather than statistical.

Primary task performance

Table 3 shows the mean number of correct identifications of each of the 12 alarms during the learning phase, Phase 1. This demonstrates that some of the alarms were recognised more accurately than others across the learning phase. Post-hoc *t*-tests on the mean scores for each of the 12 alarms revealed that V produced significantly fewer correct responses than A, B, F, I, O, S, R and HM. Analysis also showed that Hum, P and PM produced significantly fewer correct responses than A, HM, S, I and R ($p < 0.01$).

Two confusions, between P and V, and between HM and R, were significant at $p < 0.01$ level. Several confusions were found at $p < 0.05$ level: between A and B, A and O, B and F, B and HM, B and V, HM and P, F and S, I and P, O and OP, O and PM, and between PM and V.

Figure 3 shows the mean number of correct responses to each alarm for Phases 2–4. A two-way within-subjects ANOVA was carried out on the correct responses across Phases 2–4. Significant effects were found for phase ($F[2, 18] = 4.06, p < 0.05$) and alarms ($F[11, 18] = 3.27, p < 0.01$), and there was a significant interaction between phase and

Table 3. Mean number of correct responses to each of the 12 alarms, Study 2, Phase 1.

Alarm	Number of correct identifications (12)
A	9.75
B	9.30
HM	10.50
F	9.05
Hum	8.75
I	10.50
O	9.40
P	8.25
PM	8.30
S	10.20
V	7.70
R	9.70

alarm ($F[2, 11] = 1.77$, $df = 22$, $p < 0.05$), as shown in Figure 3. Though Figure 3 shows the number of alarms correctly identified for both of the secondary tasks in Phase 2, for the purposes of analysis a mean overall score was used.

Simple effects calculations for performance across the phases for each of the 12 alarms indicated that the means were significantly different only for P, S, V and R across the phases. Means for these four alarms across the three phases were 2.7 (Phase 2), 3.2 (Phase 3) and 3.5 (Phase 4) for P; 2.6 (Phase 2), 3.6 (Phase 3) and 3.7 (Phase 4) for S; 1.8 (Phase 2), 2.5 (Phase 3) and 2.7 (Phase 4) for V and 2.9 (Phase 2), 3.7 (Phase 3) and 3.5 (Phase 4) for R. For all four of these alarms, the mean was significantly higher (Tukey's HSD, $p < 0.05$) in Phases 3 and 4 than in Phase 2. There were no significant differences between Phases 3 and 4 for any of the alarms.

In terms of the phase effect, the mean number of times each alarm was identified correctly for each phase was 3.15 for Phase 2, 3.45 for Phase 3 and 3.4 for Phase 4 (all out of four presentations of each alarm). The number of correct identifications was significantly lower in Phase 2 than in the other two phases (Tukey's HSD, $p < 0.05$), but there was no difference between Phases 3 and 4. The main effect for alarms indicated that some alarms were more readily learned and recognised than others. Overall, the only significant difference across alarms was that V was correctly named significantly less often ($p < 0.05$) than any of the other alarms except P and S.

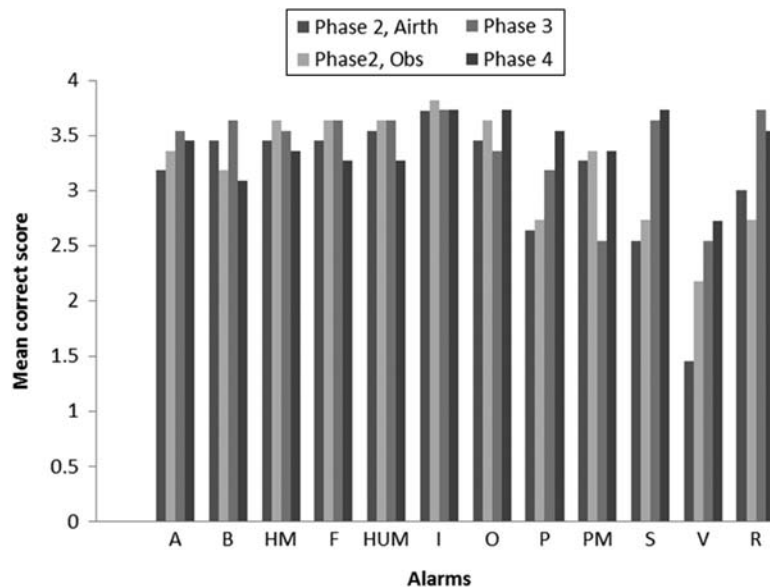


Figure 3. Alarm \times phase interaction, Study 2. This shows the mean number of correct identifications (max. = 4) for each alarm in each of the phases.

Table 4. Significant confusions between alarms, Study 2.

Alarm heard	Confused with			
	Phase 2 Arith	Phase 2 Obs	Phase 3	Phase 4
A	O	O	O	O
B		O, F	F	A
HM	R		R	
F	B		B	B
Hum		S		
I			Hum	
O				
P	V	V	V	V
PM	A			
S	P, V	P		
V	P	A, P	P	
R	HM	HM		HM

Notes: The column on the far left indicates each of the 12 alarms. The columns to the right indicate the alarms with which each of them was significantly confused with ($p < 0.01$) for each of the phases.

Primary task performance: confusions between alarms

Much of the poorer performance with some of the alarms can be explained by confusions between them. Multinomial distribution analysis (which takes account of any bias which might have occurred, e.g. in the case of asymmetric biases in alarm naming between alarms which were typically confused) was carried out on the errors made in recognising the alarms. Significant confusions ($p < 0.01$) between alarms for each of Phases 2–4 are shown in Table 4. The confusions between alarms demonstrate that whilst some persisted throughout the experiment (e.g. between A and O and between P and V), some confusions were observed only in later phases (e.g. between alarms B and A during Phase 4 only).

Secondary task performance

The mean number of correct responses across the three phases for the mental arithmetic task was 29.8 (Phase 2), 21.7 (Phase 3) and 6.4 (Phase 4) out of a possible 36. The mean number of correct responses in the observation task across the three phases was 18.75 (Phase 2), 9.6 (Phase 3) and 10.8 (Phase 4) out of a possible 40. In order to compare performance across the three phases within each of the two secondary tasks, Friedman one-way ANOVAs were carried out for each phase on the mental arithmetic and the observation task separately. For this, each participant's performance in the three phases in each of the secondary tasks (separately) was ranked from 1 to 3, with 3 being their best performance for that task and 1 being the lowest performance for that task across the three phases. For the mental arithmetic task, the corrected-for-ties ANOVA showed a significant effect across the three phases ($F_r = 18$, $df = 2$, $p < 0.01$). Post-hoc comparisons (critical z -value 11.2) showed that performance was significantly better in Phase 2 than in both Phases 3 and 4, and significantly better in Phase 3 than in Phase 4. For the observation task, the Friedman ANOVA also showed a significant effect ($F_r = 16.5$, $df = 2$, $p < 0.01$). Performance was significantly better in Phase 2 than in both Phases 3 and 4. There was, however, no significant difference between Phases 3 and 4. The data generally demonstrate that as tasks were added, performance on the secondary tasks went down.

Subjective ratings of workload

Participants rated the perceived workload of each of the phases. The mean workload rating of each phase (out of 100) was 39.5 for Phase 2, 72.2 for Phase 3 and 93.6 for Phase 4. Page's trend test showed that there were significant differences between each phase of the study, with task difficulty becoming significantly higher between Phases 2 and 3, Phases 3 and 4, and between Phases 2 and 4 ($L = 321.5$, $p < 0.001$). These ratings serve as a manipulation check, demonstrating that as tasks were added, the participants felt that their workload was increasing through the phases.

General discussion

Alarms

The data demonstrate that when alarm learning and recognition is indicated as the primary task, people are able, to some extent, to improve their recognition of a set of alarms with continued exposure to them, even when that exposure is in the

context of increasing workload. Recognition of the alarms significantly improved from Phases 2 to 3, even though workload increased from Phases 2 to 3. However, participants were also further exposed to the alarms and therefore had more opportunity to learn them. Participants did not further improve from Phases 3 to 4, suggesting either that learning had peaked or that the addition of the talking task made it more difficult for them to focus on the primary task. Comparisons with Phase 1 suggest that continued learning did take place though the effects are not consistent from alarm to alarm. The data do not address the issue of the trade-off between further exposure and the increased workload of the additional tasks, so we can only speculate on this issue. We also do not know what performance would have been like had participants not been previously exposed to the alarms (as they were in Phase 1). One interesting issue is that despite considerable exposure to the alarms, importantly, participants struggled to identify specific alarms even at the end of the study. Confusions between some alarms persisted, and participants also developed new confusions between alarms which had not necessarily been present in the earlier phases. The implication of the latter observation is that, even with prior exposure, those meanings can be confused when the participant is under workload pressure.

The evidence for the alarm learning itself demonstrates two key points. The first is that some alarms are more difficult to learn than others, and the second is that some confusions (which tend to account for the lower learning) persist despite considerable exposure to the alarms. The learning data indicated that the alarm which was easiest to learn was I and the most difficult was V (which was confused throughout the study with P). The other most persistent confusion was between alarms B and F. The nature of these confusions is consistent with those found in previous related work (Meredith and Edworthy 1994; Edworthy and Meredith 1997). For example, both P and V are regular repeated tones at a relatively high frequency, though one is higher pitched and has a slower repeating rate than the other. The conclusion is that this shared temporal regularity and high pitch was the source of this confusion, a conclusion which is supported by previous literature (e.g. Meredith and Edworthy 1994). Another persistent confusion, between B and F, is also underpinned by acoustic similarities. Both begin with a relatively long, low-pitched introductory tone and although they then differ somewhat in pitch pattern after this initial tone (one is a repeated high-pitched tone, the other is a two-tone repeated pattern), the presence of an initial long tone followed by repeated short high pitches appears to lead to significant confusion. A third persistent confusion was between HM and R, which are both continuous tones (though one is much higher than the other). *A priori* we would not predict that those two alarms would be confused because there is a considerable difference in pitch. If, however, participants used similar verbal labels for these two alarms – for example, ‘continuous tone’ – then this confusion becomes understandable. The degree to which verbal labels mediate responses to alarms is a topic on which further research could be carried out; inevitably, similar acoustic features lead to similar verbal descriptions. Experiments which are able to dissociate one from the other might provide fruitful data in clarifying the extent to which verbal labels mediate responses.

It is important to note also that the alarms which tended to the most confused were often found to be amongst the most frequently occurring in Study 1. For example, the ventilator alarm sounded frequently in Study 1, and was one of the most consistently confused alarms in Study 2.

A fourth persistent confusion, though somewhat different to the others because it was asymmetric in nature, was between A and O. In this case, A was persistently mistaken for O, but O was not persistently mistaken for A, suggesting that participants were biased towards alarm O at the expense of alarm A. The lack of significant confusion between O and any other alarm suggests that participants had a good idea as to the identity of this alarm, but tended to use that response too often, and to name other alarms as O in addition to naming it correctly. Here, we can speculate that the source of this bias might be in the availability of the response to the participant; ‘Oxygen’ is a more obvious medical label than ‘aircall’, so participants may have selected the former more readily. However, without further evidence we can only speculate on this. The two sounds do share a similar two-tone repeating pattern, which might underpin the confusion, though in one case the pitch change from the first to the second tone is upward and in one case it is downward.

Evidence that participants might indeed have been using mediating verbal labels where possible comes from the apparent lack of confusion suffered by I, which in terms of acoustic description is quite similar to some of the other alarms. On only one occasion (Phase 3) was it confused with another alarm (Hum). It is relatively easy to assign it a verbal label (‘Beethoven’s Fifth’, which is both a mnemonic for the individual pulses and a memorable name for the alarm, both factors which should make it easier to memorise and recognise).

Of course, participants’ labelling of the descriptive features of the alarms does not necessarily help them to identify the actual meaning of the alarm sound, so a further route from that label to the meaning of the sound may still be needed. The usefulness of any intermediate labels might be in allowing participants to make a link from their description, to the function of the sound – for example, the word ‘high’ as in ‘high-pitched’ can be readily linked to a high value of almost any physiological parameter, such as temperature, blood pressure and so on. Such links would serve to reduce the sound-referent distance.

The potential use of verbal labels in remembering and recognising alarms speaks of a number of pertinent issues in alarm design and in auditory cognition more generally. Ballas (1993) demonstrated that the ease with which everyday

sounds can be recognised is improved by the ease with which verbal labels can be used to describe those sounds. More generally, the ease with which listeners are able to learn alarms appears to be determined by the closeness of the sound-referent relationship (Keller and Stevens 2004; Perry et al. 2007; Petocz, Keller, and Stevens 2008). The alarms tested in the studies presented here were all abstract in nature – there was no obvious link between the alarms and their functions, and they therefore possessed either only a very weak or a non-existent sound-referent relationship – and so using mediating verbal labels in order to improve performance would serve the need to reduce the sound-referent distance and thus improve performance. There is evidence to show that verbal labels can influence the ease with which alarm sounds can be learned (Edworthy and Meredith 1997) and also that if people are given free rein to generate their own verbal labels and other ways of learning alarm sounds, this can improve recognition of the sounds to the point where learning abstract sounds is no more difficult than learning everyday sounds (Edworthy and Hards 1999). The relationship between alarm sound, its ‘official’ name and people’s own ways of labelling them, whether created through the use of description of their acoustic properties (as indicated above), mnemonic or other learning aid, is a topic which is not much explored in the literature, but is one to which further research effort could be usefully applied. Indeed, a study using different names for this same set of alarm sounds (Edworthy and Meredith 1997) showed that the names used for the sounds affected people’s ability to learn individual alarms depending on how they had been labelled. In other words, specific labels appeared to affect people’s ability to learn alarms independently of their acoustic characteristics.

A related issue is the extent to which confusions between alarms are due to there being other, similar alarms in the alarm set. For example, the major confusions between P and V, and between HM and R, may not have occurred if there had been only one example of a ‘long tone followed by higher, short tones’ and one of ‘continuous tone’ in the set. The issue of heterogeneity and uniqueness in alarm sets is one which has not commanded much attention (though see Edworthy et al. 2011), but the interactions between individual alarms within a set (and the possible confusions that will result) are an important topic for consideration, as evidenced by the impact that excessive homogeneity within an alarm set has on the learnability of the current IEC 60601-1-8 alarm sounds. All of the alarms in that set are tonal (and therefore abstract in nature) and all consist of three pulses or five pulses in two different versions (the five-tone version is the high-priority version of the equivalent three-pulse alarm). There is very little acoustical information available to the listener in order to discriminate these alarms, and generating verbal labels to distinguish between them would be almost impossible. The listener has to resort to using the melodic differences between them, which probably explains why musicians tend to fare better with these alarms than non-musicians as they have access to this knowledge and framework (e.g. Wee and Sanderson 2008). The use of mnemonics to differentiate between these alarms has been advocated. Specifically, sets of words describing the functions of the alarms have been suggested, for example ‘Temp’-ra-ture a-larm’ for the temperature alarm in this set. However, as all of the alarms in the IEC have the same number of pulses and the same rhythm, it is hard to see how these mnemonics would function. For example, mnemonics in birdsong recognition work by virtue of different bird calls having different number of ‘chirps’ and different rhythms and stress patterns. The yellowhammer’s call can be recognised by the mnemonic ‘A-little-bit-of-bread-and-no-cheese’ and the wood pigeon by ‘Take twoooo shoes, Judy’. Two different species of birds whose call was a regular ‘chirp-chirp-chirp-chirp’ would not lend themselves to discrimination via mnemonic in the same way, and the IEC alarms are the equivalent of eight birds with the same call. Wee and Sanderson (2008) found that the use of mnemonics for the IEC 60601-1-8 alarms was of little use in overall learning of those alarms, though they did create a narrower range of confusions between alarms.

There is also direct research evidence suggesting that particular acoustic features, especially temporal patterns, are important in discriminating between alarms (Edworthy et al. 2011; Edworthy and Meredith 1997; Meredith and Edworthy 1994; Patterson 1982) and yet the sharing of similar temporal patterns between alarms persists in the medical arena, both in the IEC 60601-1-8 alarms and in the alarms tested in these studies. Our data show quite clearly that the sharing of similar temporal patterns (including the continuous tone) leads to confusion between alarms, even after considerable exposure to them.

In addition, persistent in the medical arena is the use of abstract sounds as alarms *per se*, rather than alarms which can serve as a metaphor which may bring the sound closer to its referent in meaning. There is considerable research evidence which demonstrates that alarms for which there are clear relationships between sounds and their meanings (these sounds are usually referred to as ‘auditory icons’) can result in quicker learning (Keller and Stevens 2004; Leung et al. 1997; Perry et al. 2007; Ulfvengren 2003) and faster response times (Belz, Robinson, and Casali 1999; Graham 1999). The data we present in our studies demonstrate that abstract alarms can be learned, but not to perfection even with considerable exposure. Whilst there is a preference for abstract alarms in the medical arena, the issue of slow learning and confusability will persist. As approaches to alarm fatigue solutions filter through to practice (e.g. Keller et al. 2011), the relevance of the design of alarm sounds, and people’s ability to recognise and learn them, will become more apparent. There are many research issues concerning the use of abstract alarm sets which require further elaboration. These include the relative contributions of acoustic and verbal encoding in the learning of those alarms, the impact of heterogeneity of design across

an alarm set and the impact on learning of techniques expected to help with learning alarms, such as the use of mnemonics, which help to strengthen the sound-referent connection. It is likely, however, to be possible to design 'auditory icon' style alarms for the functions described in IEC 60601-1-8 which can be learned more rapidly than is typically the case for abstract alarms. The research effort which might potentially be aimed at abstract alarms needs to be seen in the context of there being other designs and types of alarms which may be more conducive to learning and which may therefore minimise the learning problem (Edworthy 2011; Wilcox 2011). Technologies which allow the rapid, online specification of alarms will also be useful in addressing the design problem, in that the relationship between design and performance can be accessed more directly and flexibly (Bennett and McNeer 2012).

Task performance

Throughout these studies the learning of and responding to alarms were presented as the primary task, though in any real medical environment it would be likely that responding to alarms would be a secondary task, particularly as most alarms are either spurious or require little immediate attention. We designated the alarm task as the primary task because our focus was on the degree to which the tasks, which mimicked those observed in the ICU, would interfere with people's ability to learn the meaning of the alarms. Performance on each of the tasks varied across participants. For the alarm-learning task, performance at the end of Phase 1 varied from close to 100% correct to approximately 55% correct alarm identifications, some participants being much more successful at learning the alarms in the learning phase than others. As we took no demographic details other than age, we do not know the reason for this variation but speculate that it might be due in part to the extent of musical training (and thus the degree to which participants are able to listen and remember sounds with a relatively small degree of variation). As none of our participants had any medical training or knowledge, we cannot attribute differences to this factor.

During Phases 2–4 there were also differential patterns of change in the correct recognition of alarms. Some participants improved, some maintained their performance and some became worse. Many participants performed particularly badly in the observation part of Phase 2, but then improved in the later phases. However, almost all participants showed the expected decline in performance on the secondary tasks as they progressed through the phases.

Our data suggest, along with most data on dual-task performance and specifically that related to dual- or multi-tasking whilst responding to alarms (Bliss, Gilson, and Deaton 1995; Bliss and Dunn 2000; Burt et al. 1995; Haas and Casali 1995), that performance falls in the secondary task but not in the primary task as the workload increases and the tasks become more demanding. The workload ratings showed very clearly that participants perceived the workload as increasing as the number of secondary tasks increased. Had the primary task been the arithmetic, observation and talking tasks, then we might have predicted that performance on the alarm task would have deteriorated, though how this would have been affected by additional exposure to the alarms is a matter of speculation. The literature on task sharing, workload and attention is enormous and often intensely theoretical (e.g. Johnson and Proctor 2004; Styles 2005) and there are many studies which have looked at dual- and multi-task performance in a variety of applied spheres (e.g. Lansdown, Brook-Carter, and Kersloot 2004; Strayer and Johnson 2001). Thus, further research in the specific area of dual- and multi-tasking when responding to alarms would seem a fruitful avenue to pursue.

Although the primary/secondary task and workload literature can give us insights into how task demands will be shared, one finding to emerge from research in this area is that, though 'attending to the alarms' can be made a primary or secondary task in research experiments, in the clinical setting, attending to alarms may be a primary, secondary or background activity, depending on the context in which the alarms occur and the observer's judgement about the purpose of that alarm. This judgement typically seems to be fairly sophisticated (Bitan et al. 2004; Gorges, Markewitz, and Westenskow 2009).

The experimental paradigm

The results of these studies do not allow us to determine the extent to which the continued exposure to the alarms mitigated the increased workload experienced from one phase to the next. We selected this design because one of our aims was to pilot this task as a potential training vehicle for people required to learn medical alarms. However, other designs would allow us to draw more specific inferences and these designs are potential candidates for future experimentation. The most obvious next steps are to repeat the experiment as a between-subjects design. For example, after a common first learning phase, participants could participate in either an alarms plus arithmetic task, alarms plus observation task, alarms plus both tasks simultaneously or alarms plus both tasks plus talking tasks (or some other design which builds on workload, such as having three conditions: one with one task, one with two and one with three). This would allow direct comparison between the variable burdens of the secondary tasks. Other between-subject designs might compare alarm learning whilst performing various combinations of the secondary tasks in comparison with learning the alarms only, and their effects on subsequent

recognition of the alarms. Here, the effect of learning in a dual-task paradigm could be compared with learning alone. There is evidence that the learning of material in a dual-task environment is enhanced in comparison with single-task learning when the stimuli which are to be learnt will subsequently appear in a dual-task rather than a single-task environment (Detweiler and Lundy 1995).

Comparison with the ICU working environment

Our studies are by necessity artificial and not necessarily reflective of the relationship between alarms and the performance of other tasks whilst in the ICU. The purpose of our study was to further the understanding of the way alarms are learned and remembered whilst performing other tasks. Throughout, we specified the alarm task as the primary task though, of course, in practice the practitioner would decide their own priorities. It is also difficult to establish the degree to which clinicians recognise, or even need to be able to recognise, individual alarms within the ICU because of both the high false alarm rate and the time gap typically found between an alarm sounding and activity which appears to be a response to that alarm. In the real clinical environment, the clinician will have many cues from the patient, visual monitors, other staff and a range of other cues, which can all be brought to bear in the process of making a clinical decision. In addition, the alarm fatigue problem clouds the alarm learning and recognition issues with which this paper is concerned. If the false alarm rate can be reduced, then observations of activity in relation to more reliable alarms will enable researchers to more precisely articulate the research questions which may or may not be important in this area. Nonetheless, the ability to recognise an alarm must be of some value. For example, if a clinician knows a particular function to produce a particularly high false alarm rate, then this information can help inform the decision as to whether or not to attend to that alarm. Unrecognisable alarms in a high false alarm rate scenario (where false alarm rates might be variable) seem to represent the worst of both worlds, and an unacceptable situation for the clinician. Reducing false alarms would also improve the usefulness of applying known features of perceived urgency in alarms (Baldwin 2010; Edworthy, Loxley, and Dennis 1991; Hellier, Edworthy, and Dennis 1993; Ljungberg and Parmentier 2012) to their design, which might result in more intuitive responses which would be less reliant on learning.

References

- Baldwin, C. 2010. "Verbal Collision Avoidance Messages During Simulated Driving: Perceived Urgency, Alerting Effectiveness and Annoyance." *Ergonomics* 54 (4): 328–337.
- Ballas, J. A. 1993. "Common Factors in the Identification of an Assortment of Brief Everyday Sounds." *Journal of Experimental Psychology: Human Perception and Performance* 19: 250–267.
- Belz, S. M., G. S. Robinson, and J. G. Casali. 1999. "A New Class of Auditory Warning Signals for Complex Systems: Auditory Icons." *Human Factors* 41: 608–618.
- Bennett, C., and R. McNeer. 2012. "PT-SAFE: A Software Tool for Development and Annunciation of Medical Audible Alarms." *Anesthesia & Analgesia* 114 (3): 576–583.
- Bitan, Y., J. Meyer, D. Shinar, and E. Zmora. 2004. "Nurses' Reaction to Alarms in a Neonatal Intensive Care Unit." *Cognition, Technology & Work* 6: 239–246.
- Bliss, J., and M. Dunn. 2000. "Behavioural Implications of Alarm Mistrust as a Function of Task Workload." *Ergonomics* 43: 1283–1300.
- Bliss, J., R. Gilson, and J. Deaton. 1995. "Human Probability Matching Behaviour in Response to Alarms of Varying Reliability." *Ergonomics* 38: 2300–2312.
- Borowski, M., M. Gorges, R. Fried, O. Such, C. Wrede, and M. Imhoff. 2011. "Medical Device Alarms." *Biomedical Technology* 56: 73–83.
- Burt, J. L., D. S. Bartolome, D. W. Burdette, and J. R. Conmstock. 1995. "A Psychophysiological Evaluation of the Perceived Urgency of Auditory Warning Signals." *Ergonomics* 38: 2327–2340.
- Chambryn, M., P. Revaux, D. Calvelo-Aros, A. Jaborska, C. Chopin, and B. Boniface. 1999. "Multicentric Study of Monitoring Alarms in the Adult Intensive Care Unit (ICU): A Descriptive Analysis." *Intensive Care Medicine* 25: 1360–1366.
- Detweiler, M. C., and D. H. Lundy. 1995. "Effects of Single- and Dual-Task Practice on Acquiring Dual-Task Skill." *Human Factors* 37 (1): 193–211.
- ECRI Institute. 2010. "Top 10 Technology Hazards." *Health Devices, Nov 2009* 38 (11): 1–10.
- Edworthy, J. 2011. "Designing Effective Alarm Sounds." *Biomedical Instrumentation & Technology* 45 (4): 290–294.
- Edworthy, J. 2012. "Medical Audible Alarms: A Review." *Journal of the American Medical Informatics Association* 20 (3): 584–589.
- Edworthy, J., and R. Hards. 1999. "Learning Auditory Warnings: The Effects of Sound Type, Verbal Labelling and Imagery on the Identification of Alarm Sounds." *International Journal of Industrial Ergonomics* 24: 603–618.
- Edworthy, J., E. Hellier, K. Titchener, A. Naweed, and R. Roels. 2011. "Heterogeneity in Alarms Makes Them Easier to Learn." *International Journal of Industrial Ergonomics* 41: 136–146.
- Edworthy, J., S. Loxley, and I. Dennis. 1991. "Improving Auditory Warning Design: Relationship between Warning Sound Parameters and Perceived Urgency." *Human Factors* 33: 205–231.
- Edworthy, J., and C. M. Meredith. 1997. "Influence of Verbal Labelling and Acoustic Quality on the Learning and Retention of Medical Alarms." *International Journal of Cognitive Ergonomics* 1 (3): 229–243.

- FDA. 2011. "Alarm Monitoring Problems FDA Patient Safety News #106." January 2011. <http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/psn/transcript-rss.cfm?show=106>
- Fischler, I. S., R. A. Griggs, S. A. Warner, M. E. Sherman, and M. Levy. 1980. *Laboratory in Cognition and Perception: Student Guide*. Iowa city, IA: Conduit.
- Gillard, J., and M. Schutz. 2012. "Improving the EFFICACY of Auditory Alarms in Medical Devices by Exploring the Effect of Amplitude Envelope on Learning and Retention." In *Proceedings of 18th ICAD*, edited by M. Nees, B. Walker and J. Freeman, 240–241, Atlanta, GA: International Community for Auditory Display.
- Gorges, M., B. A. Markewitz, and D. R. Westenskow. 2009. "Improving Alarm Performance in the Medical Intensive Care Unit using Delays and Clinical Context." *Anesthesia & Analgesia* 108 (5): 1546–1552.
- Graham, R. 1999. "Use of Auditory Icons as Emergency Warnings: Evaluation within a Vehicle Collision Avoidance Application." *Ergonomics* 42: 1233–1248.
- Haas, E., and J. Casali. 1995. "Perceived Urgency and Response Time to Multi-tone and Frequency-modulated Warning Signals in Broadband Noise." *Ergonomics* 38: 2313–2326.
- Hellier, E., J. Edworthy, and I. Dennis. 1993. "Improving Auditory Warning Design: Quantifying and Predicting the Effects of Different Warning Parameters on Perceived Urgency." *Human Factors* 35: 693–706.
- IEC. 2006. *International Standard IEC 60601-1-8 (2005–2008): Medical Electrical Equipment – Parts 1–8: General Requirements for Safety – Collateral Standard: General Requirements, Test and Guidance for Alarm Systems in Medical Electrical Equipment and Medical Electrical Systems*. Geneva: International Electrotechnical Commission.
- Imhoff, M., and S. Kuhls. 2006. "Alarm Algorithms in Critical Care Monitoring." *Anesthesia & Analgesia* 102: 1525–1537.
- Johnson, A., and R. W. Proctor. 2004. *Attention: Theory and Practice*. Thousand Oaks, CA: Sage.
- Keller, J. P., R. Diefes, K. Graham, M. Meyers, and K. Pelczarski. 2011. "Why Clinical Alarms are a 'Top Ten' Hazard: How You Can Help Reduce the Risk." *Horizons (Supplement to Biomedical Instrumentation and Technology)* Spring 2011: 17–23.
- Keller, P., and C. Stevens. 2004. "Meaning from Environmental Sounds: Types of Signal–Referent Relations and their Effect on Recognizing Auditory Icons." *Journal of Experimental Psychology: Applied* 10: 3–12.
- Lacherez, P., I. Seah, and P. M. Sanderson. 2006. "Overlapping Melodic Alarms are Almost Indiscriminable." *Human Factors* 49: 637–645.
- Lansdown, T. C., N. Brook-Carter, and T. Kersloot. 2004. "Distraction from In-Vehicle Secondary Tasks: Vehicle Performance and Mental Workload Implications." *Ergonomics* 47 (1): 91–104.
- Leung, Y. Y., S. Smith, S. Parker, and R. Martin. 1997. "Learning and Retention of Auditory Warnings." Unpublished manuscript, Defense Science and Technology Organisation, Australia.
- Ljungberg, J., and F. Parmentier. 2012. "The Impact of Information and Valance on Objective and Subjective Attention Capture by Auditory Alarms." *Human Factors* 54 (5): 826–837.
- Loeb, R. G., B. R. Jones, R. A. Leonard, and K. Behrman. 1992. "Recognition Accuracy of Current Operating Room Alarms." *Anesthesia & Analgesia* 74: 499–505.
- Logan, M. K. 2011. "Signal to Noise." *Horizons (Supplement to Biomedical Instrumentation and Technology)* Spring 2011: 1.
- Meredith, C. M., and J. Edworthy. 1994. "Sources of Confusion in Intensive Therapy Unit Alarms." In *Human Factors in Alarm Design*, edited by N. Stanton, 207–220. London: Taylor & Francis.
- Momtahan, K. L., R. Hetu, and B. W. Tansley. 1993. "Audibility and Identification of Auditory Alarms in Operating Rooms and an Intensive Care Unit." *Ergonomics* 36: 1159–1176.
- Odam, S. L., and I. Ogawa. 1992. "Direct Observation of Young Children's Social Interaction with Peers: A Review and Methodology." *Behavioral Assessment* 14: 407–441.
- Patterson, R. D. 1982. *Guidelines for Auditory Warning Systems on Civil Aircraft (Paper No. 82017)*. London: UK Civil Aviation Authority.
- Perry, N., C. Stevens, M. Wiggins, and C. Howell. 2007. "Cough Once for Danger: Icons Versus Abstract Warnings as Informative Alerts in Civil Aviation." *Human Factors* 49 (6): 1061–1071.
- Petocz, A., P. E. Keller, and C. J. Stevens. 2008. "Auditory Warnings, Signal-Referent Relations, and Natural Indicators: Re-Thinking Theory and Application." *Journal of Experimental Psychology: Applied* 14 (2): 165–178.
- Raymer, K., J. Bergstrom, and J. Nyce. 2012. "Anaesthesia Monitoring Alarms: A Theory-Driven Approach." *Ergonomics* 55 (12): 1487–1501.
- Sakai, H., D. Shin, Y. Uchiyama, R. Terashima, and T. Wakita. 2011. "Slow Eye Movement as a Possible Predictor of Reaction Delays to Auditory Warnings in a Drowsy Environment." *Ergonomics* 54 (2): 146–153.
- Sanderson, P. M., A. Wee, and P. Lacherez. 2006. "Learnability and Discriminability of Melodic Medical Equipment Alarms." *Anaesthesia* 61: 142–147.
- Schmid, F., M. S. Goepfert, D. Kuhnt, V. Eichhorn, S. Diedrichs, H. Reichensperner, A. Goetz, and D. Reuter. 2011. "The Wolf Is Crying in the Operating Room: Patient Monitor and Anesthesia Workstation Alarming Patterns During Cardiac Surgery." *Anesthesia & Analgesia* 112 (1): 78–83.
- Seagull, F. J., and P. M. Sanderson. 2001. "Anesthesia Alarms in Context: An Observational Study." *Human Factors* 43 (1): 66–78.
- Siebig, S., S. Kuhls, M. Imhoff, U. Gather, J. Scholmerich, and C. Wrede. 2010. "Intensive Care Alarms – How Many do we Need?" *Critical Care Medicine* 38 (2): 451–456.
- Siegel, A. I., and N. J. Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. 2nd ed. New York: McGraw-Hill.
- Strayer, D. L., and W. A. Johnson. 2001. "Driven to Distraction: Dual-Task Studies of Simulated Driving and Conversing on a Cellular Telephone." *Psychological Science* 12 (6): 462–466.
- Styles, E. A. 2005. *Attention, Perception and Memory*. Hove: Psychology Press.

- Talley, L., J. Hooper, B. Jacobs, C. Guzzetta, R. McCarter, A. Sill, S. Cain, and S. Wilson. 2011. "Cardiopulmonary Monitors and Clinically Significant Events in Critically Ill Children." *Horizons (Supplement to Biomedical Instrumentation and Technology)* Spring 2011: 38–45.
- Ulfvengren, P. 2003. "Design of Natural Warning Sounds in Human-machine Systems." Unpublished doctoral thesis, Stockholm Institute of Technology, Sweden.
- Wee, A., and P. M. Sanderson. 2008. "Are Melodic Medical Equipment Alarms Easily Learned?" *Anesthesia & Analgesia* 106: 501–508.
- Wickens, C. D. 1984. *Engineering Psychology and Human Performance*. Columbus, OH: Charles Merrill.
- Wickens, C. D. 2002. "Multiple Resources and Performance Prediction." *Theoretical Issues in Ergonomics Science* 3: 159–177.
- Wickens, C. D., and Y. Liu. 1988. "Codes and Modalities in Multiple Resources: A Success and a Qualification." *Human Factors* 30: 599–616.
- Wickens, C. D., D. Sandry, and M. Vidulich. 1983. "Compatibility and Resource Competition between Modalities of Input Central Processing, and Output." *Human Factors* 25: 227–248.
- Wilcox, S. 2011. "Auditory Alarm Signals." *Biomedical Instrumentation & Technology* 45 (4): 284–289.
- Williams, S., and P. C. W. Beatty. 2005. "Measuring the Performance of Audible Alarms for Anaesthesia." *Physiological Measurement* 26: 571–581.