Effects of display design on performance in a simulated ship navigation environment

Juergen Sauer†*, David G. Wastell‡, G. Robert J. Hockey§,
C. Martin Crawshaw¶, Mai Ishak†† and Jonathon C. Downing¶

†Institute of Psychology, Darmstadt University of Technology, D-64289
Darmstadt, Germany

‡Information Systems Institute, University of Salford, Salford M5 4WT, UK

§School of Psychology, University of Leeds, Leeds LS2 9JT, UK

¶Department of Psychology, University of Hull, Hull HU6 7RX, UK

††University of Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia

Although automation is playing an increasing role on the ship's bridge,
empirical research on the effectiveness of alternative bridge designs is limited.
In this paper, we describe an experimental study of the benefits of integrated
information display, using a computerized simulation of a highly automated
ship's bridge. The study compared three types of interface design, which
presented radar and electronic chart information to the operator in different
ways: (a) integrated display, (b) functionally-separate display, and (c) spatially-
separate display. Effects were examined in relation to time on watch and
scenario complexity. Following extensive training on the task, 39 participants
were tested over a 4-h experimental session, during which they encountered a
sequence of collision scenarios of varying complexity. Using a dual-task
methodology, a range of measures of primary and secondary task performance
were taken, together with assessment of information sampling behaviour and
subjective operator state (workload, fatigue, anxiety and situation awareness).
The results indicated slight navigational advantages of the integrated display
over the two alternative display types, although it also incurred higher levels of
operator cost, particularly fatigue. There were no marked effects of time on
watch, but more complex scenarios were associated with impaired perfor-
mance, increased workload and reduced situation awareness. Overall, the
findings have suggested some benefits of integrating primary information
sources in a ship's bridge environment. The study further confirms the value of
experimental simulations as tools for investigating design issues for ship's
bridge automation.

1. Introduction

Over a number of years, increasing levels of automation have led to major changes in many work environments that have had profound effects on working patterns and job performance (Woods *et al.* 1997). The maritime domain has been no exception, although the pace of introduction of automatic systems has been slower in comparison with domains such as aviation and industrial process control. Significant increases in automation have nonetheless occurred, evidenced by consistent reductions in staffing levels and the strong trend towards the one-man bridge (National Research Council 1990). Although automation has undoubtedly yielded benefits for overall human-machine performance, it has also introduced a new class of human performance problems (Sarter and Woods 1995), reflecting the shift in the operator's role from direct control of a target system to a more remote, supervisory relationship (Sheridan 1997). Problems associated with this shift (dubbed by Bainbridge (1987) as 'ironies of automation') have been noted in a range of work domains, for example, mode errors in which a control action is inappropriate for the current state of the automated system (Lee and Sanquist 1996). There is no reason to suppose that ship's bridge operation will be immune from these issues, although this domain has not been as intensively researched as other environments. The aim of this paper is to attempt to study issues relating to the design and impact of automation specifically in the area of the automated ship's bridge.

The shift to supervisory control has led to two general types of performance problem. First, humans have to carry out more monitoring activities, a task at which they are not particularly good, as demonstrated by a great deal of research on vigilance (Warm *et al.* 1996). Second, the supervisory role has made it difficult for the operator to adopt manual control if the automatic systems fail (Bainbridge 1987). This is because the operator needs time to construct an up-to-date mental model of the current system state that is the basis for effective intervention. Increasing automation has also tended to extend the human operator's 'span of control', as it becomes possible for more systems to be managed by a single individual. In the nautical environment, watchkeepers are increasingly responsible for secondary task functions such as engine and cargo control, in addition to the primary task of navigation. This requires that cognitive resources are deployed across several tasks to maintain overall performance (Metzger *et al.* 2000). In addition to the dramatic increase in monitoring load and the overheads of managing divided cognitive resources, the need to control a broader range of engineering functions also places new demands on the knowledge and skills of the watchkeeper (Lee and Sanquist 1996).

As a result of automation, workload levels in the maritime environment are thus likely to become, as in other domains, more, rather than less, extreme (Raby and Lee 2001). In the main, fewer control activities will be required than in the past because most functions will be carried out by on-board automatic systems, leaving a largely supervisory role for the watchkeeper during normal system operation. Although these periods of routine activity represent low cognitive demand, such underload is in itself undesirable from a human factors point of view (Wickens 1992). At other times, levels of workload may become excessive, for example, during collision encounters or high traffic density in confined waters. Failures of automated systems may further increase workload levels, as the management of several sub-systems (navigation, engine control, cargo control, etc.) may be allocated to a single operator. Automation thus has the effect of polarizing workload, with both extremes

of workload, underload or overload, constituting a threat to performance. In ship navigation the vast majority of accidents and critical incidents (about 90%) have occurred in confined waters (Cockroft 1984), suggesting that high workload represents the larger problem. Nonetheless both extremes need to be considered.

The need for the integration of primary information sources in ship control (radar and electronic chart) has been discussed within the maritime community for many years (Smeaton *et al.* 1995). Lee and Sanquist (1996) suggest that there are two primary motivations for this development: 'perceptual augmentation' and 'control integration'. Perceptual augmentation aims to improve the operator's perception of the ship's environment (i.e. situation awareness), while control integration aims to enhance the operator's ability to control the ship (e.g. through an autopilot). Since their introduction about 20 years ago, Advanced Radar Plotting Aids (ARPA), which provide a predictive display of local marine traffic, have superseded the manual plotting of radar echoes. They are now in universal use on vessels of any appreciable size. At the present time Electronic Chart Display Information Systems (ECDIS) are also being introduced to replace paper charts and the necessity for manual plotting of position. In most sectors of the marine industry the rate of adoption is slow, at least partly because regulatory authorities have indicated that they will continue to require paper charts for some time to come. Nevertheless, ECDIS will steadily increase in use, and a question arises as to the best method of combining chart information with ARPA within the ship's bridge environment to support the seafarer's task. Since both ARPA and ECDIS are synthetic displays of the same geographical environment, there is the possibility of integrating them into a single display. This would be technologically feasible and would seem to be a natural and appealing development. In particular, as this would correspond to the proximity compatibility principle, which argues that if a task requires high processing proximity (i.e. different information sources need to be sampled), display integration should be high (Wickens and Carswell 1995). Wickens (2000) also points out the heavy demands on cognitive resources associated with visual scanning of separate displays and mental integration of the information. However, the maritime community is divided over the issue, with some accepting the case for an integration of the two functions, while others fear that this may lead to information overload, resulting in the operator having more difficulty in identifying targets on the integrated—possibly cluttered—display (Smeaton *et al.* 1995).

While we are not aware of any systematic research examining display formats in the maritime environment, a number of experiments have compared integrated and separate displays in a generic laboratory-based context (Gillie and Berry 1995, Hansen 1996). The work of Gillie and Berry showed better performance with the integrated display in a control task but no such advantage was observed on a fault detection task. Hansen (1996) found that for integrated displays the hit rate was higher on a fault detection task but so was the false alarm rate. A major concern for integrating displays is the increase in clutter and the possibly performance-degrading effects of this (Ververs and Wickens 1998). Interestingly, decluttering by hiding irrelevant information is not necessarily better than alternative measures, such as dimming redundant information (Yeh and Wickens 2000). In particular, as temporarily eliminating information from a display has the disadvantage of attracting additional decision-making costs of what kind of information to hide.

The general aim of this paper is to develop our understanding of human-machine performance in relation to the design of automatic systems for ship's bridge

operation. We will advocate the use of the laboratory-based simulation as a tool for evaluating system design parameters and the influence of environmental factors on operator performance. In the real world, opportunities for observing and under-standing human-machine performance are more limited because it does not allow critical incidents to be fashioned at will, and the parameters of these scenarios to be accurately manipulated.

The paper will describe a major experiment, carried out to examine the impact of display design and extended task involvement on operator performance and situation awareness. Specifically, we will investigate the potential benefits and possible adverse impacts of integrated displays, where static chart information and dynamic radar-based data are combined on a single display. In comparison to standard separated displays, an integrated display may have both advantages and disadvantages, inasmuch as it affects both workload and situation awareness. The integration of multiple sources of information may well make navigation and collision avoidance more reliable (the primary goal), although it may also increase workload. In this case, there may be negative consequences in terms of increased fatigue and work strain, which could lead to impairment in the response to sub-systems such as engine room and cargo operations. Situation awareness (Endsley 1995) refers to the accuracy of an operator's mental picture of the operational environment, and their ability to respond appropriately to changes in this. It may be assumed that integrated displays will enhance situation awareness, although this has, to our knowledge, not been formally tested. Since the maritime environment is characterized by a considerable level of uncertainty about the operational movements of other vessels (Hockey *et al.*, 2002), any enhancement of situation awareness is a major goal for the development of bridge systems.

2.   The simulation task

2.1. *Theoretical background*

A computer-simulated ship's bridge environment, known as CABOT (Computer-Aided Bridge Operation Task), was developed for the experimental work. CABOT is an example of a micro-world (Brehmer and Dörner 1993)—a rich computer-based simulation of a complex real work environment, modelling all critical aspects of that environment. The development of CABOT has also built on the work carried out with another computer simulation of a ship's bridge environment (Hockey *et al.* 2002). It followed a generic methodological approach, which is described in detail in Sauer, Wastell and Hockey (2000). The methodology highlights two main design issues. First, there is a need for ecological validity (ensuring that the simulation realistically reflects key features of the real work environment). Second, the design of the simulation should be based on an explicit theory of human performance.

The long duration of watches is clearly relevant to ecological validity. Ship's bridge operators typically spend long hours on watchkeeping duty (normally two watches of 4 h, separated by an 8-h rest period) before being relieved (Raby and Lee 2001). Since this is an activity that makes considerable attentional demands, long periods of work may give rise to vigilance decrements, which are frequently reported in such tasks (Wickens 1992). Task complexity is another dimension that must be incorporated into the simulation. Studies of human vigilance have typically involved very simple task scenarios and it has been argued that vigilance decrements are more common in laboratory tasks than in real work environments because the simplicity of the task fails to engage the interest and motivation of the operator (Mackie 1984,

Hockey 1997). In carrying out our experimental work it is thus vital that the task scenarios reflect the complex demands of a real ship's bridge operation.

The task environment used in the present study has been guided both by general workload theory (Wickens 1992) and the more specific framework provided by Hockey's model of *cognitive-energetical control* (Hockey 1993, 1996, 1997). The central argument of this model is that task performance is goal-regulated, and operators usually maintain adequate levels of performance on high priority (primary) tasks, even during periods of peak work demands and stress. This is achieved by a compensatory process in which additional resources are recruited to the primary task, maintaining performance on top level goals (navigation and collision avoidance functions), although only at the expense of lower priority activities (such as monitoring cargo status or changes in platform functions). Although this compensatory process may succeed in maintaining primary performance, there is a *psychological cost* to be paid in terms of increased effort and strain, particularly an increase in fatigue (Hockey 1997). A human factors analysis of the impact of changes in ship's bridge design and watchkeeping functions therefore needs to consider both primary and subsidiary aspects of the task, as well as the underlying subjective state of the operator. The simulation environment that we have designed reflects these theoretical concerns. It includes an integrated set of primary and secondary task requirements, together with measures to assess workload, situation awareness and subjective state. Given the compensatory nature of cognitive dynamics, it is the secondary tasks and subjective measures that are expected to be more sensitive to the effects of interface design and other task variables.

In the present study the main research questions concern display design and the extent to which primary information sources should be integrated. The use of complexity and task periods as further independent variables allow for the test of their possible moderating effects on display differences.

2.2. *Technical description of task*

2.2.1. *General features of CABOT*:   CABOT is a generic simulation of the essential remote displays of an advanced ship's bridge environment, where a single operator controls three vessel functions: *navigation*, *engine control* and *cargo control*. These functions represent only a selection of duties that watchkeepers carry out on real ship's bridges. While there are further important tasks, such as communication activities (e.g. gaining permission to enter waters, making distress calls), these were not modelled in the CABOT environment. A number of further environmental variables (such as weather conditions and heaviness of swell) were not included to limit the complexity of the task environment. Otherwise, time requirements would have become too great for participants to be sufficiently trained.

The navigation function is supported by ARPA, ECDIS and a control panel. ARPA and ECDIS may be integrated in one display or be displayed on separate screens. Vessel navigation is defined as the primary task in this multiple-task environment, with two secondary tasks in addition. All operator actions are recorded automatically in a results file, allowing for a quantitative evaluation of performance on primary and secondary tasks. There is also a 'playback' facility, where all completed scenarios can be re-run in accelerated motion for a qualitative assessment of performance on the navigation task. Figure 1 presents the main display of the task environment.
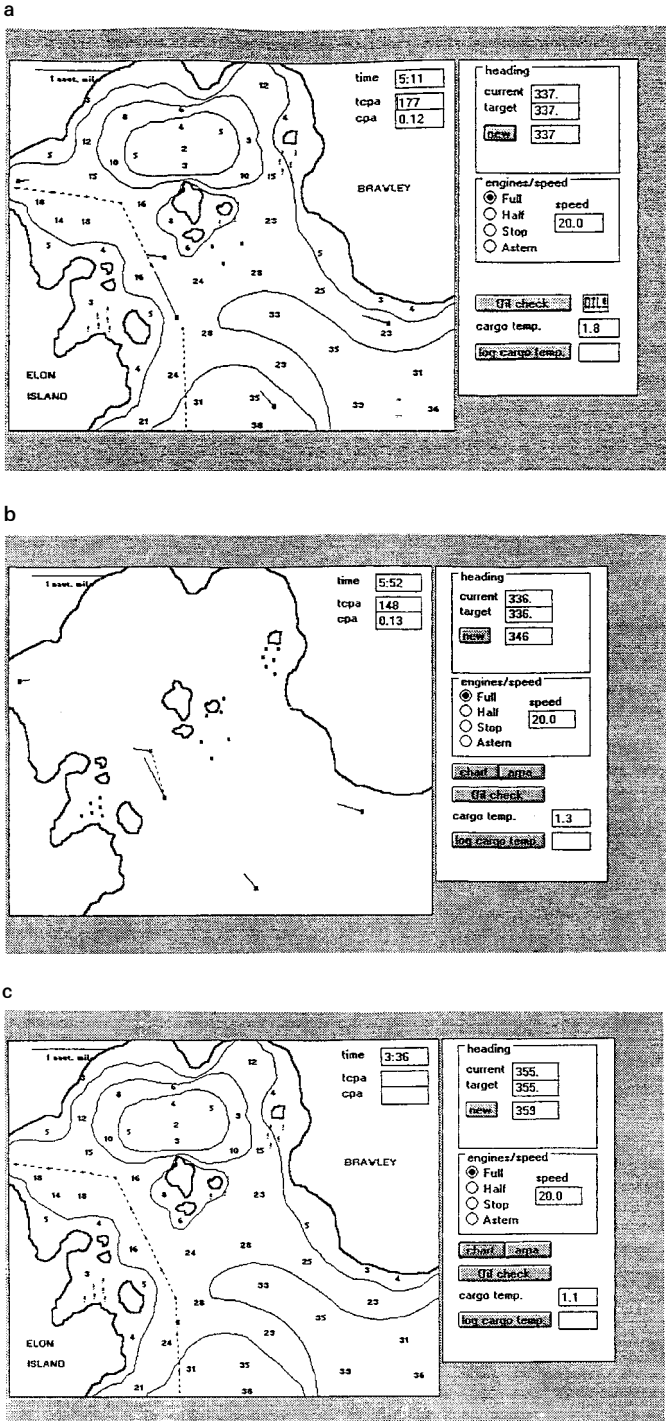
Figure 1.   Main display of Computer-Aided Bridge Operation Task (CABOT): (a) Integrated display: ARPA and ECDIS; (b) Separate display: ARPA; (c) Separate display: ECDIS (ARPA = Advanced Radar Plotting Aids; ECDIS = Electronic Chart Display Information Systems).

ARPA is a North-up true motion display (with true motion vectors) providing information about position of own vessel, true heading of own vessel (indicated by velocity vector), true heading of target vessel (also indicated by velocity vector), position of temporary navigational hazards (e.g. fishing vessels) and position of shoreline. Furthermore, ARPA indicates the closest point of approach for target vessels (i.e. predicted passing distance in nautical miles) and time to closest point of approach in seconds. ECDIS is also a North-up true motion display, providing information about position of own vessel, plotted course of own vessel, shorelines, depth contours, spot depths and permanent navigational hazards (e.g. buoys).

2.2.2. *Experimenter interface*: CABOT is highly flexible, permitting the experimenter to create scenarios of many different types, in terms of both complexity of topography (coastline, islands, channel, open sea) and encounter configurations (number and type of vessels, speed and direction of other ships, etc.). The following features can be included in the scenario: coastline, depth contours, spot depths, up to four threat vessels, up to 50 fishing vessels (stationary obstacles), up to 50 buoys, the track envelope and the plotted course of own ship. In making these features available, the software is sufficiently versatile to address a number of relevant issues in the design of a highly automated ship's bridge environment.

2.2.3. *Navigation task*: The watchkeeper has a control panel available for navigation. It consists of two sections controlling separate navigational functions: *heading* and *speed*. Furthermore, the control panel permits access to secondary task functions, such as engine and cargo control. The control panel for the 'separate' condition also provides the facility for switching between ARPA and ECDIS.

(1) *Heading*. The top section of the panel allows course changes. The new desired heading is entered into the dialogue box (figure 1), followed by clicking the 'new' button. The heading entered will then appear in the box 'target heading' with the vessel slowly turning to its new course (with a rate of turn of 9° per 5-s cycle). The reading in the 'current heading' box allows the rate of turn to be observed.
(2) *Speed*. Four different engine speeds may be selected: (1) *full speed*: 20 knots, (2) *half speed*: 7.6 knots, (3) *stop*, and (4) *astern*. Stopping distance, acceleration/deceleration and turning circle all vary with speed.

2.2.4. *Engine and cargo control task*: The two secondary tasks are engine and cargo control, and the control panel provides dialogue boxes for them.

(1) *Oil temperature*. Operators are required to monitor the engine oil temperature and to correct any deviations from normal operational limits. Oil temperature can be checked by clicking on the dialogue box 'oil check', which evokes a graphical display indicating current and previous levels of this parameter. The display shows two red lines, representing upper and lower temperature safety levels, allowing the watchkeeper to detect unsafe levels quickly. The software is programmed for a temperature drift to occur at any time but on average every 4 min, requiring the watchkeeper to check the temperature at regular intervals. The drift from a safe (target) state to an unsafe level takes approximately 30 s. In the event of an observed drift towards unsafe limits, the oil temperature has to be returned to safe levels by

clicking the 'reset' button. Since the oil temperature graph cannot be viewed simultaneously with the main ARPA/ECDIS display, the operator must switch between the temperature display and the navigational function to check safe temperature levels. Measurement of the frequency and duration of this sampling behaviour can therefore be used as an index of momentary workload on the primary (navigation) task.

(2) *Cargo temperature*. This secondary task requires the operator to record the temperature of the cargo at fixed intervals. It is a prospective memory task, in which the operator has to remember to carry out an action at a prearranged time in the future (Brandimonte *et al.* 1996, Sauer 2000). At 2-min intervals (indicated by the ever-present system clock), the operator has to take a reading of the cargo temperature, and type it into the dialogue box to log it. Timing of the logging is recorded for later analysis. Again, this can provide an indirect measure of primary task load, through failures to carry out the action, or errors in timing.

It is acknowledged that the two secondary tasks have a higher importance in the CABOT environment than on a real ship's bridge. For the purpose of this experiment, it was necessary to modify both tasks to increase event frequency and task difficulty. Inevitably, this has somewhat reduced the level of fidelity of the simulation compared to work design on a real vessel.

3. Method

3.1. *Experimental design*

The experiment used a mixed $3 \times 2 \times 4$ design. Display interface (3 levels) was varied between-subjects, while scenario complexity (2 levels) and period of watchkeeping (4 levels) were varied within subjects.

One group of watchkeepers worked with a fully integrated display (ID), in which ARPA and ECDIS information were superimposed. The second group had a functionally-separate display (FSD). Although ARPA and ECDIS were presented on the same VDU, they were selected sequentially by toggling between different screens. The third group had spatially-separate displays (SSD), ARPA and ECDIS being presented on separate (although adjacent) VDUs. The SSD interface most closely resembles current ship's bridge design, where radar and chart display units are provided on separate screens. The three conditions allowed us to address two primary questions. First, is mental integration of radar and electronic chart information facilitated by presenting the information on the same screen (ID vs. FSD/SSD)? Second, if the two sources are separated, is it better to present information sequentially on the same screen or rather on separate VDUs (FSD vs. SSD)?

Scenario complexity was manipulated by varying the number of relevant vessels and the complexity of the terrain. A complex scenario was defined as one with a high traffic density (three potential threat vessels) and a difficult terrain (complex topographical features and restricted freedom of movement). Low complexity scenarios were designed with only one potential threat vessel and a simpler/less restricted terrain. An example of each type is shown in figure 2.

A prolonged testing session of 4 h (divided for scoring purposes into 4×1 h periods) was used to simulate watchkeeping conditions at sea. This also allowed us to test for interactions of interface design with factors such as fatigue and strain that encourage performance degradation.
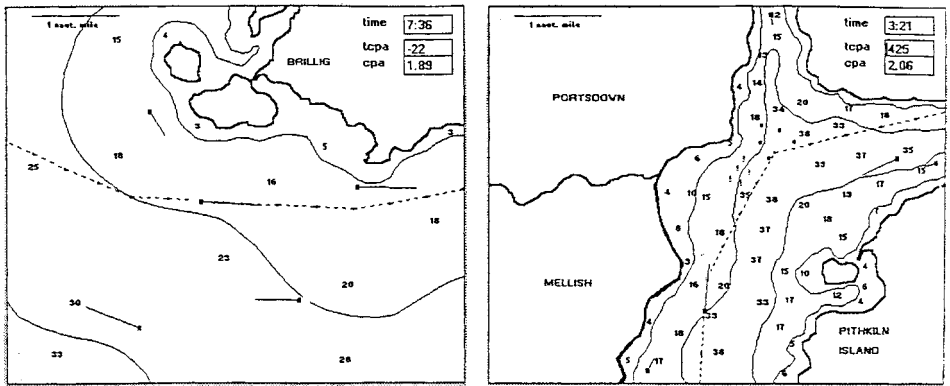
Figure 2.   Examples of low and high complexity scenario.

3.2. *Participants*

Thirty-nine participants (35 male, 4 female) were recruited for the experiment. Most were students at Hull University. Their ages ranged from 19–30 years (mean = 23.4 years). Since they were recruited partly on the basis of having a high degree of computer literacy and some understanding of automated control systems, the great majority came from engineering or computer science departments. Participants with a different background were carefully screened to ensure that they met these requirements.

3.3. *Training*

Participants received a total of 4.5 h of training, typically in groups of three, over three sessions. The first training session comprised three parts. They were first given an introductory lecture, which provided them with a context in terms of recent technological developments in the maritime industry. No information was given about any specific hypotheses under investigation. They were then taught the standard collision regulations (the 'Rules of the Road at Sea') to enable them to take appropriate action during collision encounters. Models of ships and charts were used to demonstrate different collision situations, outlining the appropriate manoeuvres to avoid a collision. During the introduction to the system, they tried out the controls of the system while practising some basic manoeuvres. This allowed them to develop an understanding of the behavioural characteristics of the vessel (e.g. stopping distance, turning circle) while having some practice with the different system controls. At the end of the training session they were given a 5-page document, summarizing the collision regulations, which they were asked to rehearse before being tested at the next session. The second training session comprised four parts. To ensure that participants were familiar with the collision avoidance regulations, they first took a written examination. If knowledge gaps were detected, the experimenter repeated the relevant section to be sure that the Rules of the Road were well understood. Training went on to cover collision assessment techniques, which were practised by the participants in a specially created scenario. They were then introduced to the two secondary tasks, and given practice on them. The final part of the training session consisted of combining primary and secondary task elements. The third training session represented a dress rehearsal for the experimental session, participants working through four scenarios

during a 1 h session. This allowed the participants to gain some more practice on the task and to develop a stable strategy for integrating the different task elements.

3.4. *Task measures*

A range of measures were taken during the experimental session to examine performance on tasks of different priority, operator's information sampling behaviour, and subjective workload assessment.

3.4.1. *Primary task measures*: In maritime performance research, cross track error (XTE) is the measure generally used for track-keeping performance. This is given by the mean deviation in nautical miles (nm) of the vessel from the plotted course during a scenario. Collision avoidance is typically measured by identifying the closest point of approach (CPA) to other vessels. We employed a similar measure to CPA, which also took into account the duration during which a prescribed 'safe distance' from other vessels (defined here as a minimum of 0.25 nm from other vessels) was transgressed. This is referred to as safe distance error, given by the percentage of time during the scenario in which the safe distance criterion was breached. An expert rating of primary task performance was also carried out, allowing a qualitative and hence more holistic assessment of watchkeeping performance. The two experts had master's tickets and more than 15 years of experience at sea. One was a pilot at a busy British port, while the other was a master with a British ferry company. Using the replay facility of CABOT, the assessors evaluated each scenario according to five criteria, using a 1–6 Likert scale. Performance was rated for the operator's implied situational awareness, appropriateness of CPA, compliance with Rules of the Road, appropriateness of alterations of course and speed, and overall performance. The assessors did not know the experimental conditions when rating a scenario. Inter-rater reliability coefficients for the different scales were satisfactory ($r = 0.7$ or higher, $p < 0.01$). A principal components analysis showed the presence of a single factor, accounting for 75% of the variance. On the basis of this, in addition to the individual scale items, analysis was also carried out on the average rating. Unfortunately, the ratings could only be completed for ID and FSD, as the expert assessors were not available for the final set of data.

3.4.2. *Secondary task measures*: There were three secondary task measures. Performance on the oil temperature monitoring task was measured in terms of both omissions and false alarms. Omissions were defined as the percentage of failures to detect the 'drift' (error) states in the oil temperature display. False alarms refer to incorrect resets of the display. Prospective memory failures were defined as the percentage of cargo temperature recordings not carried out during the required time limit (i.e. ± 5 s of scheduled time).

3.4.3. *Information sampling behaviour*: Information management activity in the secondary task was measured by three indices of sampling behaviour. The most general measure was sampling time, defined as the percentage of time for each scenario during which the oil check display was sampled. Other measures included the mean duration and frequency of checks made of the oil temperature display.

3.4.4. *Subjective state measures*: Subjective operator assessment was carried out at the end of each 15-min scenario by means of a short embedded questionnaire using

bipolar visual analogue scales. Participants were required to move a cursor on the screen to select a point on a 100 mm line to indicate how they felt during the scenario, in terms of four measures. Anxiety (relaxed–tense) and fatigue (alert–tired) have been found to provide valuable information about subjective costs in previous studies (Hockey *et al.* 1998). Mental workload (MWL) was measured using a three-item scale based on SWAT (Reid and Nygren 1988), and included three items: How much effort did you put into the task? How difficult did you find the task? How much did you feel under time pressure? Participants were also asked to complete a single item measure of situation awareness (SA). Based on Endsley's (1995) definition of SA, they were asked to rate how well they were able to identify targets (e.g. threat vessels), and to plan and predict the development of the situation as a function of available information. To minimize the disruption of the task session, only a small number of items was used to measure the concepts above. There is clearly a conflict between measuring psychological concepts thoroughly and minimizing task disruption. However, the scales used in the present case had satisfactory reliability scores, as demonstrated by previous analyses (Cronbach's $\alpha > 0.85$, Hockey *et al.* 1998).

3.5. *Procedure*

During the 4-h experimental session each participant completed 16 scenarios of 15 min each. Scenarios were randomized between participants with the constraint that two simple and two complex scenarios were presented in each hour. All sessions began at 12:30 h and took place in a laboratory, in groups of two or three participants. Each participant worked individually, with screens separating the work stations. Participants were reminded that their task was to steer a vessel through different kinds of terrain on a 4-h voyage while obeying the Rules of the Road. Furthermore, they were to complete the secondary tasks, as practised during training.

4. Results

The data were analysed using a 3-way mixed analysis of variance. For all primary and secondary task and information sampling measures, a log-transformation was required to reduce variance heterogeneity. For clarity, only untransformed data are presented in the text.

4.1. *Primary task*

The data for the various measures of primary task performance are summarized in table 1.

Table 1. Primary task performance as a function of interface and complexity

Measure	ID		FSD		SSD	
	LC	HC	LC	HC	LC	HC
Cross track error (10^{-2} nm)	1.3	8.1	1.7	10.6	2.2	8.5
Safe distance error (%)	0.6	40.7	1.6	35.1	1.7	56.7
Expert rating of performance (1–6)	5.5	3.3	5.2	3.2	*	*

*No ratings were obtained for the SSD condition. (ID = integrated display; FSD = functionally-separate display; SSD = spatially-separate display; LC = low complexity; HC = high complexity).

4.1.1. *Cross track error (XTE)*: As table 1 shows, overall course accuracy was very high, although tracking is best for ID, followed by SSD and FSD. Despite the small differences, the effect of interface was significant ($F = 4.88$; df = 2,36; $p < 0.05$). LSD tests showed that XTE was lower for ID than for the other two groups ($p < 0.01$). There was also a strong effect of scenario complexity ($F = 458.0$; df = 1,36; $p < 0.001$). As expected, XTE was worse for high complexity (HC) than for low complexity (LC). There was no effect of periods ($F < 1$), performance remaining relatively stable over the 4-h working session. None of the interactions was significant.

4.1.2. *Safe distance error*: Collision avoidance was measured in terms of the percentage of time that CPA fell below the safe distance of 0.25 nm. There were no significant differences between the three display types ($F < 1$), although safe distance errors were, as expected, much greater for the more demanding HC condition than for LC ($F = 111.5$; df = 1,36; $p < 0.001$). While small improvements in collision avoidance performance were found with increasing time on watch (29% in period 1; 19% in period 4), the main effect was not significant ($F = 2.33$; df = 3,108; $p > 0.05$). However, there was a significant interaction between periods and scenario complexity ($F = 3.29$; df = 3,108; $p < 0.05$), more marked improvements occurring over time with HC. No other significant effects were found.

4.1.3. *Expert rating*: Table 1 also shows expert ratings for navigation performance for ID and FSD. There were no significant differences, average ratings being very similar ($F = 1.46$; df = 1,23; $p > 0.05$). There was a slight but significant improvement in rated performance over the watch (from 4.0 to 4.7; $F = 4.1$; df = 3,69; $p < 0.01$), and better performance for simple than for complex scenarios ($F = 88.5$; df = 1,23; $p < 0.001$). Although the overall rating was used for the main analysis, we also analysed differences for individual items. This showed that ID was rated more highly (ID = 4.41; FSD = 4.04) for 'appropriateness of alterations of course and speed' ($F = 4.49$; df = 1,23; $p < 0.05$), but there were no other differences.

4.2. *Secondary task*
The data for the various secondary task measures are summarized in table 2.

4.2.1. *Omissions*: Omissions were defined as the percentage of undetected drifts of the oil temperature display. There were no effects of interface ($F < 1$), but more omissions for HC than for LC ($F = 5.84$; df = 1,36; $p < 0.05$). Omissions also

Table 2. Secondary task performance and information sampling behaviour as a function of interface and complexity

	ID		FSD		SSD	
Measure	LC	HC	LC	HC	LC	HC
Omissions in checking task (%)	8.1	8.5	6.1	8.1	9.4	12.3
Incorrect resets in checking task (%)	0.9	0.9	1.1	1.4	0.5	0.5
Prospective memory failures (%)	39.9	46.0	45.6	52.6	35.2	39.2
Sampling time (%)	18.5	13.9	13.7	10.0	22.0	15.2

(ID = integrated display; FSD = functionally-separate display; SSD = spatially-separate display; LC = low complexity; HC = high complexity).

increased generally over the watch ($F = 3.07$; df = 3,108; $p < 0.05$). No other main effects or interactions were significant.

4.2.2. *False alarms*: The false alarm rate on the monitoring task was defined as the percentage of incorrect resets made. Although the false alarm rate was very low ($<1\%$), incorrect resets were more common under FSD. The effect of interface was significant ($F = 4.11$; df = 1,36; $p < 0.05$), LSD tests indicating the source of the effect to be in the difference between FSD and SSD ($p < 0.01$). There were no other effects.

4.2.3. *Prospective memory failures*: Prospective memory was evaluated by counting the number of cargo temperature recordings that were either omitted or made at the wrong time (outside the limits of ± 5 s). Again, there were no effects of display interface ($F < 1$), although the data in table 2 show that performance was best for SSD, and worst for FSD. Prospective memory failures were also higher for HC than LC ($F = 13.74$; df = 1,36; $p < 0.001$), but there were no effects of periods ($F < 1$), or any interactions.

4.3. *Information sampling behaviour*
4.3.1. *Sampling time*: As table 2 shows, the overall percentage of time the oil temperature display was inspected was highest for SSD and lowest for FSD, although the difference just failed to reach significance ($F = 3.12$; df = 2,36; $p = 0.056$). As expected, scenario complexity had a very strong effect on sampling behaviour ($F = 74.0$; df = 1,36; $p < 0.001$), less time being spent on the secondary task under HC. There was also a small increase in sampling time (from 14 to 17 %) over the watch ($F = 2.96$; df = 3,108; $p < 0.05$).

4.3.2. *Frequency and duration of oil check*: Separate analyses were carried out of the mean frequency and duration of sampling. These also failed to reveal main effects of interface, although there was an interaction between interface design and time on task for sampling frequency ($F = 2.44$; df = 6,108; $p < 0.05$). The rate of sampling dropped over the session for ID (from 3.15 to 2.84), increased for FSD (from 2.43 to 2.67), and remained unchanged for SSD (2.89 to 2.93).

4.4. *Subjective ratings*
Subjective measures are summarized in table 3. Two participants (one from each of the FSD and ID groups) did not use the visual analogue scale appropriately, giving

Table 3.   Subjective ratings (0–100) as a function of interface and complexity

Measure	ID		FSD		SSD	
	LC	HC	LC	HC	LC	HC
Anxiety	27.3	34.3	17.3	27.9	19.6	28.6
Fatigue	58.3	58.5	43.4	39.4	41.9	44.7
Situation awareness	74.6	65.4	81.7	67.3	80.1	71.1
Mental workload	22.4	41.0	15.4	37.6	16.5	34.4

(ID = integrated display; FSD = functionally-separate display; SSD = spatially-separate display; LC = low complexity; HC = high complexity).

ratings of either 0 or 100 for most scale items, and had to be dropped from the analysis. Table 3 suggests generally higher levels of strain (anxiety and fatigue) and mental work load for the integrated display interface, as well as reduced situation awareness. It also indicates strong effects of complexity across all measures except fatigue. Results of analyses for each of the four measures are summarized below.

4.4.1. *Anxiety*:   The apparent effect of interface was not significant ($F = 1.59$, df $= 2,34$, $p > 0.05$). There were also no interactions with periods ($F = 1.29$, df $= 6,102$, $p > 0.05$), complexity ($F < 1$) or the 3-way interaction ($F < 1$). There was a marked effect of complexity, however ($F = 36.76$, df $= 1,34$, $p < 0.001$), but not of periods ($F < 1$) or any other interactions.

4.4.2. *Fatigue*:   The effect of interface was significant for fatigue ($F = 3.98$, df $= 1,34$, $p < 0.05$). LSD tests confirmed that fatigue was greater for ID than for either FSD or SSD (both $p < 0.05$). There was also a strong increase in fatigue over the 4-h work session ($F = 24.36$, df $= 3,102$, $p < 0.001$). The pattern of results is summarized in figure 3. There was no interaction of interface with either complexity ($F = 1.33$, df $= 2,34$, $p > 0.05$) or periods ($F = 1.70$, df $= 6,102$, $p > 0.05$). No other effects were significant.

4.4.3. *Situation awareness*:   Although SA appears lower for ID, the effect of interface was not significant ($F = 1.04$, df $= 2,34$, $p > 0.05$). Furthermore, there was no interaction of interface with complexity ($F < 1$) or task periods ($F = 1.20$, df $= 6,102$, $p > 0.05$). Situation awareness was most strongly influenced by complex-
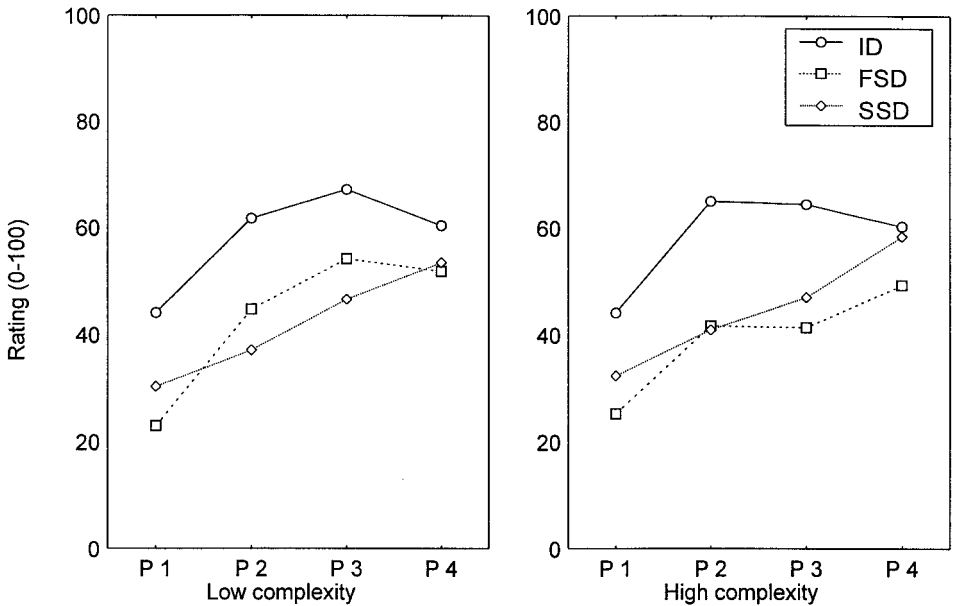

Figure 3.   Subjective ratings of fatigue (0–100) as a function of interface, complexity and task period (ID = integrated display; FSD = functionally-separate display; SSD = spatially-separate display).

ity ($F = 33.14$, df $= 1,34$, $p < 0.001$), being markedly reduced for more complex scenarios. There was also an increase over time ($F = 4.85$, df $= 3,102$, $p < 0.01$), although this occurred mainly for HC scenarios (significant complexity × periods interaction: $F = 4.83$, df $= 3,102$, $p < 0.01$).

4.4.4. *Mental workload*:   Mean MWL ratings are higher for ID than for the other interfaces, but the effect is, again, not significant ($F = 1.14$, df $= 2,34$, $p > 0.05$). There are also no interactions of interface with other variables (all F values close to 1). As for SA, there is a very strong main effect of complexity ($F = 114.81$, df $= 1,34$, $p < 0.001$), and a small effect of periods ($F = 4.83$, df $= 3,102$, $p < 0.01$). There is also a near-significant interaction between periods and complexity ($F = 2.48$, df $= 3,102$, $p = 0.07$). MWL levels are considerably lower for LC scenarios, but fall more over time for HC.

5.   Discussion

The principal research issue addressed in the study concerned the potential benefits of integrated displays, as compared with separate displays. Time on watch and complexity were used as moderators for the effect of interface design, and to enhance ecological validity. The results suggested that performance on the primary task of navigation was slightly enhanced by the integrated display. Cross track error was significantly reduced and the expert assessors adjudged course and speed changes to be more proficient under ID. In addition, this increased tracking performance was achieved without compromising collision avoidance criteria. Safe distance was maintained as effectively under ID as under the other two interfaces. Overall, no aspect of primary task performance was impaired by the integrated display.

There was also no impairment in secondary task measures. This may indicate that the improvement in navigation performance was not associated with marked increases in workload. In fact, although MWL ratings were highest for ID, the differences were not significant. Oddly, this apparently helpful display did not lead to an increase in SA, one of the reasons sometimes given for introducing such a change into bridge systems (Lee and Sanquist 1996). On the negative side, however, there is a subjective cost of the ID interface, as ratings of fatigue were higher, compared to the other systems. The enhanced navigational performance, and maintenance of an adequate response to secondary task activities thus appears to be achieved at the expense of increased concentration and task engagement (Hockey 1997).

Overall, we may conclude that the integrated display, by removing the need to mentally integrate the radar and chart pictures, provided genuine benefits in terms of supporting ship navigation, although it may also attract some costs in terms of increased workload and fatigue. A larger study with many more participants is needed to provide the necessary statistical power to detect the small effect observed for MWL (for this study, partial $\eta^2 = 0.06$, and $(1-\beta) = 0.235$). By contrast, the power statistics for the fatigue effect were: partial $\eta^2 = 0.19$, and $(1-\beta) = 0.674$. While there is no directly comparable research on this question, an evaluation of the integrated ARPA/ECDIS display was carried out on a high-fidelity ship simulator (Smeaton *et al.* 1995). This reported advantages of the integrated display over traditional bridge systems for assessing collision risk and possible manoeuvres against the navigational chart.

Interestingly, of the other two displays, the FSD interface emerged as the least attractive alternative. Performance was poorer on the oil temperature monitoring

task (both higher FA rate and reduced sampling duration). In addition to these statistically significant effects, the FSD condition also showed the poorest performance on the primary task measures, although only the comparison with ID for XTE was significant. These problems for FSD do not seem to be due to an increased experience of workload, since the MWL mean for FSD was actually the lowest of the three. Taken together, the findings suggest that the sequential provision of information on one display is detrimental, compared to either an integrated display or, indeed, the conventional solution of two displays.

The comparison between FSD and SSD is also of interest. It appears better to have all information available simultaneously, despite having to switch spatially between displays (spatial integration) than to have a single physical source with sequential switching (temporal integration). One factor may be the greater 'visual momentum' (Wickens 1992) for SSD, facilitating more easily the transition between different displays. A critical factor appears to be sampling speed. In a single display facility with temporal switching, any system usually needs a finite period (normally a few seconds) to build up the screen image (as in our experiment). In contrast, in the dual screen facility the information is continuously available. This requires only an eye or head movement from the operator, which takes only a fraction of a second. Therefore, if the operator needs to switch several times between displays (e.g. to re-check information), sampling speed may be severely slowed in the single display facility. Indeed, observations during the course of the experiment confirmed that participants sometimes switched very rapidly between ARPA and ECDIS during the SSD condition.

Although the main effect of time on watch was not a primary research question, the results are nevertheless of interest, as the strain caused by long watches is a topic of general concern. This is particularly relevant in the context of increasing levels of automation, requiring the watchkeeper to carry out more intensive monitoring (Lee and Sanquist 1996). According to Hockey's model, we would not expect to see serious decrements in primary task performance as a result of these long watches, although impairments in secondary task performance are predicted. The results of the experiment are broadly in line with these predictions. Primary task performance showed no significant change over task periods. Indeed situation awareness was better later in the session. This may reflect the general benefit of experience for our novice navigators who, despite extensive training, were still relatively inexperienced in terms of operational scenarios. It is also likely that situation awareness grows naturally within a watch as navigators (experienced or otherwise) become more familiar and involved in the specific ambient setting. As expected, fatigue levels increased over the watch, although there were no dramatic secondary task decrements. The detrimental effects of extended task involvement typically observed in vigilance studies (e.g. Parasuraman 1986) were thus not observed here. This may be attributed to the comparatively high event rate in the task environment, and the temporal structure of the session, which provided participants with a new scenario every 15 min.

Regarding the effects of complexity, it is not surprising to find high complexity associated with worse primary performance. The task is somewhat 'data limited' (Norman and Bobrow 1975), inasmuch as the quality of navigation performance is constrained by opportunities for manoeuvring without making large course changes. This means that both XTE and collision avoidance are more likely to be poorer, however much effort participants make. More interesting are the effects on secondary task performance, system management strategies and subjective state variable. These provide insights into the pattern of adaptive changes made by

participants in managing task demands. The compensatory mechanisms predicted by Hockey (1996) are clearly seen in the data, manifested in both secondary task decrements and changes in information management strategy. More complex scenarios were associated with increased omissions (failures to detect drift states) and increased prospective memory errors, as well as both shorter and less frequent checks of the oil temperature display. Under some circumstances such strategy changes may represent increased optimization of information sampling. In others they may also be seen as short-cuts, and linked with increased risk-taking in other work environments (Holding 1983, Hockey 1993). This latter interpretation is more likely in this case, because of the observed impairment in the detection of drift states. Reduced monitoring behaviour as a function of increasing workload has also been observed in another ship navigation study (Kerstholt *et al.* 1996).

The present experiment is, to our knowledge, the first well-controlled laboratory study that has compared different methods of navigational display integration on ship's bridges. Since the results provided evidence for considering the integrated display as a better option than the two alternative display types, it is difficult to see why such an integration should not be made at source, provided that there is a possibility of reducing clutter on the display. Since our laboratory-based simulation of the task environment can only model generic features of the work environment, real systems of the different display types need to be assessed separately (Kaufmann and Eaton 1994). This is due to the non-generic features of real systems, which also impinge on performance but could not be modelled in our study.

Finally, we will comment on the utility of using simulations in this kind of human factors research. The use of such 'micro-worlds' has been advocated by a number of researchers as they allow complex task environments to be investigated under well controlled experimental conditions (Brehmer and Dörner 1993). In the present study, we would make two methodological points regarding micro-worlds. First, they should be designed to test specific research questions based on a systematic theory of human performance (Sauer, Wastell and Hockey 2000). A second point relates to the external validity of micro-world findings. Although a rigorous methodology can guarantee a high degree of external validity for the micro-world, it is necessary to corroborate the primary research results in larger scale, high fidelity simulations, with stronger ecological validity (Smeaton *et al.* 1995). The present study has gone some way towards meeting the first of these goals, confirming the benefits found in our previous use of the micro-world approach in the context of manned spaceflight (Hockey *et al.* 1998, Sauer, Hockey and Wastell 2000). The second is a necessary next step before we can have real confidence in the effectiveness of integrated bridge systems in real ships.

References

BAINBRIDGE, L. 1987, Ironies of automation, in J. Rasmussen, K. Duncan and J. Leplat (eds), *New Technology and Human Error* (Chichester: John Wiley).

Brandimonte, M., Einstein, G. O. and McDaniel, M. A. 1996, *Prospective Memory: Theory and Applications* (Hillsdale, NJ: Lawrence Erlbaum).

Brehmer, B. and Dörner, D. 1993, Experiments with computer-simulated microworlds: escaping both the narrow straits of the laboratory and the deep blue sea of the field study, *Computers in Human Behavior*, **9**(2–3), 171–184.

Cockroft, A. N. 1984, Collisions at sea, *Safety at Sea*, June, 17–19.

Endsley, M. R. 1995, Measurement of situation awareness in dynamic systems, *Human Factors*, **37**(1), 65–84.

Gillie, T. and Berry, D. 1995, Object displays and control of dynamic systems, *Ergonomics*, **37,** 1885–1903.

Hansen, J. P. 1996, An experimental investigation of configural, digital and temporal information on process displays, *Human Factors*, **37,** 539–552.

Hockey, G. R. J. 1993, Cognitive-energetical mechanisms in the management of work demands and psychological health, in A. D. Baddeley and L. Weiskrantz (eds), *Attention, Selection, Awareness and Control: A Tribute to Donald Broadbent* (Oxford: Oxford University Press).

Hockey, G. R. J. 1996, Skilled performance and mental workload, in P. Warr (ed.), *Psychology at Work* (London: Penguin).

Hockey, G. R. J. 1997, Compensatory control in the regulation of human performance under stress and high workload: a cognitive-energetical framework, *Biological Psychology*, **45,** 73–93.

Hockey, G. R. J., Wastell, D. and Sauer, J. 1998, Effects of sleep deprivation and user interface on complex performance: a multilevel analysis of compensatory control, *Human Factors*, **40,** 233–253.

Hockey, G. R. J., Healey, A., Crawshaw, C. M., Wastell, D. and Sauer, J. 2002, Cognitive demands of navigation and collision avoidance in simulated ship control, *Human Factors*. Manuscript submitted for publication.

Holding, D. H. 1983, Fatigue, in G. R. J. Hockey (ed.), *Stress and Fatigue in Human Performance* (Chichester: John Wiley).

Kaufmann, R. and Eaton, R. M. 1994, User trials on the colours and symbols for electronic charts, *Applied Ergonomics*, **25,** 366–372.

Kerstholt, J. H., Passenier, P. O., Houttuin, K. and Schuffel, H. 1996, The effect of a priori probability and complexity on decision making in a supervisory control task, *Human Factors*, **38**(1), 65–78.

Lee, J.-D. and Sanquist, T. F. 1996, Maritime automation, in R. Parasuraman and M. Mouloua (eds), *Automation and Human Performance: Theory and Applications* (Mahwah, NJ: Lawrence Erlbaum), 365–384.

Mackie, R. 1984, Research relevance and the information glut, in F. Muckler (ed.), *Human Factors Review* (Santa Monica, CA: Human Factors Society).

Metzger, U., Duley, J. A., Abbas, R. and Parasuraman, R. 2000, Effects of variable-priority training on automation-related complacency: performance and eye movements, *Proceedings of the 14th Triennial Congress of the International Ergonomics Association and the 44th Annual Meeting of the Human Factors and Ergonomics Society*, San Diego, CA, 30 July–4 August. (Santa Monica, CA: HFES).

National Research Council 1990, *Crew Size and Maritime Safety* (Washington, DC: National Academic Press).

Norman, D. A. and Bobrow, D. 1975, On data-limited and resource-limited processing, *Journal of Cognitive Psychology*, **7,** 44–60.

Parasuraman, R. 1986, Vigilance, monitoring and search, in K. R. Boff, L. Kaufman and J. P. Thomas (eds), *Handbook of Perception and Human Performance* (New York: John Wiley).

Raby, M. and Lee, J. D. 2001, Fatigue and workload in the maritime industry, in P. A. Hancock and P. A. Desmond (eds), *Stress, Workload and Fatigue* (London: Lawrence Erlbaum).

Reid, G. B. and Nygren, T. E. 1988, The subjective workload assessment technique: a scaling procedure for measuring mental workload, in P. A. Hancock and N. Meshkati (eds), *Human Mental Workload* (Amsterdam: North Holland), 185–218.

Sarter, N. B. and Woods, D. D. 1995, How in the world did we ever get into that mode? Mode error and awareness in supervisory control, *Human Factors*, **37**(1), 5–19.

Sauer, J. 2000, Prospective memory: a secondary task with promise, *Applied Ergonomics*, **31,** 131–137.

Sauer, J., Hockey, G. R. J. and Wastell, D. 2000, Effects of training on short- and long-term skill retention in a complex multiple-task environment, *Ergonomics*, **43,** 2043–2064.

Sauer, J., Wastell, D. and Hockey, G. R. J. 2000, A conceptual framework for designing micro-worlds for complex work domains: a case study of the Cabin Air Management System, *Computers in Human Behavior*, **16,** 45–58.

Sheridan, T. B. 1997, Supervisory control, in G. Salvendy (ed.), *Handbook of Human Factors* (New York: John Wiley), 1295–1327.

Smeaton, G. P., Dineley, W. O. and Tucker, S. M. 1995, Display requirements for ECDIS/ARPA overlay systems, *Journal of Navigation*, **48**(1), 13–28.

Ververs, P. M. and Wickens, C. D. 1998, Head-up displays: effects of clutter, display intensity, and display location on pilot performance, *International Journal of Aviation Psychology*, **8,** 377–403.

Warm, J. S., Dember W. N. and Hancock, P. A. 1996, Vigilance and workload in automated systems, in R. Parasuraman and M. Mouloua (eds), *Automation and Human Performance: Theory and Applications* (Mahwah, NJ: Lawrence Erlbaum), 183–200.

Wickens, C. D. 1992, *Engineering Psychology and Human Performance* (Columbus, OH: Merill).

Wickens, C. D. 2000, Human factors in vector map design: the importance of task-display dependence, *Journal of Navigation*, **53**(1), 54–67.

Wickens, C. D. and Carswell, C. M. 1995, The proximity compatibility principle: its psychological foundation and relevance to display design, *Human Factors*, **37,** 473–494.

Woods, D. D., Sarter, N. and Billings, C. 1997, Automation surprises, in G. Salvendy (ed.), *Handbook of Human Factors* (New York: John Wiley), 1926–1943.

Yeh, M. and Wickens, C. D. 2000, Attention filtering in the design of electronic map displays: a comparison of color-coding, intensity coding, and decluttering techniques. Technical report (ARL-00-4 / FED.LAB-00-2) to US Army Research Laboratory, Institute of Aviation, University of Illinois.