THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Resolving contentious hospital observation chart design decisions using a behavioural experimental approach

Melany Jean Christofidis

BSc (Hons I)

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2015

School of Psychology

Abstract

A new paper-based hospital observation chart has been developed using human factors design principles. This novel design, compared to previous charts, yielded fewer errors and faster response times in chart-users' detection of patient physiological deterioration compared to other Australian observation charts that were in use at the time. In recent clinical studies, the chart has also been associated with an 11% mortality reduction amongst intensive care unit admissions, as well as a 45% reduction in the incidence of cardiac arrests. However, there are a number of points of contention as to whether this design can be regarded as best practice. First, it is unclear whether the chart offers performance benefits to users highly experienced with alternative chart designs. Second, clinicians have questioned particular features that were designed to help users detect abnormal vital sign observations. For example, there is a dispute as to whether blood pressure and heart rate graphs should be presented as separate or overlapping plots. Third, disagreement surrounds the optimal design layout to facilitate users' calculation of summary scores that represent the physiological state of a patient. In the absence of expert consensus, this thesis sought to address each of these points of contention using behavioural experiments. In general, findings supported the design choices associated with the new observation chart. Specifically, in relation to the detection of abnormal observations, it was found that (1) even users experienced with alternative chart designs performed better with the new chart; (2) blood pressure and heart rate were better presented as separate graphs (even for chart-users who preferred plots that overlap); and (3) users' performance with drawn-dot observations, an integrated colour-based scoring-system, and grouped scoring-rows was consistent with apriori human factors design principles. One design aspect of the new chart was not supported: users were found to be less accurate calculating patient deterioration summary scores when the design involved recording interim steps in the calculation. Overall, it is argued that these experiments demonstrate the value of using behavioural experiments to assess best design practice, rather than relying solely on expert opinion.

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly authored works that I have included in my thesis. I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award. I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School. I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.

Publications during candiature

Publications and presentations involving material from the thesis

Peer-reviewed papers

1. **Christofidis, M. J.**, Hill, A., Horswill, M. S., & Watson, M. O. (2013). A human factors approach to observation chart design can trump health professionals' prior chart experience. *Resuscitation, 84*(5), 657-665.
2. **Christofidis, M. J.**, Hill, A., Horswill, M. S., & Watson, M. O. (2014). Observation charts with overlapping blood pressure and heart rate graphs do not yield the performance advantage that health professionals assume: an experimental study. *Journal of Advanced Nursing, 70*(3), 610-624.
3. **Christofidis, M. J.**, Hill, A., Horswill, M. S., & Watson, M. O. (2016). Observation chart design features affect the detection of patient deterioration: A systematic experimental evaluation. *Journal of Advanced Nursing*, *72(1),* 158-172.
4. **Christofidis, M. J.**, Hill, A., Horswill, M. S., & Watson, M. O. (2015). Less is more: the design of early-warning scoring systems affects the speed and accuracy of scoring. *Journal of Advanced Nursing, 71(1),* 1573-1586.

Conference presentations

1. **Christofidis, M. J.**, Hill, A., Horswill, M. S., & Watson, M. O. (2012). *Human factors design and observation charts*. Paper presented at the 7th Annual International Conference on Rapid Response Systems and Medical Emergency Teams. Sydney, NSW.
2. **Christofidis, M. J.**, Hill, A., Horswill, M. S., & Watson, M. O. (2012). *Human factors principles of patient observation charts: Examining the utility of the Seagull Sign.* Paper presented at the 4th International Conference on Applied Human Factors and Ergonomics. San Francisco, CA.
3. **Christofidis, M. J.**, Hill, A., Horswill, M. S., & Watson, M. O. (2013). *Human factors design and observation charts.* Paper presented at the Managing the Deteriorating Patient conference. Melbourne, VIC.

Invited government presentations

1. **Christofidis, M. J.**, Horswill, M. S., Hill, A., & Watson, M. O. (2012). *Human factors design and observation charts*: Paper presented at the Recognising and Managing the Deteriorating Patient (RMDP) Reference Group Meeting. Brisbane, QLD.

Other related publications and presentations

Published research reports

1. **Christofidis, M. J.**, Horswill, M. S., Hill, A., McKimmie, B. M., Visser, T., & Watson, M. O. (2012). *Task analysis and heuristic analysis of insulin charts.* Final report prepared for the Australian Commission on Safety and Quality in Health Care. Retrieved from http://www.safetyandquality.gov.au/wp-content/uploads/2012/06/56679-Insulin-charts-heuristic-analysis-2-Feb-2011-Final-Report.pdf
2. Horswill, M. S., Preece, M. H. W., Hill, A., **Christofidis, M. J.**, & Watson, M. O. (2010). *Recording patient data on six observation charts: An experimental comparison.* Report prepared for the Australian Commission on Safety and Quality in Health Care's program for Recognising and Responding to Clinical Deterioration. Retrieved from http://www.safetyandquality.gov.au/wp-content/uploads/2012/01/35980-RecordingData.pdf
3. Horswill, M. S., Preece, M. H. W., Hill, A., **Christofidis, M. J.**, Karamatic, R., Hewett, D., & Watson, M. O. (2010). *Human factors research regarding observation charts: Research project overview.* Report prepared for the Australian Commission on Safety and Quality in Health Care's program for Recognising and Responding to Clinical Deterioration. Retrieved from http://www.safetyandquality.gov.au/wp-content/uploads/2012/01/35986-HumanFactors.pdf
4. Preece, M. H. W., Hill, A., Horswill, M. S., Dunbar, N., Adams, L. M., Stephens, J. L., **Christofidis, M. J.**, & Watson, M. O. (2010). *Developer's guide for observation and response charts.* Report prepared for the Australian Commission on Safety and Quality in Health Care's Program for Recognising and Responding to Clinical Deterioration. Retrieved from http://www.safetyandquality.gov.au/wp-content/uploads/2012/02/ORC-Developers-Guide-4-Oct-2010.pdf

Other government publications

1. Horswill, M. S., Hill, A., **Christofidis, M. J.**, & Watson, M. O. (2012). *How to conduct a behavioural study to test chart modifications* (information sheet). Sydney, Australia: Australian Commission on Safety and Quality in Health Care. Retrieved from http://www.safetyandquality.gov.au/wp-content/uploads/2012/08/EE1-ORC6-Fact-sheet.pdf
2. Horswill, M. S., Hill, A., **Christofidis, M. J.**, & Watson, M. O. (2012). *Why is it crucial to test any non-approved ORC modifications?* (information sheet). Sydney, Australia: Australian Commission on Safety and Quality in Health Care; 2012. Retrieved from http://www.safetyandquality.gov.au/wp-content/uploads/2012/08/ee1-orc-5-fact-sheet.pdf

Conference presentations

1. Dunbar, N., Preece, M. H. W., Horswill, M. S., Hill, A., Karamatic, R., **Christofidis, M. J.**, Hewett, D. G., & Watson, M. O. (2010). *A human factors approach to observation chart design can improve the detection of clinical deterioration.* Paper presented at the 6th International Symposium on Rapid Response Systems and Medical Emergency Teams. Pittsburgh, PA.
2. Preece, M. H. W., Horswill, M. S., Hill, A., **Christofidis, M. J.**, & Watson, M. O. (2010). *A human factors approach to observation chart design can improve the detection of clinical deterioration.* Paper presented at the 46th Annual Conference of the Human Factors and Ergonomics Society of Australia. Sunshine Coast, QLD.

Publications included in this thesis

The following publications have been directly incorporated into the thesis chapters:

Christofidis, M. J., Hill, A., Horswill, M. S., & Watson, M. O. (2013). A human factors approach to observation chart design can trump health professionals' prior chart experience. *Resuscitation, 84*(5), 657-665.

Contributor	Statement of contribution
Christofidis, M.J. (Candidate)	Designed the experiment (10%) Conducted data collection (100%) Analysed and interpreted the data (80%) Drafted the paper (100%)
Hill, A.	Conceived the study (100%) Designed the experiment (30%) Analysed and interpreted the data (5%) Reviewed and edited the manuscript drafts (70%)
Horswill, M.S.	Designed the experiment (30%) Analysed and interpreted the data (10%) Reviewed and edited the manuscript drafts (15%)
Watson, M.O.	Designed the experiment (30%) Analysed and interpreted the data (5%) Reviewed and edited the manuscript drafts (15%)

Christofidis, M. J., Hill, A., Horswill, M. S., & Watson, M. O. (2014). Observation charts with overlapping blood pressure and heart rate graphs do not yield the performance advantage that health professionals assume: an experimental study. *Journal of Advanced Nursing, 70*(3), 610-624.

Contributor	Statement of contribution
Christofidis, M.J. (Candidate)	Conceived the study (100%) Designed the experiment (60%) Designed the experimental materials (100%) Conducted data collection (100%) Analysed and interpreted the data (70%) Drafted the paper (100%)
Hill, A.	Designed experiment (30%) Analysed and interpreted the data (15%) Reviewed and edited the manuscript drafts (70%)
Horswill, M.S.	Designed experiment (5%) Analysed and interpreted the data (10%) Reviewed and edited the manuscript drafts (15%)
Watson, M.O.	Designed experiment (5%) Analysed and interpreted the data (5%) Reviewed and edited the manuscript drafts (15%)

Christofidis, M. J., Hill, A., Horswill, M. S., & Watson, M. O. (2016). Observation chart design features affect the detection of patient deterioration: A systematic experimental evaluation. *Journal of Advanced Nursing*, *72(1),* 158-172.

Contributor	Statement of contribution
Christofidis, M.J. (Candidate)	Conceived the study (100%) Designed the experiment (80%) Designed the experimental materials (100%) Conducted data collection (100%) Analysed and interpreted the data (80%) Drafted the paper (100%)
Hill, A.	Designed experiment (10%) Analysed and interpreted the data (5%) Reviewed and edited the manuscript drafts (70%)
Horswill, M.S.	Designed experiment (5%) Analysed and interpreted the data (10%) Reviewed and edited the manuscript drafts (15%)
Watson, M.O.	Designed experiment (5%) Analysed and interpreted the data (5%) Reviewed and edited the manuscript drafts (15%)

Christofidis, M. J., Hill, A., Horswill, M. S., & Watson, M. O. (2015). Less is more: the design of early-warning scoring systems affects the speed and accuracy of scoring. *Journal of Advanced Nursing, 71(1),* 1573-1586.

Contributor	Statement of contribution
Christofidis, M.J. (Candidate)	Conceived the study (100%) Designed the experiment (80%) Designed the experimental materials (100%) Conducted data collection (100%) Analysed and interpreted the data (80%) Drafted the paper (100%)
Hill, A.	Designed experiment (10%) Analysed and interpreted the data (5%) Reviewed and edited the manuscript drafts (70%)
Horswill, M.S.	Designed experiment (5%) Analysed and interpreted the data (10%) Reviewed and edited the manuscript drafts (15%)
Watson, M.O.	Designed experiment (5%) Analysed and interpreted the data (5%) Reviewed and edited the manuscript drafts (15%)

Contributions by others to the thesis

Associate Professor Mark S. Horswill (primary advisor), Dr Andrew Hill (associate advisor), and Associate Professor Marcus O. Watson (associate advisor) contributed to the design, data analysis, interpretation and publication of all four experiments. Their individual contributions to each experiment are outlined at the beginning of each relevant chapter. Associate Professor Horswill, Dr Hill and Associate Professor Watson also critically revised drafts of this thesis.

Statement of parts of the thesis submitted to qualify for the award of another degree

None.

Acknowledgements

First, I thank my advisors, Associate Professor Mark Horswill, Associate Professor Marcus Watson, and Dr Andrew Hill. Mark, for your positivity, optimism and patience. Marcus, for your wisdom and advice beyond the thesis. And Andrew for your humour, great attention to detail and all the time and help you have given me over the years. You are a wonderful team and I am forever grateful to each of you for your kindness, understanding and support.

I also thank Heather McKay, Lynette Adams, David Collard and staff of The Canberra Hospital, Mt Isa Base Hospital, Logan Beaudesert Hospital and The University of Queensland's School of Psychology for help with various administrative aspects of the project. Thanks also go to the health professionals and students who participated in my studies. I wish to acknowledge the Australian Postgraduate Award and the Smart Futures PhD Scholarship for providing a much-appreciated stipend during candidature. I also thank the journal editors and reviewers of my papers, whose comments were insightful and contributed to the overall quality of the final versions.

On a personal note, I want to thank my parents, George and Rhonda, as well as friends and family; especially Dargie, Lachie and Matty. Thank you for putting up with me in your own special ways. Finally, I want to thank Ricky, Steve and Mr. K. Dilkington for getting me through hundreds of hours of charting.

Keywords

human factors, hospital observation charts, design, patient deterioration, usability, nursing

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 179999, Psychology and Cognitive Sciences not elsewhere classified, 90%
ANZSRC code: 111099, Nursing not elsewhere classified, 10%

Fields of Research (FoR) Classification

FoR code: 1799, Other Psychology and Cognitive Sciences, 90%
FoR code: 1110, Nursing, 10%

Table of Contents

List of Figures

List of Tables

Preface

This thesis is presented in a non-traditional format. A set of manuscripts, largely as they were submitted to print publication, has been directly incorporated into the thesis as Chapters 2, 3, 4 and 5. The first page of each respective chapter outlines the submission history of the article. Each manuscript details an experimental study that was written up for a medical or nursing-oriented journal. Consequently, Chapters 2 to 5 focus on the clinical relevance of chart design and evaluation. These journals (as opposed to journals based in human factors, psychology or cognitive ergonomics) were selected so that the experimental findings and recommendations would reach a greater number of health professionals motivated to improve the detection of patient deterioration. However, as a consequence, the papers themselves contain limited detail regarding the underlying human factors rationales for each experiment and the links between the manuscripts are not entirely explicit. To better convey the human factors context of the research project, Chapter 1 provides an overview of the background, hypotheses and aims of the studies. Chapter 6 then discusses the broad implications of the findings, in addition to limitations and potential avenues for future research.

Chapter 1

Expert decisions

In domains such as healthcare, many key decisions are based on expert opinion, founded on what individuals know from their training, practices and experience. Expert knowledge is typically considered to be the best source of information, especially in unfamiliar situations where definitive data is not available (McBride & Burgman, 2012; Mumpower & Stewart, 1996). However, relying on expert opinion can be problematic. First, expertise is typically confined to a narrow field: outside of this, experts are subject to the same limited reasoning processes as 'non-experts'(McBride & Burgman, 2012). Experts can also be susceptible to cognitive biases even when operating within their area of expertise. Overconfidence, for example, can lead to poor decision-making (Phua & Tan, 2013). An overconfident expert may not recognise the uncertainty in their knowledge about a variable, and could fail to account for relevant and pertinent information (McBride & Burgman, 2012). There is a tendency for people to perceive their performance according to predetermined ideas about their abilities. This can limit personal insight, where individuals' perceptions of their competence fail to correlate with their actual performance (Dunning, Johnson, Ehrlinger, & Kruger, 2003). Cognitive biases such as this can be amplified in the absence of quality feedback. If expertise is acquired in an environment where mistakes are often not immediately costly (at least from the perspective of the expert) and feedback is not fast or frequent, then individuals can form inaccurate beliefs about their judgments (McBride & Burgman, 2012). Finally, experts do not always agree with one another. Although there can be many reasons for expert disagreement, it is often simply attributable to the different ways in which individuals think about a problem. Novel scenarios can facilitate varied judgments and conclusions from even the most competent and critical experts (Mumpower & Stewart, 1996).

Within the healthcare domain, one important area that has traditionally been guided by expert opinion is the design of paper-based hospital charts. Hospital chart design has typically been based on the intuition and clinical experience of staff at individual institutions or health services who have been perceived as having some knowledge of chart design (Chatterjee, Moon, Murphy, & McCrea, 2005; Knight, Calvesbert, Clarke, & Williamson, 2002; Preece, Hill, Horswill, & Watson, 2012b; Zeitz & McCutcheon, 2006). Similarly, in assessing the efficacy of individual chart designs, health professionals have tended to rely on their own (and their colleagues') subjective judgments. This is problematic for several reasons.

First, a clinician's realm of expertise is often limited to their area of medical specialisation. Health professionals do not typically receive training in how to design or redesign hospital charts,

thus their input is primarily focussed on the clinical aspects of the design rather than usability. Second, health professionals may also be overconfident in the design process, failing to account for critical information that lies outside of their field of knowledge (for example, important design techniques from the cognitive engineering domain). This overconfidence may be reinforced by a common perception in healthcare that the more experienced the clinician, the more expert they are. This possibility is especially concerning in light of empirical evidence which suggests that, after the first couple of years, clinicians' level of performance typically does not appear to improve with experience (Ericsson & Ward, 2007). Systematic reviews (Choudhry, Fletcher, & Soumerai, 2005) have demonstrated that, in many cases, performance can decline with greater experience, especially in the absence of continued training (Ericsson & Ward, 2007).

Third, clinical staff are also unlikely to receive accurate feedback about the effect of their chart design on staff performance or clinical outcomes; especially regarding specific design elements. This lack of quality feedback may lead chart designers to acquire inaccurate beliefs about their designs, which may carry over to future iterations. Even when more formal chart design methodologies have been adopted, feedback has been almost entirely subjective. For example, in developing a paper-based chart for monitoring critical care patients in a hospital accident and emergency department, Knight et al. (2002) employed four nurses with varying levels of seniority. Based on their clinical experiences and observations, the group created a list that prioritised what were regarded as essential and desirable chart elements and arranged chart sections into various layouts to enable discussion of new ideas. In an attempt to address and resolve potential pitfalls, hard copies were placed in the hospital staff room for comments from health professionals at all levels. Without trialling the new design, either experimentally or in practice (reportedly due to issues involving chart printing), Knight et al.'s (2002) chart was approved and introduced into the department. The problem with this approach to chart design is that, without any objective evidence to support the efficacy of the end product, its implementation could lead to more errors and time delays compared to the chart that it is replacing.

Finally, expert clinicians do not always agree on chart design decisions. Conflict can arise, for example, when chart designers disagree over what particular design elements can be regarded as best practice. This thesis focuses on major points of contention that surround the design of a ubiquitous and critically important paper-based hospital chart: the general observation chart.

Observation charts

General observation charts, which are traditionally kept at the end of a patient's bed during their stay, are used primarily to document physiological vital signs (Nwulu, Westwood, Edwards,

Kelliher, & Coleman, 2012). Observation charts usually incorporate observations for respiratory rate, oxygen saturation, blood pressure, heart rate, temperature and level of consciousness (ACSQHC, 2010; Lockwood, Conroy-Hiller, & Page, 2004) (for the sake of brevity, this thesis will use the term 'vital signs' to also encompass other physiological parameters that typically feature on observation charts; for example, oxygen flow rate). These vital signs are the typical predictors of adverse outcomes in medical admissions (Bright, Walker, & Bion, 2004; Buist, Bernard, Nguyen, Moore, & Anderson, 2004; Goldhill, White, & Sumner, 1999a).

Other clinical information can also be included on observation charts depending on the preferences of the institution (e.g., urine output and analysis, weight, blood sugar levels and pain scores) (ACSQHC, 2009). Vital sign observations are typically initiated when a patient is admitted to a healthcare facility to establish baseline data (ACSQHC, 2010) and are continued to monitor the patient's physiological condition at a frequency prescribed by a clinician or by hospital policy. This data can then be used to: (a) plan and implement appropriate interventions (e.g., medications); (b) evaluate a patient's response to interventions (e.g., before and after a surgical procedure); and, most importantly, (c) identify when a patient's general physical condition deteriorates (Koutoukidis, Stainton, & Hughson, 2012). It is important that health professionals have the means to accurately and efficiently identify clinical deterioration, as its prevalence within hospitals is widespread and increasing due to the changing characteristics of patients (e.g., ageing populations and an increased proportion of patients having complex medical issues) and healthcare systems (e.g., shorter hospital stays and increased bed occupancy) (ACSQHC, 2008; Bright et al., 2004; Johnstone, Rattray, & Myers, 2007; Robb & Seddon, 2010). Fortunately, observable derangements in vital signs often precede deterioration and therefore many serious adverse events are predictable (Buist et al., 2004; Goldhill, Worthington, Mulcahy, Tarling, & Sumner, 1999b; Jacques, Harrison, McLaws, & Kilborn, 2006; Kause et al., 2004).

For instance, in-hospital cardiorespiratory arrests have been shown to have markedly discernible clinical antecedents. One study that examined the observation charts of patients who experienced cardiorespiratory arrest in hospital found that 84% of the charts (n = 54) documented acute physiological deterioration within the eight hours prior to the arrest (Schein, Hazday, Pena, Ruben, & Sprung, 1990). Within-hospital deaths can also be associated with precursory derangements in vital signs. Hillman et al. (2001) found that around half of a sample of patients who died in hospital (whose mortality was not preceded by cardiorespiratory arrest or admission to an intensive care unit; n = 66) had serious vital sign abnormalities within eight hours of death, while one third had abnormalities within the preceding 48-hour period. Indeed, within emergency departments, 98% of deaths have been associated with abnormal vital signs or altered levels of consciousness for a significant period leading up to the point of death (Roller, Prasad, Garrison, &

Whitley, 1992). These studies suggest that observable derangements in vital signs can assist in the early recognition of patient deterioration, which may in turn minimise complications that may otherwise arise from delayed management and an inappropriate level of intervention.

Despite their clinical value, particularly as potential predictors of deterioration, vital signs are not always adequately measured and/or recorded. Consequently, deteriorating patients can go unnoticed (Leuvan & Mitchell, 2008; Odell, Victor, & Oliver, 2009). Failure to recognise and act upon deterioration even occurs when abnormal observations are documented appropriately (ACSQHC, 2008). Several studies have demonstrated that abnormal physiological values are often charted in the hours preceding cardiorespiratory arrest, unaccompanied by appropriate clinical action (Endacott, Kidd, Chaboyer, & Edington, 2007; Franklin & Mathew, 1994; Goldhill et al., 1999a). Many factors can contribute to health professionals' failures to recognise and respond to a patient who deteriorates. These include: (a) a poor understanding of why vital signs are measured; (b) limited knowledge of the symptoms and signs that can signal deterioration; (c) failures in communication (including uncertainty in whether it is appropriate to seek assistance); (d) unclear roles and responsibilities; and (e) inadequate skills and expertise (Cioffi, Salter, Wilkes, Vonu-Boriceanu, & Scott, 2006; Endacott et al., 2007; Robb & Seddon, 2010). However, another potential key contributor is often overlooked: the design of the observation chart itself. Despite an increased focus in the literature on physiological predictors of deterioration and the response of health professionals (e.g., the effect of rapid response systems), there has been little empirical research to investigate the tools with which deterioration is detected (ACSQHC, 2008, 2009; Odell et al., 2009; Preece, Hill, Horswill, Karamatic, & Watson, 2012a). As previously mentioned, observation chart design choices are typically based on the intuition and clinical experience of staff at individual institutions or health services (Chatterjee et al., 2005; Knight et al., 2002; Zeitz & McCutcheon, 2006). This non-standardised, unempirical and subjective approach has led to redundancies in effort and considerable variation in chart design (Chatterjee et al., 2005; Preece et al., 2013).

In Australian hospitals, for example, observation charts can be classified into two main categories: those with a track-and-trigger system and those without (see Figure 1). Track-and-trigger systems combine the routine 'tracking' of vital sign observations with 'triggers' to prompt chart-users to act on deterioration according to pre-determined criteria (Gao et al., 2007). Charts that incorporate a track-and-trigger system can be subdivided by the type of alerting system:

> (a) single parameter systems, where vital sign observations are compared with a simple set of criteria and a response algorithm activates when any single criterion is met (e.g., the calling criteria for a Medical Emergency Team) (see Figure 2);

(b) multiple parameter systems, where response algorithms require more than one criterion to be met, or differ according to the number of criteria met;

(c) aggregate scoring systems, where weighted scores, assigned to physiological values, are compared with predefined trigger thresholds (see Figure 3); and

(d) combination systems, where single or multiple parameter systems are used in combination with aggregate weighted scoring systems (see Figure 4) (ACSQHC, 2009; NICE, 2007b).

Observation charts with a track-and-trigger system can also differ in the type of abnormality alert used. Some use grey shading or lines to indicate abnormal ranges or thresholds for abnormality. Others use different coloured areas on the chart to reflect levels of physiological abnormality that are linked to either weighted scores (in aggregate scoring systems) or specific triggers (in single or multiple parameter systems). Observation charts can also differ on a number of other factors including paper size, orientation, display format of vital signs, the use of colour to signal abnormalities, and the presentation of vital signs relative to one another (Preece et al., 2013).

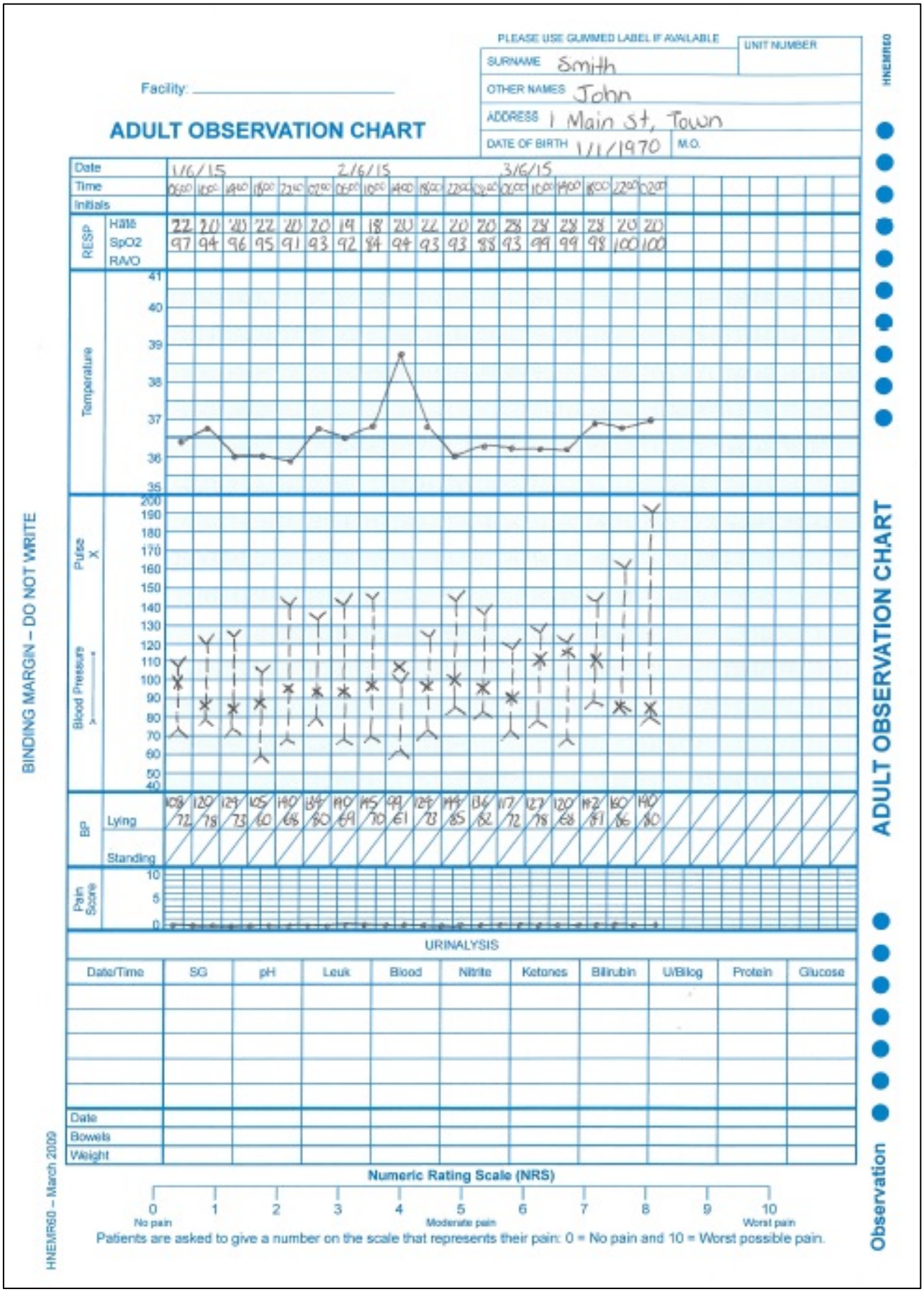


Figure 1. An existing chart (A4 size) without a track-and-trigger system. (All identifying markings have been removed.)

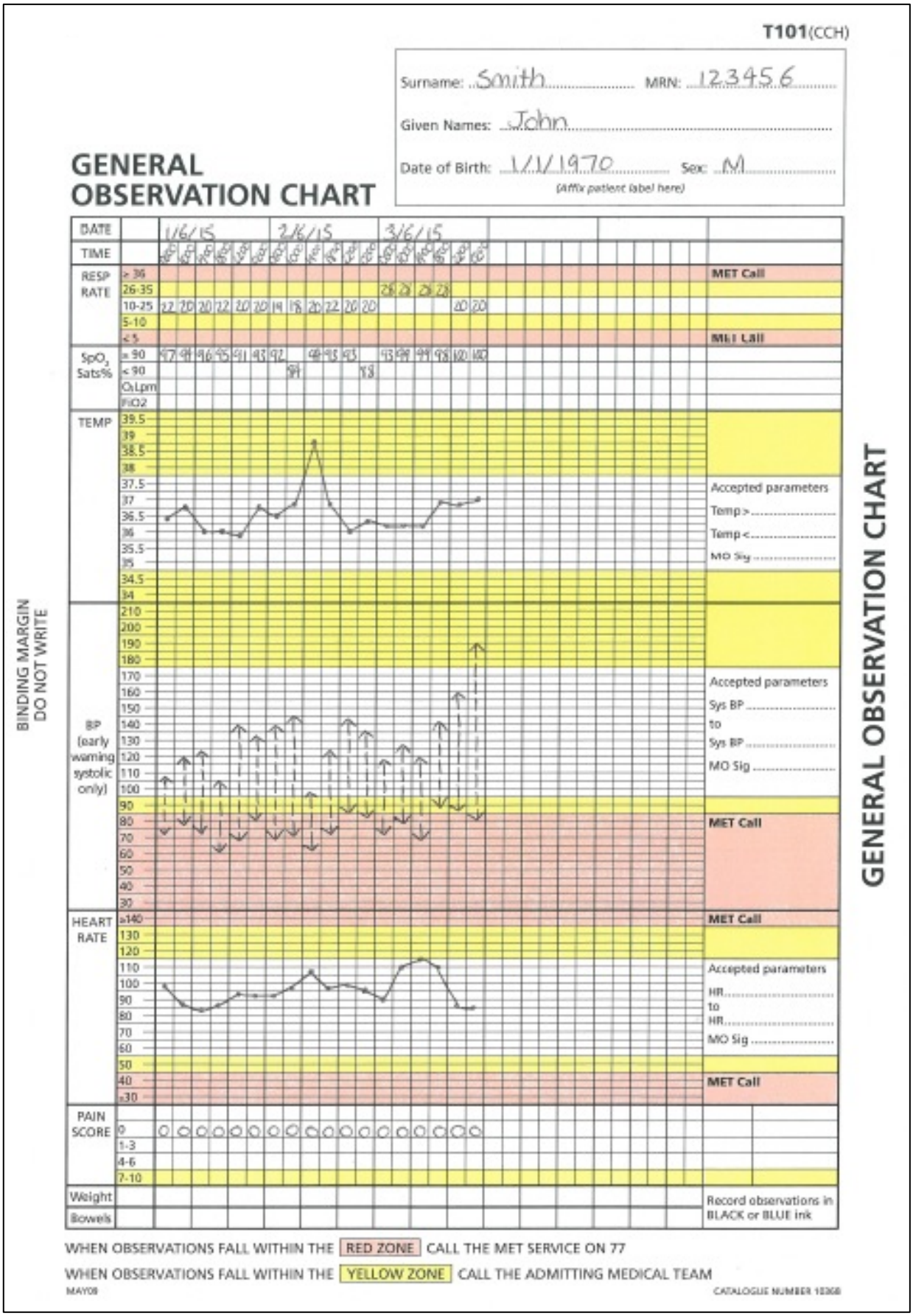


Figure 2. An existing chart (A4 size) with a single parameter track-and-trigger system, where vital sign observations are compared with a simple set of criteria and a response algorithm activates when any criterion is met. (All identifying markings have been removed.)

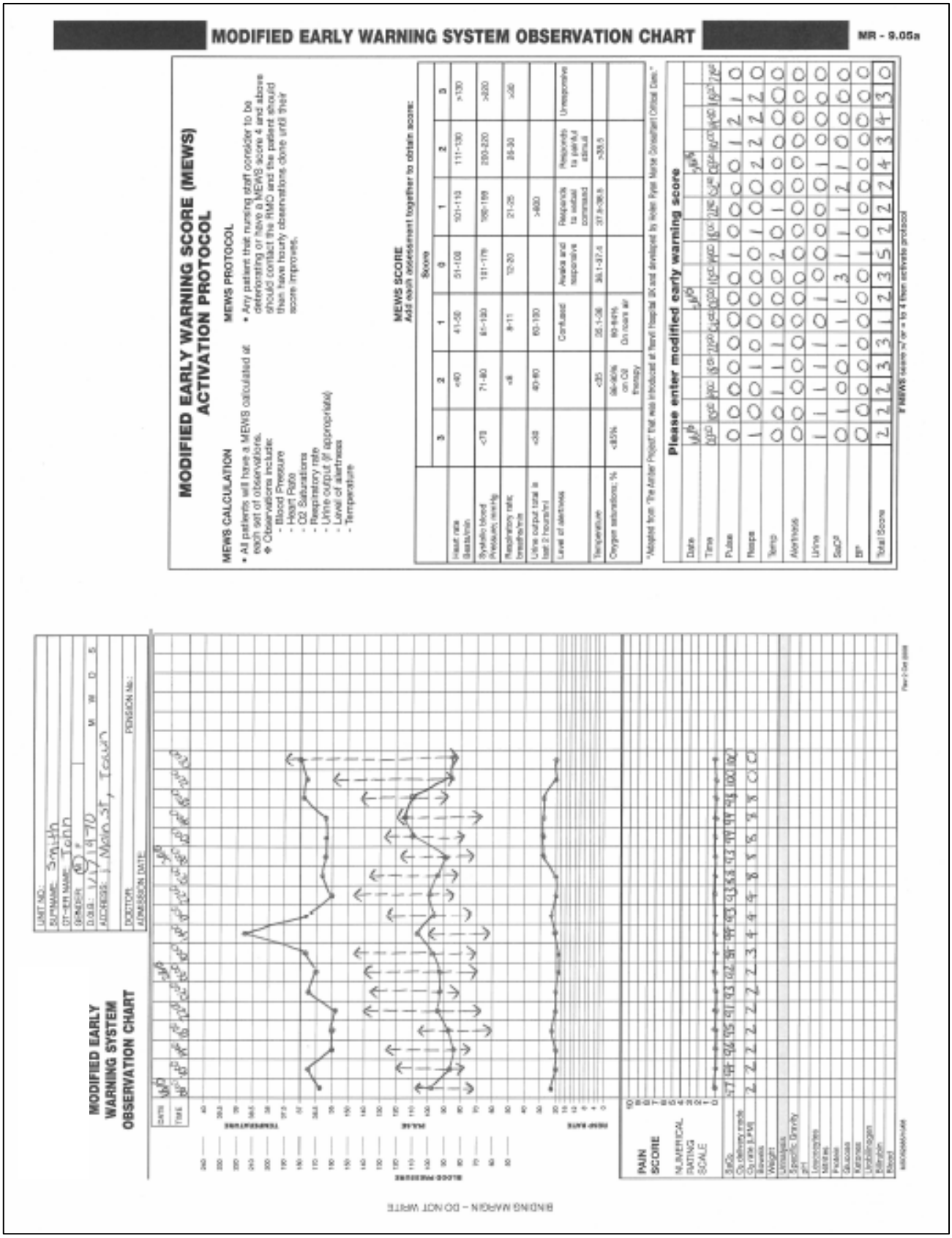


Figure 3. An existing chart (A3 size) with an aggregate scoring track-and-trigger system, where weighted scores are compared with predefined trigger thresholds. Note that the chart has been rotated 90 degrees to fit the page. (All identifying markings have been removed.)

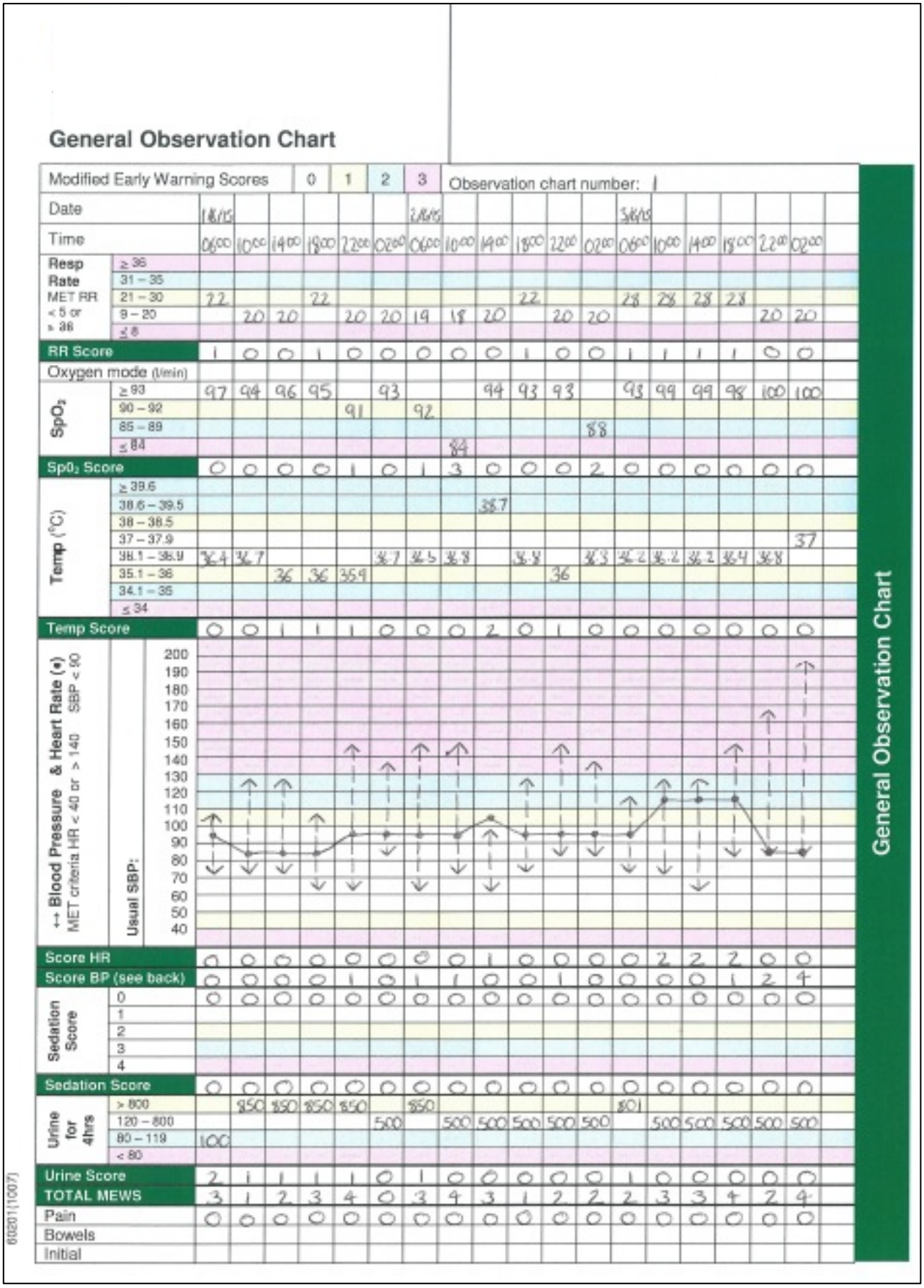


Figure 4. An existing chart (front page; A4 size) with a combination track-and-trigger system, where a multiple parameter system is used in combination with an aggregate weighted scoring system. (All identifying markings have been removed.)

The impact of design

Design variability raises the possibility that some charts might be more dangerous to patients than others. Indeed, after observing the concurrent use of five different chart designs within a 250-bed hospital, Chatterjee et al. (2005) hypothesised that the design of an observation chart could measurably influence its function. To test this, Chatterjee and colleagues evaluated the effectiveness of varying designs in clinical practice. Doctors and nurses (n = 63), who were presented with real physiological data recorded on the five existing charts, were asked to identify abnormal vital sign observations. After comparing detection rates across designs, the indicative results suggested potential benefits of particular chart design features. These objective experimental findings and the authors' subjective preferences guided the design of a new observation chart. Following the introduction of the novel design to hospital wards, re-evaluation revealed substantial improvements in the recognition of deranged respiratory rate, oxygen saturation, heart rate and temperature observations. Chatterjee et al. (2005) concluded that poor design can significantly undermine health professionals' recognition of clinical deterioration. Subsequent studies further demonstrated that chart redesign can significantly impact performance (Hammond et al., 2013; Kansal & Havill, 2012; Mitchell et al., 2010; Robb & Seddon, 2010). However, these studies are limited in that they do not systematically compare design features between novel and existing charts. Consequently, the authors are unable to attribute the superior performance of their novel chart to particular design decisions. Arguably, these studies are also limited in that each redesign team was comprised exclusively of health professionals. As previously mentioned, expertise is typically confined to a narrow domain. Outside their specific clinical area (e.g., nursing, critical care medicine), these experts' reasoning processes and judgments are subject to the same frailties as those of non-experts (McBride & Burgman, 2012).

In the redesign of an interface, a more appropriate expertise may be that possessed by individuals trained in systems design. For instance, a group of human factors researchers recently undertook a multiphasic project to develop an adult observation chart that supported the recognition of clinical deterioration (ACSQHC, 2008; Preece, Horswill, Hill, & Watson, 2010c). The research team created a new observation chart by amalgamating several sources of information including: (a) a set of design rules that they adapted from existing software and web design usability heuristics (Gerhardt-Powals, 1996; Nielsen & Mack, 1994; Zhu, Vu, & Proctor, 2005); (b) the design features that they judged to represent best practice in a heuristic evaluation of 25 existing charts (Preece et al., 2013); (c) the preferences of surveyed health professionals (n = 347) (Preece et al., 2012a); and (d) the design elements of a paediatric chart that was under development at the time (Horswill, Preece, Hill, Christofidis, & Watson, 2010; Preece et al., 2013). Two versions of the Adult

Deterioration Detection System (ADDS) chart were designed: one that allowed a patient's usual systolic blood pressure to be taken into account (see Figures 5 and 6), and one that lacked this facility.

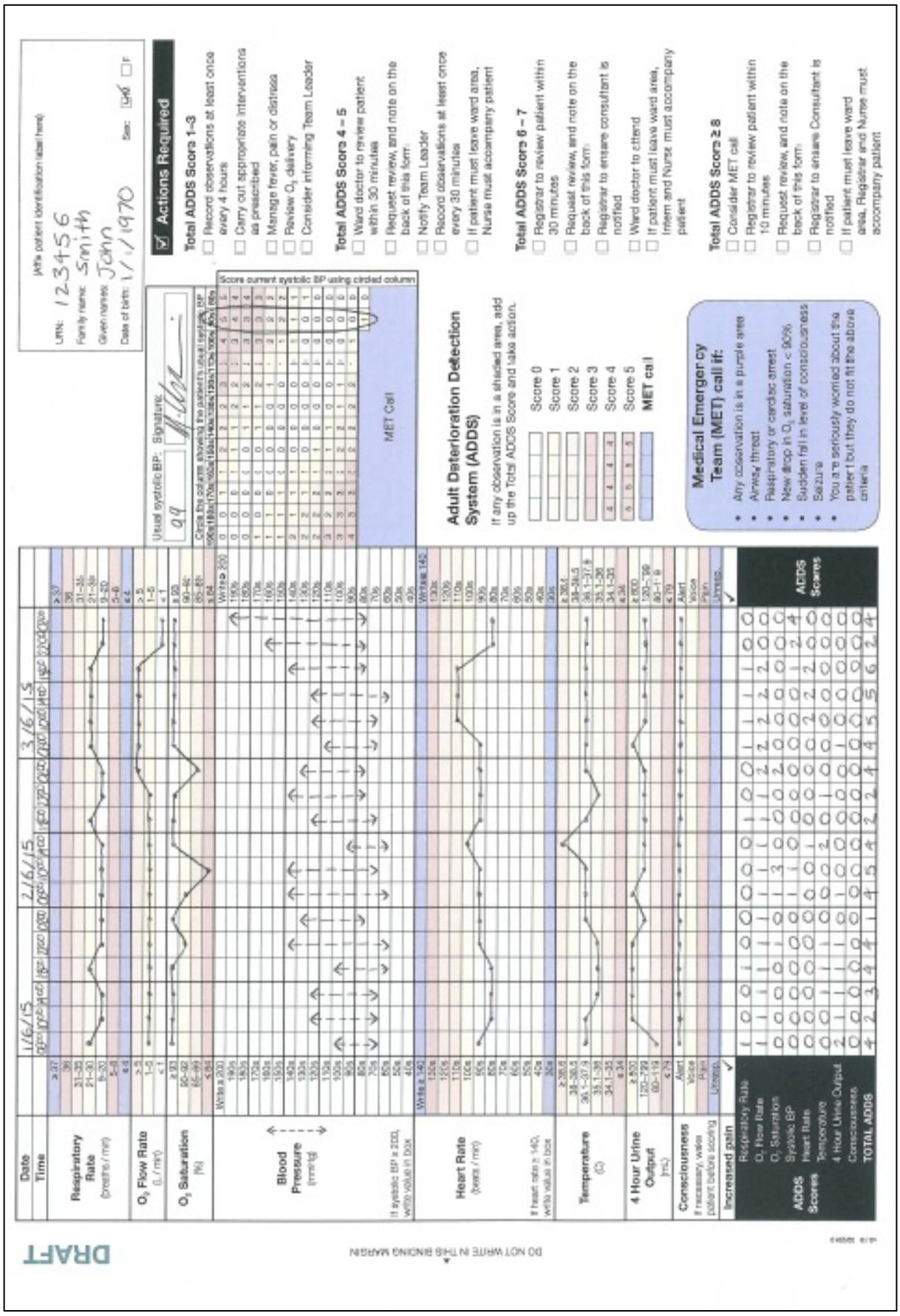


Figure 5. The inside page of the ADDS chart (A3 size) with a systolic blood pressure table. Note that the chart has been rotated 90 degrees to fit the page.

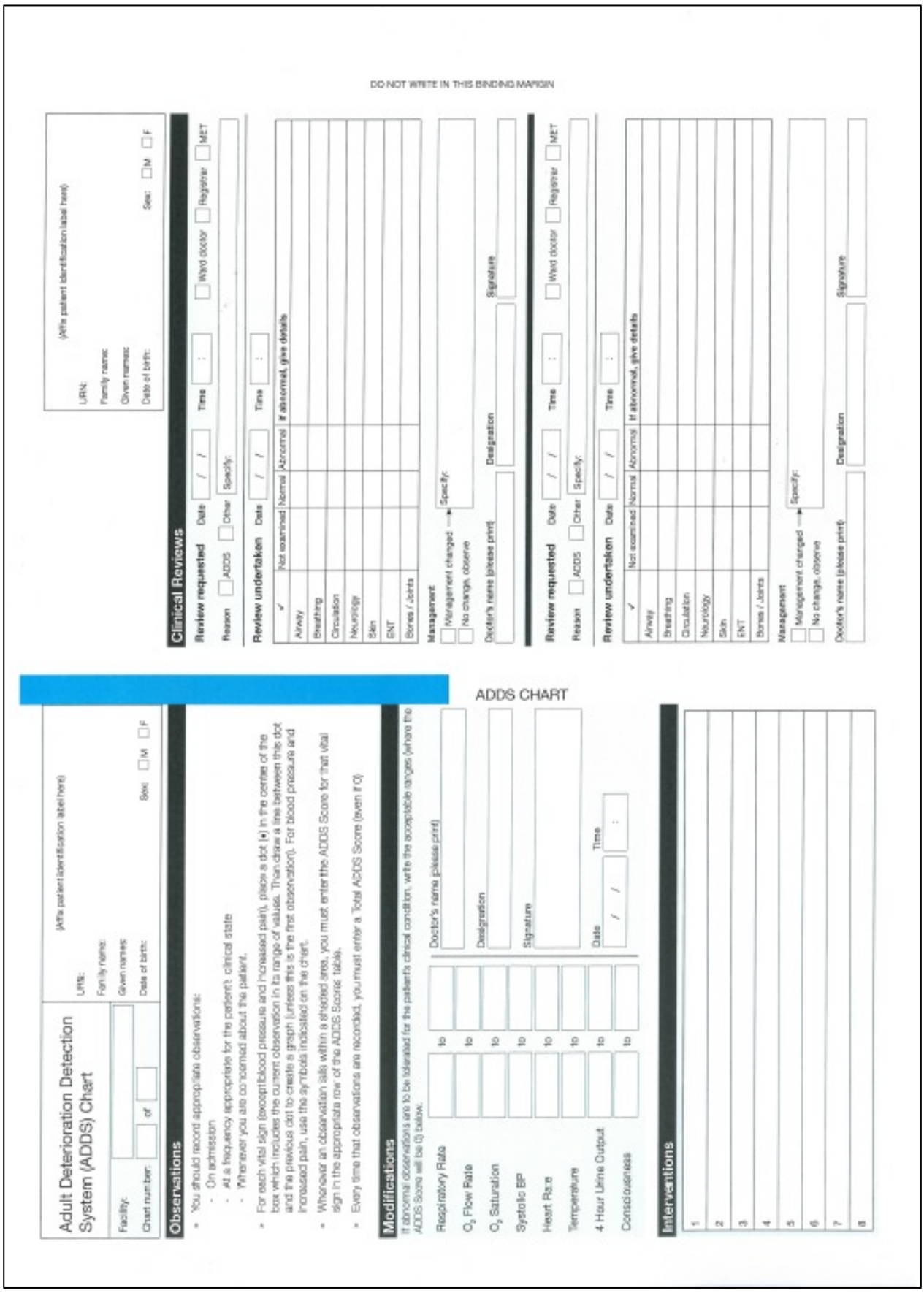


Figure 6. The outside page of the ADDS chart (A3 size). Note that the chart has been rotated 90 degrees to fit the page.

Preece et al. (2012b) assessed the performance of the ADDS chart designs in a behavioural experiment. Four existing observation charts, which had been identified in the heuristic evaluation (Preece et al., 2013), were selected for comparison. These were classified as being either reasonably well designed (n = 2), of average design quality (n = 1), or poorly designed (n = 1). Experienced health professionals (doctors and nurses; n = 45) and novices (individuals unfamiliar with observation charts; n = 46) each completed 48 trials, in which they were shown realistic patient data transcribed onto one of the charts. Each chart was used on eight trials, four times with normal data and four times with one abnormal vital sign observation (this could be either a derangement in oxygen saturation, blood pressure, heart rate or temperature). On each trial, health professionals and novices judged whether or not any of the vital signs were abnormal (all participants were required to memorise the normal physiological ranges for each vital sign prior to the task). For each of the six charts, two outcome measures were scored for each participant: error rate (the proportion of trials where the participant correctly identified a normal case or correctly indicated which vital sign was abnormal, as applicable) and response time (the average time taken to view the chart and make the judgment). The results revealed that both participant groups made significantly fewer errors and responded faster when using the ADDS charts versus the other designs, suggesting that observation chart design can significantly affect both health professionals' and novices' decision accuracy and response times in detecting deterioration. The findings also demonstrated that, in this instance, a chart designed by researchers with expertise in human factors performed better than several other charts that had been designed by teams of clinicians. A subsequent before-and-after evaluation of a version of the ADDS chart in a hospital setting also demonstrated a 45% reduction in the incidence of cardiac arrests (Drower, Mckeany, Jogia, & Jull, 2013). A later variation of the ADDS chart (the Q-ADDS form) was also found to reduce the severity of patient illness at admission to the intensive care unit as well as the average length of stay. This retrospective audit also revealed an 11% decrease in mortality amongst intensive care unit admissions (Joshi, Landy, Anstey, Gooch, & Campbell, 2014). These findings suggest that central to good observation chart design is an understanding of human limitations and affordances, and that this understanding is not necessarily intuitive.

Human factors design principles behind the ADDS chart design

In the design or redesign of an interface, human factors specialists can guide their processes using several sources of information. Some of these sources provide highly specific advice (e.g., data compendiums, industry standards, published empirical studies) that can be particularly advantageous when they are relevant to the given domain. However, in situations where relevant

standards or empirical findings do not exist, or for novel situations where existing standards or precedents are too domain-specific to solve a particular design problem, human factors experts must look to more abstract principles (Wickens, Lee, Liu, & Gordon Becker, 2004). The infancy of the human factors approach to patient chart design led the ADDS chart designers to adapt existing usability heuristics from the domains of software and web design: namely, those of Gerhardt-Powals (1996), Nielsen (1993) and Zhu et al. (2005). The following sub-sections describe the way in which the ADDS chart designers applied these usability heuristics (for a full description of each ADDS chart feature and the rationale behind its use, refer to Preece et al. (2013)).

Display information to match users' tasks

To give one example of a usability heuristic referenced by the ADDS chart designers, Nielsen (1993) recommended that interfaces should present pertinent information at the exact time and place where users need it. Matching the interface with the user's task was argued to minimise the need to search for information. In line with this principle, the ADDS chart includes succinct instructions on how to use the chart (e.g., when to measure vital signs, how to record observations, and how data relates to the track-and-trigger system) positioned on the outside front page as close as possible to the top of the page (see Figure 7). This design decision was made so that the instructions are available when a user first looks at the chart (Preece et al., 2010a), as English-reading people tend to search from top to bottom and left to right in organised visual spaces (Wickens et al., 2004).

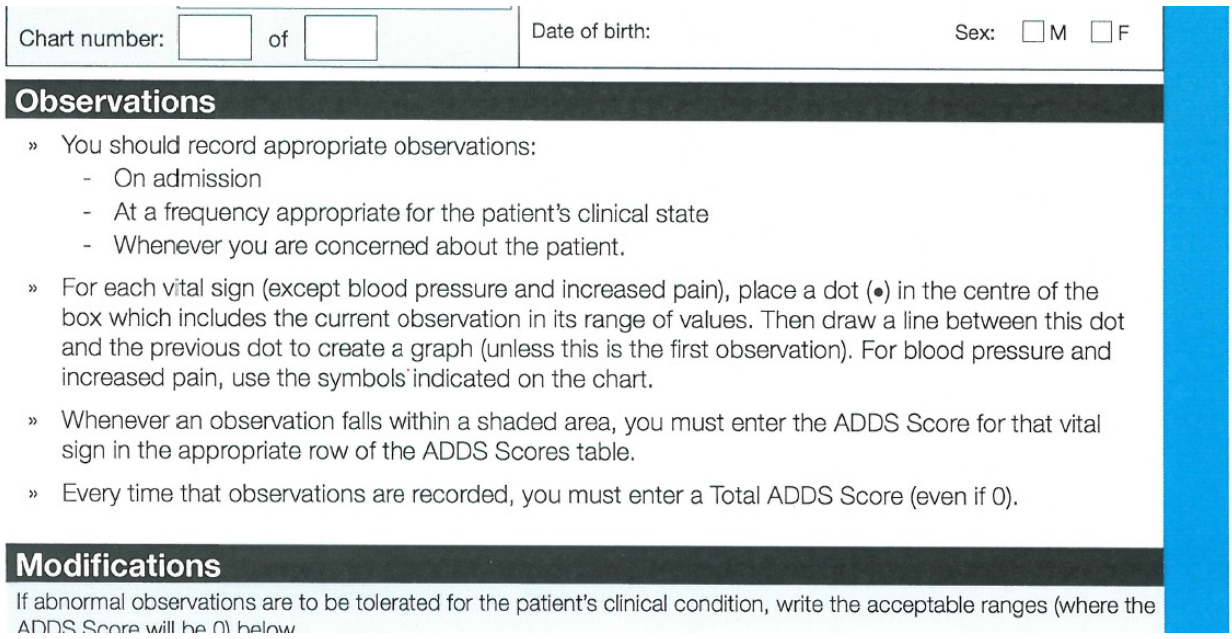


Figure 7. The ADDS chart positions general instructions near the top of the outside front page.

Nielsen (1993) further suggested that interface elements should be accessed in an order that maps on to the way in which a task will most effectively and efficiently be carried out. For example, to simplify a user's task, an interface can indicate a suggested sequence (e.g., the order implied by the listing of elements from top to bottom). Accordingly, the ADDS chart arranges vital signs according to their importance (see Figure 5). The most deterioration-relevant vital signs (e.g., respiratory rate) are placed where users will first look as a result of English reading conventions: towards the top left-hand side of the page (Nielsen, 1993). In contrast, urine output and pain are positioned towards the bottom of the chart, as they are comparatively less important for identifying potential deterioration (Preece et al., 2010a). The sequencing of vital signs is also logical from a clinical perspective. For instance, oxygen flow rate is contiguous with respiratory rate and oxygen saturation because an abnormal oxygen flow rate may indicate a deteriorating respiratory system when oxygen saturation sits within the normal reference range (Preece, Horswill, Hill, & Watson, 2010c). This also aligns with the next principle, to display information that will be used together close together.

Display information that will be used together close together

Gerhardt-Powals (1996) and Nielsen (1993) both highlighted the importance of grouping data meaningfully to decrease information search time. For example, information that will be used together (or that is contextually relevant) should be displayed close together, while contrasting information should be positioned with some separation. The combination track-and-trigger system of the ADDS chart, which comprises single and multiple parameter systems and an aggregated weighted scoring system (ACSQHC, 2009; NICE, 2007b) adheres to this principle in several ways. First, the single parameter system requires a Medical Emergency Team (MET) call when any individual observation is outside a given range, as indicated by purple range rows. The list of MET call criteria is perceptually linked to the purple range by being positioned adjacent to the vital sign recording area in a text box (that is also coloured purple to reinforce the perceptual link) (Preece et al., 2010c) (see Figure 8).

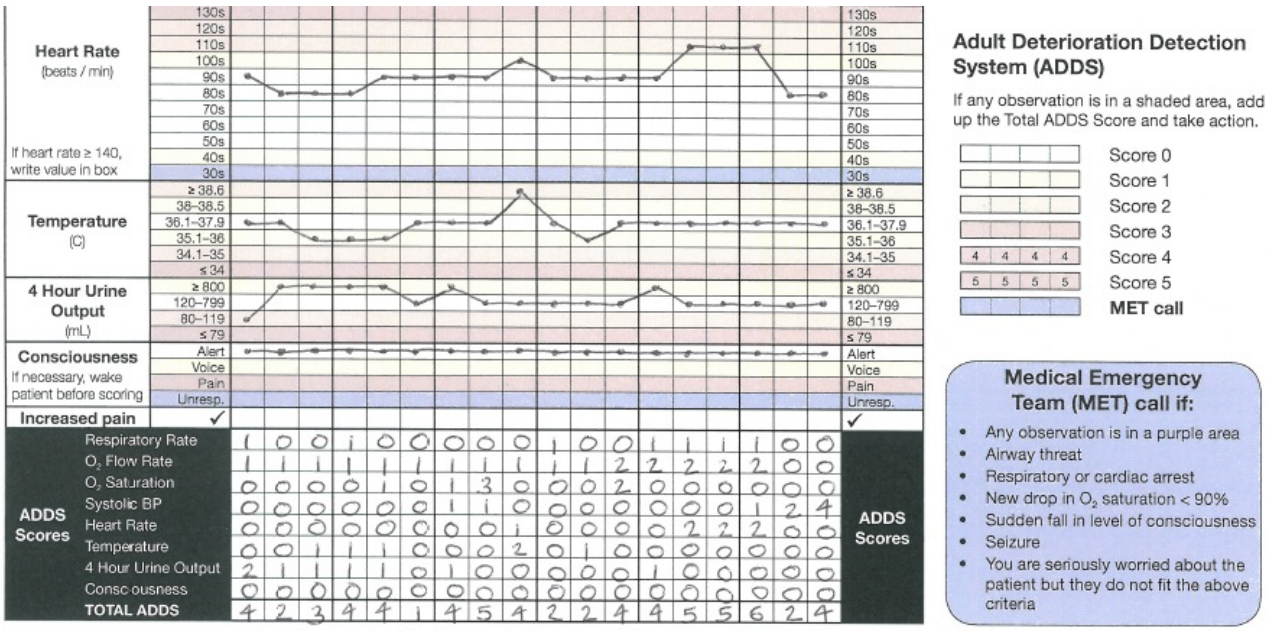


Figure 8. The ADDS chart positions the call criteria of the single parameter track-and-trigger system close to the vital sign observations.

Second, users can compare observations for each vital sign with a set of colour-coded criteria to determine whether any vital signs have reached predefined threshold levels of abnormality. The key for the colour-coded criteria is positioned adjacent to the vital sign data, so that users do not need to memorise this information (i.e., somewhat arbitrary pairings of colours with numbers) in order to use the chart successfully (Preece et al., 2010c) (see Figure 9).

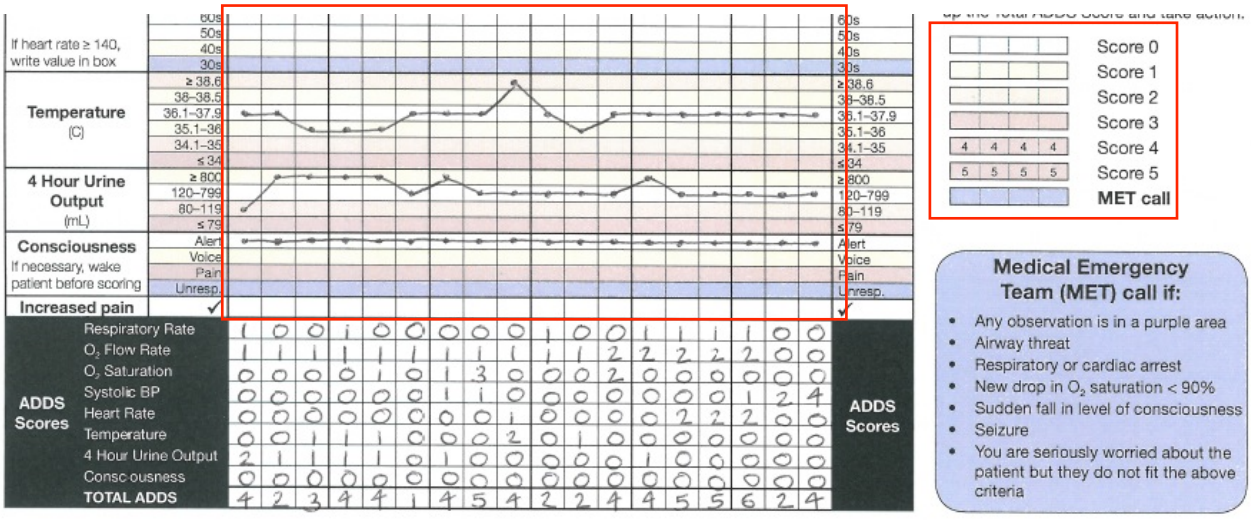


Figure 9. The ADDS chart positions the colour key of the multiple parameter track-and-trigger system close to the vital sign observations (highlighted by the boxed areas).

Third, the aggregate weighted scoring system incorporated into the chart involves assigning a Total ADDS Score to each set of vital sign observations. This score describes the patient's overall level of derangement across multiple vital signs. In this particular system, scores for each individual

vital sign are grouped together underneath the vital sign recording area (rather than, for example, being arranged separately immediately beneath the corresponding vital sign data). This allows users to: (a) independently assess the observations without interference from the early-warning scores, and (b) sum the early-warning scores without interference from the observations. Finally, the ADDS chart also includes a list of staff-initiated actions if a patient shows signs of deterioration. Again, because users need to correlate the total summed early-warning score with the list of actions, these two features are positioned as close together as possible (Preece et al., 2010a) (see Figure 10).

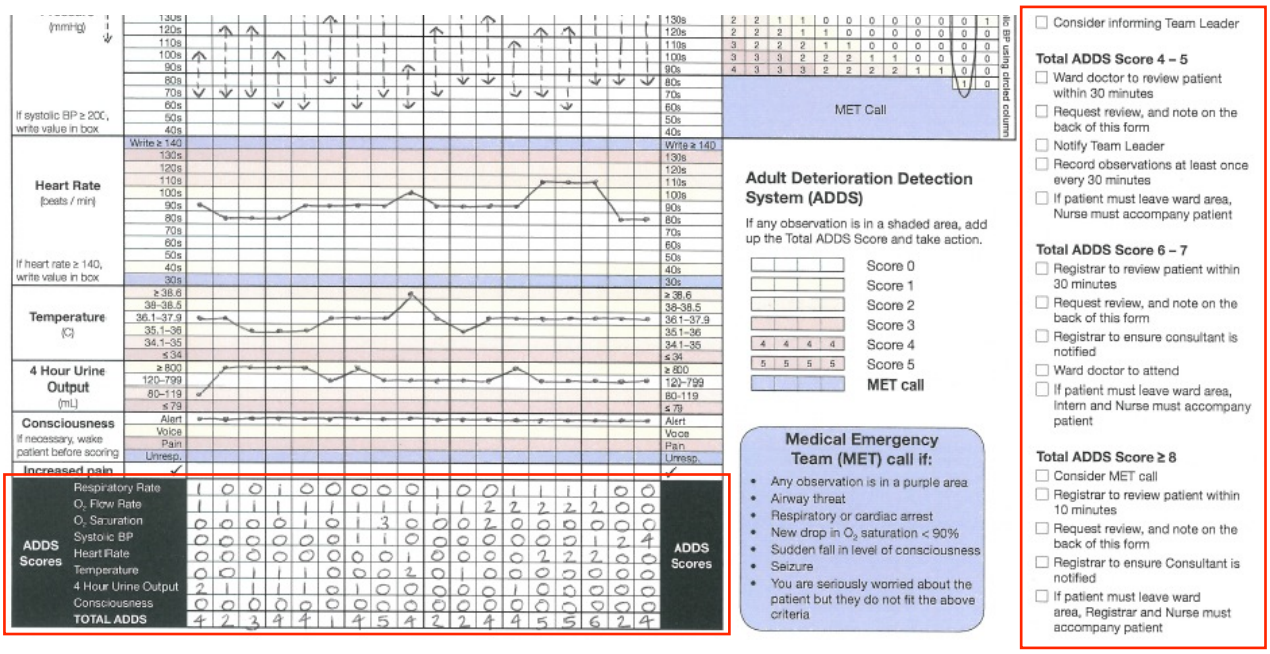


Figure 10. The ADDS chart groups scoring-rows together at the bottom of the page, and positions the row of summed early-warning scores close to the list of staff-initiated actions (highlighted by the boxed areas).

Minimise users' cognitive load

Another relevant human factors principle is to design the system to minimise users' cognitive load (Gerhardt-Powals, 1996; Nielsen, 1993). One strategy for achieving this is through the use of colour in a display to direct users' attention to particular elements, increasing the efficiency with which appropriate information is identified (Karwowski, 2006). For instance, the ADDS chart includes coloured range rows to signal the points at which observations cross particular thresholds of abnormality (Preece et al., 2010a). For example, if a patient's heart rate is 108 beats per minute, the recorded value will fall within the yellow reference range row, allowing users to immediately recognise that the observation is abnormal (rather than having to, say, recall the normal range from memory or look it up in a table). That is, the cognitive workload to detect an abnormality is reduced (Nielsen, 1993). This helps to reduce the need for mental comparisons and unnecessary thinking (Gerhardt-Powals, 1996). The ADDS chart's use of drawn-dot vital sign observations also

eliminates a potential source of unwanted workload (e.g., by preventing automatic reading of written number observations and/or preventing the comparison of numerical observations with clinical criteria stored in memory), freeing users' cognitive resources for higher-level tasks (Gerhardt-Powals, 1996) such as overall diagnostic evaluation of the patient (see Figure 11).

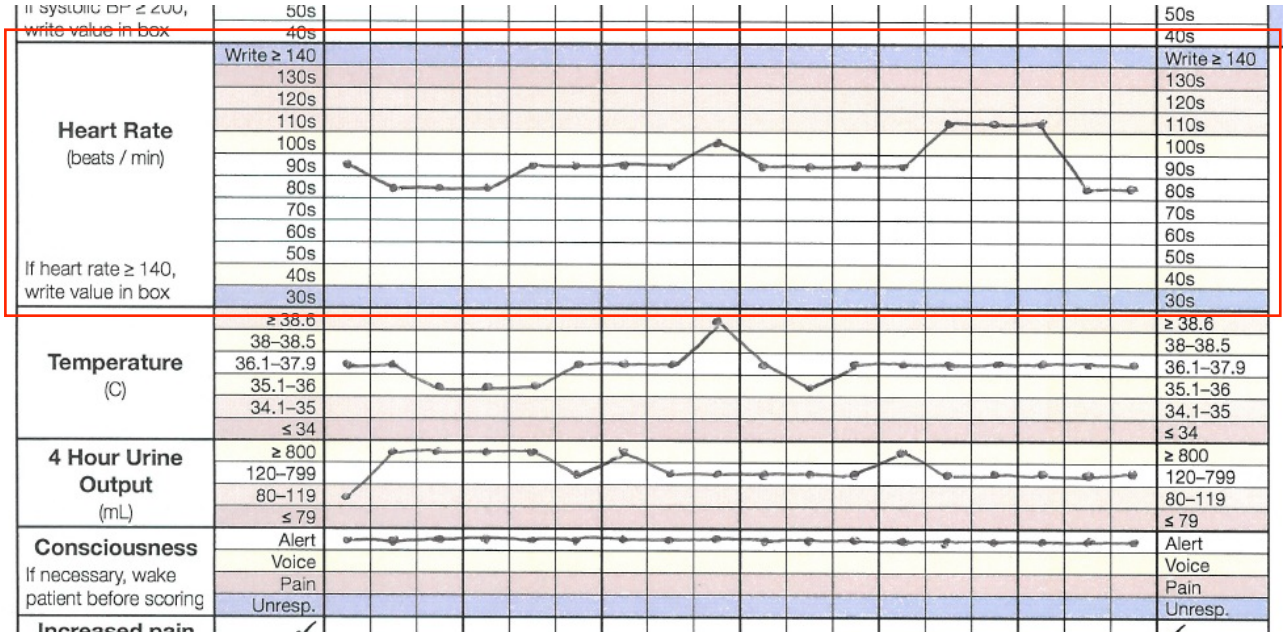


Figure 11. The ADDS chart uses vital signs graphs with drawn-dot observations, where thicker horizontal lines separate adjoining vital sign graphs (highlighted by the boxed area).

Limit data-driven tasks

It has been proposed that data-driven tasks should be limited in time-critical environments with high information loads (Gerhardt-Powals, 1996). The presence of an integrated colour-based system and the use of drawn-dots also ensures that the task of searching for abnormal observations is not unnecessarily data-driven, potentially reducing the time that chart-users need to spend assimilating raw vital sign observations. Drawn-dots (in contrast to written-numbers) also make it easier to detect trends in the data, especially if consecutive data points are connected with lines (see Figure 11) (Wickens & Hollands, 2000), without imposing costs on focused attention (Salvendy, 1997).

Display relationships

The relationships between elements on (in this case) a patient chart can be highlighted using principles of graphic structure. For example, it has been proposed that items can be seen as belonging together if they are closely positioned, are enclosed (e.g., with boxes or lines) or look similar (Nielsen, 1993). The design of the ADDS chart involves applying all of these perceptual grouping strategies to help users understand the structure of the chart. First, adjoining vital signs are

visually separated by thicker horizontal lines (Preece et al., 2010c) to ensure that unrelated elements do not appear to belong together (Preece et al., 2013). Second, the date row is ruled off every 24 hours, signaling the separation of date information into days (Preece et al., 2010c) (see Figure 12).

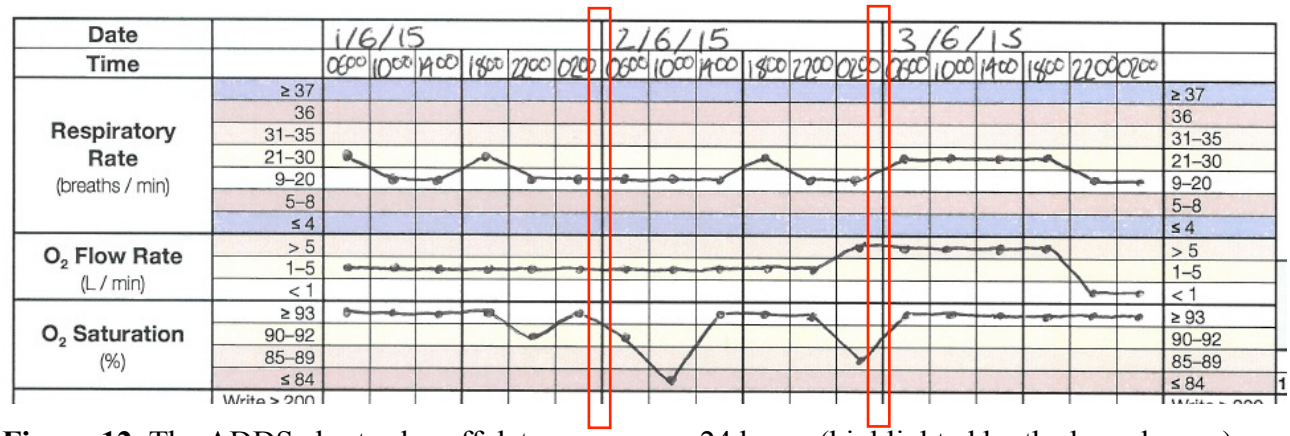


Figure 12. The ADDS chart rules off date rows every 24 hours (highlighted by the boxed areas).

To mitigate 'column shift' errors, where chart-users enter or read data from the wrong time-point column, thick vertical lines are placed after every three columns (Preece et al., 2010c). This was designed to facilitate easier tracking, by making adjacent columns more visually distinct (with either a thick line on the left, or on the right, or no thick line at all). As a result, the columns on either side of the column for any given time-point will appear visually dissimilar (Preece et al., 2013) (see Figure 13).

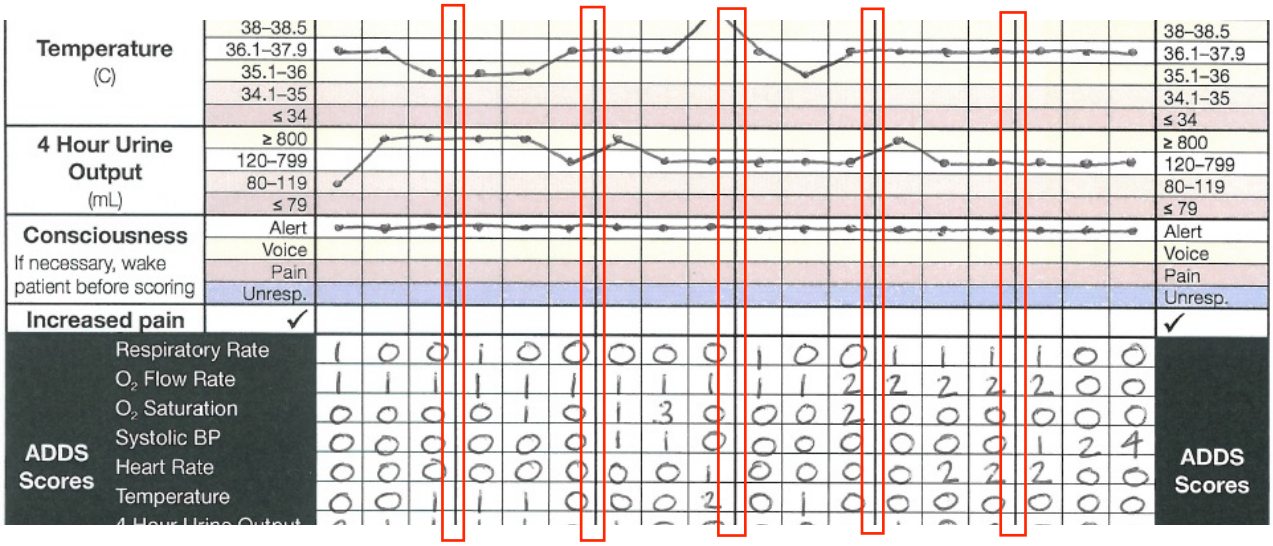


Figure 13. The ADDS chart uses thick vertical lines after every three time-point columns (highlighted by the boxed areas).

Use colour appropriately

Another human factors principle is to limit an interface to no more than 5 to 7 consistently applied colours (Nielsen, 1993). If colour is unrestrained within a display, it can create clutter and

increase visual search times (Karwowski, 2006). Accordingly, only five background colours (including white space) are used in the ADDS chart's observation area (Preece et al., 2010c) to reduce the risk of a visually cluttered display. These background colours are light pastel shades to ensure that observations written into the coloured cells are easily visible under a wide range of lighting conditions (Preece et al., 2013). Because colour can also be used to rank order items on a scale (Karwowski, 2006), the ADDS chart uses an intuitive progression of colour densities that correlates with the level of physiological derangement (Preece et al., 2010a). This feature also provides a redundant cue that is of particular use to colour-blind users (Preece et al., 2013).

Speak users' language

It has been proposed that terminology (e.g., words, abbreviations, and icons) should take into account users' existing vocabulary and understanding (Nielsen, 1993). Preece et al. (2012a) collected quantitative data on experienced chart-users' preferred terminology, which subsequently influenced the research team's design decisions. Three hundred and forty-seven health professionals (approximately two thirds of whom reported using observation charts daily) responded to an online survey that featured questions about the design of charts, including the comprehensibility of common abbreviations and preferred vital sign terminology. For example, the ADDS chart designers adopted the label 'O_2 Flow Rate' (over 'O_2 Delivery') because it was preferred by more surveyed health professionals (Preece et al., 2010c).

Maintain consistency

To facilitate recognition, it has been proposed that consistent formatting should be used for similar information (Nielsen, 1993). The ADDS chart uses the same formatting for labels of the same level of importance (e.g., the label for each vital sign) to avoid related elements appearing as if they belong to different categories (Preece et al., 2013). Accordingly, different formatting is used for unrelated elements. For instance, graph labels are formatted differently from their corresponding vertical axis scales (Preece et al., 2010c).

Contentious design decisions

Human factors design principles, including those of Gerhardt-Powals (1996), Nielsen (1993) and Zhu et al. (2005), are intended to act as guides rather than hard and fast rules (Proctor & Van Zandt, 2008; Wickens et al., 2004), and experienced designers are expected to carefully consider how to apply principles with regard to context (Nielsen, 1993; Wickens et al., 2004). However, this purposeful abstraction means that principles can sometimes conflict, even for skilled interface

designers. For example, a small display, created to minimise the effort that users exert to access information, can lead to a display that is less legible. Similarly, a display with redundancy, included to increase the chance that a message will be interpreted, can lead to a display that is visually cluttered (Wickens et al., 2004). In situations where human factors design principles clash (i.e., when certain principles support one design option, and other principles support another), there is often no simple resolution as to which principle(s) should take precedence. In other cases, a single principle can be implemented in multiple ways and guidelines to direct a decision between alternatives may not exist (Wickens et al., 2004). Consequently, interface designers cannot mechanistically apply general human factors design principles to determine good design.

The potential conflict between principles also means that, after an interface has been designed, certain display features can still be contentious from a human factors perspective. For example, although the ADDS chart has demonstrated significant benefits within laboratory and clinical settings, there are several points of contention as to whether this chart can be regarded as best practice. Clinicians have specifically questioned whether the novel ADDS chart design: (1) will remain beneficial to users highly experienced with alternative chart designs; (2) should present blood pressure and heart rate graphs as separate plots; (3) should use drawn-dot observations, an integrated colour track-and-trigger system, and grouped scoring-rows to support users; and (4) adopts a design layout that best facilitates users' calculation of summary scores. As with previous work (Chatterjee et al., 2005; Hammond et al., 2013; Kansal & Havill, 2012; Mitchell et al., 2010; Robb & Seddon, 2010), the study by Preece et al. (2012b) compared charts that varied on more than one dimension. This means that, while Preece et al.'s results support the efficacy of the ADDS chart's overall design, they do not constitute evidence to support any of the specific design decisions employed.

In the absence of expert consensus and objective evidence, this thesis seeks to resolve some of these points of contention through the use of behavioural experiments.

Will the novel design benefit users highly experienced with alternative chart designs?

The first point of contention relates to whether the ADDS chart remains beneficial to users highly experienced with alternative chart designs. When designing or redesigning an interface, the end-user is of paramount importance. Particular attention should be paid to their individual characteristics, such as age, work experience, education level and familiarly with existing systems (Drews & Kramer, 2012; Kalyuga, Ayres, Chandler, & Sweller, 2003). This is especially germane to hospital settings, which are inhabited by large and diverse groups of health professionals.

Observation charts, for example, can be used by the full spectrum of healthcare staff, from enrolled nurses on their first ward rotation to senior specialists.

Novice users

Essential to chart design is consideration of the novice user. Every year, cohorts of graduate nurses and doctors enter the healthcare system where they use general observation charts for the first time. It has been proposed that one way in which novices differ from more experienced individuals is that they lack task-relevant schemas (Kalyuga et al., 2003); in the case of this thesis, newly qualified health professionals do not have schemas associated with chart-related tasks. Schemas, which are mental constructs that can reduce cognitive load, permit the organisation of multiple sub-elements of information as a single entity in working memory. Without relevant schemas, novices are more limited by the capacity of working memory, which can only handle a few elements of information at a time. When a system fails to provide guidance for dealing with new units of information, novices can experience cognitive overload. Cognitive overload can be defined as when the requirements for a particular cognitive task exceed the capacity of an individual's working memory (Kalyuga, 2007; Kalyuga & Renkl, 2010; Oksa, Kalyuga, & Chandler, 2010; Salden, Aleven, Schwonke, & Renkl, 2010; Schnotz, 2010; van Gog, Ericsson, Rikers, & Paas, 2005).

To illustrate this point, consider how a novice user might detect deterioration using a chart (without a track-and-trigger system) where observations are recorded as written numbers within a table (i.e., where each column represents a different vital sign and each row corresponds to a time-point) (see Figure 14). This type of design, used in some Australian hospitals (Preece et al., 2013), requires domain-specific knowledge such as the normal and abnormal reference ranges for each vital sign, the clinical relevance of the degrees of abnormality, as well as the significance of trends in the data (which hinges on the ability to decipher the trends). Each of these tasks necessitates substantial conscious effort that could potentially overload users' working memory. For instance, to assess the (largely interdependent) relationship between blood pressure and heart rate (columns 'BP' and 'P' in Figure 14, respectively), a novice chart-user would need to: (a) find the appropriate recorded heart rate observation; (b) determine if the given heart rate observation fell out of the normal reference range by retrieving it from memory and comparing; (c) hold this judgment in working memory; (d) find the corresponding blood pressure observation (i.e., recorded at the same time-point as the heart rate observation); (e) decide if the blood pressure observation was out of range, again by retrieving the normal physiological reference range from memory and comparing; and (f) assess the relationship between heart rate and blood pressure in the clinical context of the patient's condition. If the relationship required continuous monitoring, the novice would need to

hold each consecutive comparison in memory as they worked through the time-points, progressively increasing the load on working memory. Tabular displays may also require chart-users to mentally visualise recorded observations in a graphical format to detect trends in the data (Preece et al., 2013). These kinds of cognitive demands can lead to error, as can improper simplifications that may result from a lack of experience (Proctor & Van Zandt, 2008).

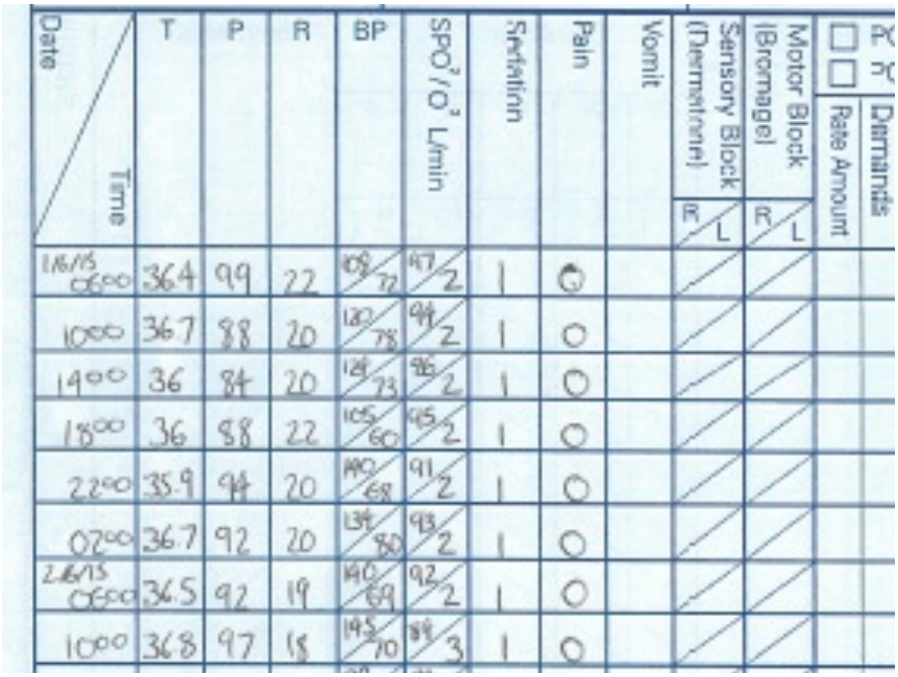


Figure 14. An extract of an existing chart with a tabular display of data.

The ADDS chart was designed with the explicit aim of facilitating chart-users' detection of patient deterioration in a user-friendly way (Preece et al., 2012b). For example, as previously discussed: (a) chart instructions are situated towards the top of the outside front page so that they are available when a user first looks at the chart; (b) the most important vital signs are positioned towards the top left-hand side of the chart, where users are likely to first look; (c) components of the combination track-and-trigger system are displayed close together so that users do not need to search extraneously for information; (d) colour-coded reference range rows allow users to recognise abnormal observations, rather than having to remember the normal reference ranges for each vital sign; and (e) drawn-dot observations prevent users from automatically reading the numbers and/or comparing them with clinical criteria stored in memory. Although these design features were utilised to assist all chart-users, they may particularly help inexperienced users by acting as a substitute for novices' missing schemas. Indeed, these design techniques may have contributed to the superior performance of the ADDS chart among novice chart-users in the empirical study by Preece et al. (2012b). In contrast, the poorest performing charts may have led novices to engage in cognitively inefficient problem-solving strategies that imposed a heavy working memory load.

Experienced users

Unlike novices, experienced chart-users are able bring acquired schemas, held in long-term memory, to a task. The idea is that implementing a schema requires substantially less working memory capacity than individually implementing the many lower-level elements that it incorporates, thus mitigating the processing of overwhelming amounts of information. Although conscious effort is required to control the use of schemas, after enough practice they can operate more automatically. Consequently, experienced users are able to circumvent the limitations of working memory capacity (Kalyuga et al., 2003). Acquired schemas brought to a task (in this case, the task of detecting deterioration) may facilitate higher-level strategies by experienced chart-users, relative to novices' piecemeal approach (Gerhardt-Powals, 1996). Returning to the example of the tabular chart (see Figure 14), it could be the case that highly experienced users can access the memorised normal ranges required to detect deterioration with sufficient ease that they can perform to the same level with these charts as with the ADDS chart (that is, it is possible that they would gain little advantage from the load-reducing strategies employed by the ADDS chart). For instance, with a tabular chart (Figure 14), experienced users may be able to visualise the ebb and flow of observations down each vital sign column while concurrently noting observations that are of clinical concern. Even across time-point rows, experienced users may be able to assess a patient's condition in a broader sense (i.e., consider the relationship between all vital signs at a particular time). Indicative findings support the idea that there might be some distinction between novice and experienced chart-users. Preece et al. (2012b) found that health professionals detected deterioration significantly faster than novices using designs without track-and-trigger systems (i.e., where participants had to rely on their memory of normal vital sign ranges).

The potential for performance differences between novice and experienced users became a point of contention during the development of the ADDS chart. Some clinicians expressed concerns that, although the design may advantage novices, it might be problematic for experienced clinicians who are accustomed to other chart formats. This anecdotal concern may be consistent with findings within the human factors literature; specifically, the 'expertise reversal effect' where instructional approaches found to be ideal for novices are sometimes counterproductive for more experienced users. For novices, guidance provides users with information that explains the concepts and procedures that they need to learn, while using strategies that are compatible with their cognitive abilities and limitations (e.g., working memory capacity) (Kalyuga, Chandler, & Sweller, 1998; Kirschner, Sweller, & Clark, 2006). Thus, instructional guidance can act as a substitute for missing schemas. If effective, this guidance can even help to construct schemas (Kalyuga et al., 2003).

Although instructional design principles typically succeed at reducing novice users' extraneous cognitive load (Rey & Buchwald, 2011), the same principles may not be as helpful for experienced users. As previously mentioned, individuals with experience are able bring acquired schemas, held in long-term memory, to a task. If these users are unable to avoid attending to an interface's instructional information (which is often difficult to ignore), both schema-based and instruction-based guidance are available for dealing with the same material. Overlap ensues if users try to relate the corresponding components, which can lead to the recruitment of additional working memory resources and, potentially, cognitive overload (Kalyuga et al., 2003). Thus, a system high in instructional guidance may hinder experts' processing of information, relative to instruction that relies more on pre-existing schemas for direction.

The expertise reversal effect, extensively described within the instructional learning literature, has been observed across several domains and experimental conditions (Brunstein, Betts, & Anderson, 2009; Kalyuga, Chandler, Tuovinen, & Sweller, 2001; Kalyuga & Sweller, 2004; Kyun, Kalyuga, & Sweller, 2013; Nückles, Hübner, Dümer, & Renkl, 2010; Reisslein, Atkinson, Seeling, & Reisslein, 2006; Tuovinen & Sweller, 1999). For example, in a two-stage experiment, inexperienced trainees in electronics (n = 15) were found to perform significantly better using diagrams of electrical circuits that integrated textual explanations (where they were unable to understand the diagrams without the text). More experienced electronics trainees (n = 15), on the other hand, were found to benefit more from diagrams without explanation. These participants also reported expending less mental effort with the diagram-only format. This finding suggested that for the more expert trainees, the additional text was a redundancy that should be eliminated from the diagram (Kalyuga et al., 1998). Although interpreting an observation chart is not directly comparable to learning something through instruction, it may be possible that the same psychological processes operate. That is, a chart designed for novices (i.e., the ADDS chart) may not be ideal for experienced users who bring incongruent expectations and understandings to the novel chart.

Hypotheses and behavioural experiment (Christofidis, Hill, Horswill, & Watson, 2013)

Two competing hypotheses related to user experience are presented. First, it is proposed that user-friendly design will outweigh prior experience, such that chart-users will make fewer errors and respond more quickly when using the ADDS chart, compared to chart designs with which they are highly experienced. Alternatively, it can be proposed that chart-users will perform better when presented with the chart that they regularly use in their occupational role, demonstrating that prior experience outweighs design. Chapter 2 aims to resolve this contentious issue. Although the results of Preece et al. (2012b) already revealed that doctors and nurses (in addition to the aforementioned

novices) performed significantly better using the ADDS chart designs, almost all of these participants were experienced in using multiple charts of different design with varying levels of instructional guidance and, as a result, may have developed flexible and adaptive schemas. Arguably, it is more crucial to discern if the ADDS chart will disadvantage staff who are more likely to be subject to the expertise reversal effect: those with rigid and repetitive schemas that have resulted from extensive experience with a single chart design. To assess this, we recruited two groups of health professionals who were experienced with either a multiple parameter track-and-trigger chart or a graphical chart without a track-and-trigger system.

Should the novel design present blood pressure and heart rate observations as separate plots?

The second point of contention involves a particular aspect of the ADDS chart design: namely, whether blood pressure and heart rate observations should be presented as separate plots (as in the ADDS chart), or overlapping plots on the same axes (as is the case in many existing charts). As previously mentioned, human factors design principles are not concrete rules (Proctor & Van Zandt, 2008; Wickens et al., 2004). Conflict can arise when one principle can be applied in several ways and there are no guidelines to direct a decision between alternatives. In this instance, the principle of 'proximity compatibility' can be implemented in multiple ways and the existing literature does not allow a conclusion to be reached as to which application results in better performance in detecting patient deterioration.

Integrative processing

The widely documented proximity compatibility principle (Wickens & Carswell, 1995) proposes a relationship between two dimensions: processing proximity (the extent to which two information sources are used within the same task, e.g., compared or integrated) and display proximity (how close two display components are in a user's perceptual space) (Wickens & Hollands, 2000). The principle suggests that close processing proximity benefits from close display proximity. It is reasoned that when two sources of information are presented close together in space, their integration and comparison can be made easier with a reduction in visual search cost (i.e., the time users spend moving their attention from one source to the other) (Wickens & Carswell, 1995). When separated, users must retain the information relevant to one source (often by rehearsal), move their attention to access the second source, and then compare or combine the information. The time it takes to access the second source can degrade a user's memory for the first source: even more so if the second source is found within a cluttered field (Lee, Kirlik, & Dainoff, 2013; Wickens &

McCarley, 2007).

Of specific relevance to this thesis is the application of the proximity compatibility principle to graphical displays. Wickens et al. (2004) suggested that graphs which require the integration or comparison of components can benefit from being constructed close together in space because excessive visual search effort can hinder graph interpretation. The authors suggested that it can be advantageous, for example, to keep two graph lines on the same panel (rather than separate panels) if they require comparison. With regard to blood pressure and heart rate observations, a clinical rationale for plotting these two vital signs together in close proximity (i.e., on the same axes) (see Figure 15) relates to their interrelationship, where a decrease in blood pressure can lead to a reflex increase in heart rate (and vice versa). This relationship is sphygmoidal in nature: a small change in blood pressure can cause a large change in heart rate, within the responsive range of the physiological baroreceptor reflex (i.e., the steep portion of the curve) (Smith & Fernhall, 2011). This association, recorded on observation charts with overlapping blood pressure and heart rate graphs, may assist chart-users to detect deterioration faster and with less cognitive demand. For instance, a systolic blood pressure observation on the lower end of the normal reference range may signal users to examine this time-point more carefully, where they may in turn notice an abnormal heart rate observation. Overlapping graphs of blood pressure and heart rate is a design feature that was ubiquitous across Australasian hospitals at the time when the ADDS chart was developed and was preferred by health professionals (Preece, Horswill, Hill, Karamatic, & Watson, 2010b).

Emergent features

Close display proximity can yield another usability advantage. When multiple elements of a display are grouped together, a new feature can emerge that is not inherent in any of the elements themselves. Emergent features can benefit task performance because their salience facilitates more direct perception, allowing users to inspect a display globally rather than focussing on the individual parts. This can reduce the cognitive effort and attentional demands needed for a multi-element display (Lee et al., 2013; Proctor & Vu, 2006; Wickens & Carswell, 1995). The observation chart ‘Seagull Sign’, a visual cue that can occur when systolic blood pressure and heart rate are graphed as overlapping plots on the same axes (Darby, Mitchell, Van Leuvan, Kingbury, & McKay, 2012), can be conceptualised as an emergent feature in a display of close proximity. The Seagull Sign highlights a likely physiological abnormality when a patient’s heart rate is plotted above their systolic blood pressure at the same time-point (Darby et al., 2012) (see Figure 15). Arguably, when these two vital signs are graphed as overlapping plots on the same axes, the occurrence of a heart rate observation plotted above a systolic blood pressure observation (at the same time-point) is visually salient. The Seagull Sign may allow chart-users to engage in a more

efficient parallel visual search process, compared to a slower serial search where each individual element (i.e., observation) is inspected for a target (i.e., derangement) (Drews & Kramer, 2012).

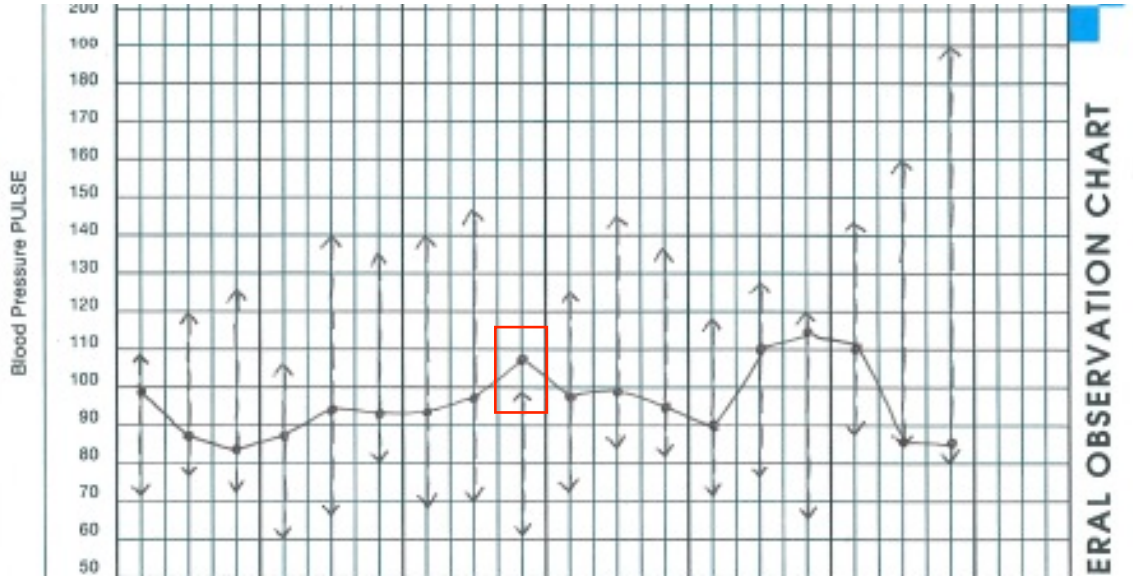


Figure 15. An extract of an existing chart with overlapping blood pressure and heart rate plots illustrating an example of the Seagull Sign (highlighted by the boxed area).

The emergence of the Seagull Sign may be attributable to what is known as gestalt grouping. Gestalt principles suggest that when similar objects are perceived as a group, a dissimilar object (the 'anomaly') becomes a focal point (Drews & Kramer, 2012; Zheng & Xue, 2009). On an observation chart with overlapping vital sign axes, the relationship between systolic blood pressure and heart rate is usually relatively stable: that is, blood pressure observations (typically marked by a 'v') are consistently plotted above heart rate observations (marked by a dot). From a gestalt perspective, consecutive occurrences of this consistent relationship (i.e., a 'v' above a dot) may be perceived as a group. If this were the case, an anomalous occurrence of a heart rate observation plotted above a systolic blood pressure observation would become particularly salient to the user.

Independent processing

The principle of proximity compatibility *also* proposed that display elements should be separated if independent processing is preferable (e.g., tasks that require two or more variables to be processed independently, or a variable that requires focused attention) (Wickens & Carswell, 1995). In the ADDS chart, blood pressure and heart rate are plotted on separate graphs. The rationale behind this design decision was that health professionals need to independently process both vital signs to determine which observation(s) are abnormal (see Figure 16). (Even in the presence of a Seagull Sign, users need to discern whether one or both vital signs are deranged; for example, a patient with normal blood pressure, but an abnormally high heart rate.)

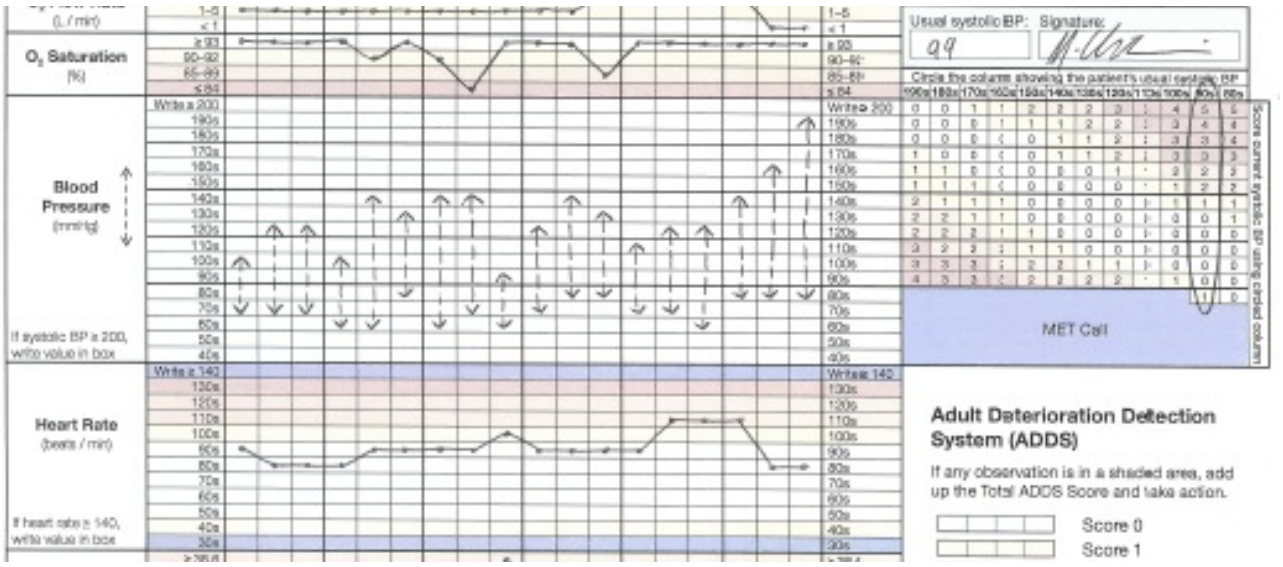


Figure 16. The ADDS chart plots blood pressure and heart rate on separate graphs (this case is equivalent to the case in Figure 15).

Considering the potential drawbacks of close processing proximity, Preece et al. (2013) also questioned clinicians' preferences for overlapping plots. Although this arrangement facilitates use of the Seagull Sign, potentially mitigating the need for slower serial search processes, emergent features can come at a cost (particularly if they are highly salient). The emergent feature can distract users from paying attention to other components of the display. This is especially problematic if other elements require analysis (Wickens & Carswell, 1995).

Cases of extreme spatial proximity can also result in visual clutter (Lee et al., 2013). When parts of a display overlap or are within a degree or so of visual angle from each other, the presence of one part can negatively impact the independent processing of the other. Additional processing demands are imposed when both are relevant to the task because it can be be difficult for users to perceptually parse each component from one another. Excess clutter can consequently slow the time it takes for users to search for, find and read items within a display (Wickens et al., 2004). This issue is well documented in the research literature on head-up displays, where air navigational information is projected onto a transparent screen between the pilot and windshield (FAA, 2008; Wickens & Hollands, 2000). Although this superimposition can help pilots concurrently compare head-up display information with the outside world (e.g., aligning the guidance with the true runway during a landing), the overlapping clutter can reduce the readability of the information and the ability to see unanticipated outside elements (e.g., a vehicle parked on the runway). Analogously, overlapping plots on a patient observation chart may lead to cluttered graphing, where deranged observations for one vital sign may be difficult to separate perceptually from observations for the other vital sign. Decreased discriminability of components, through contiguous or overlapping displays, is especially evident when extreme spatial proximity is used to emphasise

emergent features (Wickens & Carswell, 1995). Thus, although the Seagull Sign may attract users' attention to a particular time-point, the close proximity of the blood pressure and heart rate observations may make it more difficult and time-consuming for users to accurately identify which observation(s) are abnormal.

Another important cost/benefit consideration relates to the use of colour. Although colour can be valuable in a display because it can direct users' attention to particular elements, if its use is unrestrained, it can create clutter and increase visual search times (Karwowski, 2006). This risk is especially pertinent to the aforementioned overlapping blood pressure and heart rate plots. For example, there is one Australian observation chart that includes overlapping blood pressure and heart plots (see Figure 4) as well as a coloured track-and-trigger system. On this chart, the coloured ranges on the heart rate/blood pressure graph refer only to heart rate and the user must refer elsewhere for blood pressure ranges. This might result in increased cognitive load for the user.

Hypotheses and behavioural experiment (Christofidis, Hill, Horswill, & Watson, 2014)

In line with the ADDS chart design, the prediction was made that charts-users will perform significantly worse using charts with overlapping (vs. separate) plots, especially in the presence of an integrated colour-based track-and-trigger system. It was also hypothesised that the Seagull Sign will confer no advantage: that is, chart-users who are trained to use the Seagull Sign will not perform faster nor make fewer errors when using charts with overlapping (vs. separate) plots. The aim of the study reported in Chapter 3 is to test these hypotheses and hence resolve the contention surrounding whether the ADDS chart should present blood pressure and heart rate as separate plots. The experiment specifically examines whether experienced and novice chart-users can better recognise abnormal observations on separate or overlapping graphs and if the emerging feature of the Seagull Sign assists users to detect patient deterioration when plots do overlap.

Should the novel design use drawn dot observations, an integrated colour track-and-trigger system, and grouped scoring-rows to support users?

The third point of contention pertains to three related ADDS chart design features: (1) the use of drawn-dot observations, (2) the use of an integrated colour track-and-trigger system, and (3) the use of grouped vital sign scoring-rows. Arguably, it is possible that these three features could actually be hindering the usability of the ADDS design (i.e., where the chart demonstrates efficacy in spite of the above design decisions, and not because of them).

Data-recording format

The ADDS chart designers elected to use drawn-dot observations in light of several apriori human factors design principles within the literature (Gerhardt-Powals, 1996; Nielsen & Mack, 1994; Zhu et al., 2005). As described earlier, Preece et al. (2012b) reasoned that drawn-dot (vs. written number) vital sign observations would minimise cognitive load (Gerhardt-Powals, 1996; Nielsen, 1993) by precluding the possibility of users automatically reading the numerical values and/or comparing them to their memory of clinical criteria. Preventing this automatic reading was also reasoned to minimise unwanted data-driven searches (Gerhardt-Powals, 1996) for abnormal vital sign observations. Drawn-dots were also adopted because the alternative, written-number observations, may mislead chart-users. Vital sign measurements can vary due to: (a) natural steady-state variability; (b) transient perturbations (e.g., pain, anxiety); and (c) health professionals' technique (Reisner, Chen, & Reifman, 2012). This can lead to 'micro trends' in the physiological data that are simply aberrations of measurement error. However, human factors arguments can also be made for written-number observations, which feature on many other paper-based charts (see Figure 4 for an example of an existing chart). Written-number observations add redundancy due to the direct repetition of content in a different format. This may increase the chance that information will be noticed (Wickens et al., 2004). Where numbers are recorded in 'quasi-graphs', a deranged observation may be more noticeable in written-number form (vs. drawn-dot) because of its position within an abnormal reference range row *and* its abnormally high or low value. Also, in cases where an observation is recorded within the wrong reference range row, the written-number provides chart-users with an opportunity to correctly interpret the observation (as compared to a drawn-dot recorded in an incorrect range row). This practice is arguably in line with Gerhardt-Powals's (1996) recommendation to practice judicious redundancy. Including more information than may be needed at a given time (Gerhardt-Powals, 1996) can be a simple and effective method to increase the likelihood of a user detecting errors and correcting them (Salvendy, 1997).

Scoring-system integration

Again in line with apriori principles (Gerhardt-Powals, 1996; Nielsen & Mack, 1994; Zhu et al., 2005), the ADDS chart uses an integrated colour-based track-and-trigger system. Similar to drawn-dot observations, Preece et al. (2012b) reasoned that a colour-based system would minimise users' cognitive load and reduce data-driven aspects of the deterioration detection task (Gerhardt-Powals, 1996; Nielsen, 1993). The coloured range rows mean that users do not have to remember or look up (e.g., in a reference table) normal vital sign ranges. Thus, when an observation crosses a particular threshold of abnormality the user simply has to notice that the observation is recorded

against a coloured background, hence making the task of detection more automated and less data-driven. However, many other observation charts use a non-integrated tabular system (see Figure 3 for an example of such a chart) for which alternative human factors arguments can be made as to why this design decision might be superior. For example, Nielsen (1993) and Gerhardt-Powals (1996) recommend a 'less is more' approach to displays, where only information that is needed by the user is included. Both authors suggest that interfaces should be simplified as much as possible, arguing that every extra item within a display is an additional piece of information to learn, search through and possibly misunderstand. Extraneous information can slow down expert users, but more critically, can confuse novices (Nielsen & Mack, 1994). A non-integrated tabular system may simplify the chart display such that users (especially those who are experienced and have the vital sign reference ranges implicit in memory) can search for deranged observations without interference from the coloured range rows.

Scoring-row placement

Finally, the ADDS chart groups scoring-rows together at the bottom of the page, as Gerhardt-Powals (1996) and Nielsen (1993) both emphasise the effect that meaningfully grouped data can have on the speed with which information is accessed. The authors propose that closely positioned information will be beneficial when information needs to be used together. The scoring-rows are grouped together so that chart-users can first allocate an ADDS score to each vital sign observation, and then sum the recorded scores together to form a total score. In contrast to other observation charts that present the rows separately (i.e., directly underneath the corresponding vital sign data; e.g., see Figure 4), it was reasoned that the ADDS chart layout would save users from potential visual interference. That is, grouped scoring-rows for individual vital signs would allow users to search for abnormal observations without interference from individual vital sign scores (and also to assess individual vital sign scores without interference from vital sign observations).

Nevertheless, the use of separate scoring-rows can also be supported using a human factors rationale. Early-warning scores (if recorded accurately) can act as redundant cues in the detection of deterioration. For instance, if a nurse reviews a patient's earlier observations and fails to notice an abnormal vital sign recording, they may still detect the corresponding early-warning score. In this context, charts with separate rows provide immediate redundancy, as users presumably assess a set of observations and then consult the corresponding scoring-row immediately below (i.e., positioning the score close the corresponding observation will reduce users' search time, which again adheres to the suggestions of Gerhardt-Powals (1996) and Nielsen (1993) to group data in a consistently meaningful way). Charts with grouped rows provide comparatively delayed redundancy, as users are more likely to assess each set of vital sign observations consecutively and

then consult the early-warning score rows together as a separate task. This example highlights the aforementioned difficulties in applying human factors design principles. In this case, the purposeful abstraction of the principle to 'display information that will be used together close together', means that a single principle can be implemented in multiple ways and guidelines to direct a decision between alternatives do not exist (Wickens et al., 2004).

Hypotheses and behavioural experiment (Christofidis, Hill, Horswill, & Watson, in press)

It was hypothesised that the design features of the ADDS chart will benefit users. That is, participants' performances are proposed to be consistent with the apriori human factor principles that Preece et al. (2012b) adapted from the web and software domains (Gerhardt-Powals, 1996; Nielsen, 1993; Zhu et al., 2005). As such, it is predicted that chart designs with drawn-dot observations, an integrated colour-based scoring-system and grouped scoring-rows will yield the fastest and most accurate responses. Chapter 4 describes an experiment designed to test these hypotheses with a view to resolving the debate surrounding the effect of these specific chart features on users' recognition of patient deterioration.

Does the layout of the novel design best facilitate the calculation of summary scores?

The final point of contention relates to whether the design layout of the ADDS chart best facilitates users' calculations of patient deterioration summary scores. The ADDS chart uses a combination scoring system in which chart-users determine early-warning scores that summarise the physiological state of a patient and trigger appropriate clinical action (Mohammed, Hayton, Clements, Smith, & Prytherch, 2009; Prytherch et al., 2006). The multi-step process involved in determining an early-warning score suggests that chart designs with combination (and aggregate) scoring systems may be more susceptible to error, particularly because the accuracy of a given step depends on the accuracy of the preceding step. For instance, a correct early-warning score depends on accurate individual vital sign scores (where the ADDS chart, for example, includes eight vital signs). Individual scores depend on appropriately recorded observations, which are contingent on carefully collected vital sign measurements. A correct early-warning score *also* depends on the accurate summation of individual vital sign scores. This step is of particular interest from a human factors perspective. Although tasks that involve mathematical calculations (and their verification) are commonplace for health professionals, they are inherently prone to error (Sela & Auerbach-Shpak, 2014). Empirical evidence has demonstrated poor mathematical ability amongst qualified and student nurses, where simple arithmetical mistakes constitute one of the major sources of

mathematical error. For example, a sample of Australian second-year undergraduate nursing students scored an average of 56.1% in a test of basic mathematical and drug calculations, where over a third of total errors were arithmetical (Eastwood, Boyle, Williams, & Fairhall, 2011).

Grouped scoring-rows

As previously mentioned, Preece et al. (2012b) used the principles of Gerhardt-Powals (1996) and Nielsen (1993) to hypothesise that the use of grouped scoring-rows would remove the potential interference of individual vital sign scores when users search for abnormal observations (and similarly, remove the possible interference of recorded observations when assessing individual scores). However, given that human factors design principles are not hard and fast rules (Proctor & Van Zandt, 2008; Wickens et al., 2004), it cannot be assumed that one design decision will apply to all situations. Thus, a novel task demands a reconsideration of the principles. In this instance, we need to consider the ways in which data can be meaningfully grouped to assist chart-users to determine both individual vital sign scores and early-warning scores. Once again, the abstract nature of the principle facilitates more than one reasonable application. First, grouped scoring-rows may help users sum individual early-warning scores into a total score, because their attention can remain focused on one part of the chart. This was the rationale of the ADDS chart designers (Preece et al., 2012b). However, some health professionals have argued that this layout will impair the recording of individual vital sign scores. These clinicians have highlighted the potential for error when chart-users determine an individual score on one part of the chart (i.e., where the observation is recorded) and then switch their attention to another part of the chart to record the score (i.e., beneath all of the vital sign data). It is possible that the mental effort involved in reorienting their attention to a new visual space after a large visual switch will lead chart-users to make mistakes when recording individual vital sign scores.

Separate scoring-rows

Some of these health professionals prefer observation charts to incorporate separate scoring-rows to support users' recording of individual vital sign scores (e.g., see Figure 4). In line with the principle that meaningfully grouped data can improve the search for information (Gerhardt-Powals, 1996; Nielsen, 1993), the close proximity of each row to the corresponding vital sign data could arguably facilitate faster and more accurate determinations of individual scores. However, the ADDS chart designers reasoned that when summing separated scores (which on the ADDS chart, covers almost the whole height of an A3 page), the in-between observations may interfere with a user's visual search down the time-point column such that they may read the wrong score(s) (e.g., in an adjacent column) or skip a score (or scores) entirely.

No scoring-rows

Another variation of the ADDS chart, developed more recently by an Australian state health department, excludes individual vital sign scoring-rows altogether (see Figure 17). On this chart, users need to concurrently determine each individual vital sign score while holding a running total in mind. This design would almost certainly yield faster response times compared to those charts with scoring-rows, as it precludes users from recording 144 extra scores for every complete chart. However, this potential design solution is not without risk. Without rows to record individual scores, chart-users rely on an internal representation of the calculation process which may be compromised by the limitations of working memory (Wickens & Hollands, 2000). Working memory plays an important role in the computation of arithmetical answers, as it temporarily holds the initially presented operand(s) and the intermediate value(s) computed during the solution. However, working memory is limited in that only a small amount of information can be 'worked on' by other cognitive transformations. If a manipulation prolongs the period in which information is stored, a heavier load is placed on working memory and error can result (Campbell, 1992; Wickens et al., 2004). Thus, in summing individual early-warning scores, manipulations that prolong the storage of the initially presented operand (i.e., a determined individual vital sign score) or the intermediate computed values (i.e., the progressively summed scores) may increase the risk of arithmetical errors.

Chart designs without scoring-rows (vs. those with rows) may be more susceptible to these working memory limitations. The storage period for holding intermediate values will be comparatively prolonged because users need to simultaneously determine successive individual vital sign scores. (Note that the need to cross-reference to the systolic blood pressure table on the ADDS chart may further prolong this storage period.) This is an example of retroactive interference, where the retrieval of material-to-be-remembered is disrupted by subsequent activity. The risk of interference during the retention interval tends to increase if the material is impeded by other material of the same type (Wickens & Hollands, 2000). As such, chart-users' storage of intermediate scores during the summation process (i.e., digits that range from 0 to 8) may be made worse by the determination of subsequent individual vital sign scores (i.e., digits that range from 0 to 5).

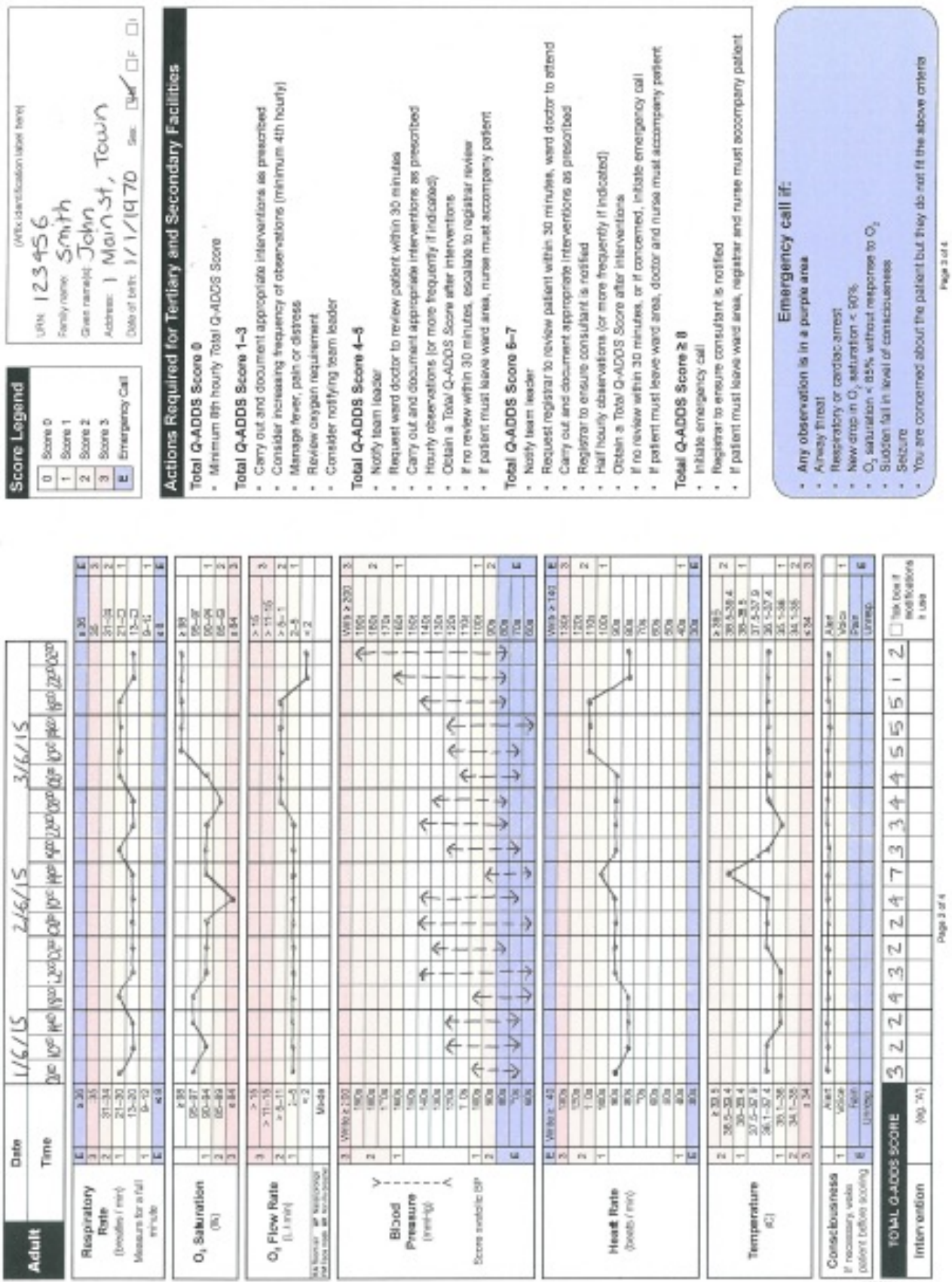


Figure 17. An existing chart (A3 size) with no individual vital sign scoring-rows. (Note that the chart has been rotated 90 degrees to fit the page.)

Excluding scoring-rows may be especially problematic in clinical environments where disruptions are commonplace. For example, if a health professional is interrupted part-way through their calculation of an early-warning score, they will either attempt to remember the intermediate score they were up to (which may be susceptible to error) or start the process over again (which will prolong task duration).

Indeed, interface designers in Israel recently suggested that paper-based charts provide an external representation of users' calculation processes (e.g., where data are presented in a table or formula). In their redesign of a radiotherapy chart, Sela and Auerbach-Shpak (2014) aimed to reduce errors associated with the calculation of radiation doses (where under-dosing can lead to ineffective cancer treatment, and overdosing can injure the exposed body area). After an initial assessment, the authors established that their institution's existing chart did not assist staff to perform calculations or integrate information. For example, to determine the total dose of a patient's radiation, users had to refer to constituent doses that were spread across different (and disorganised) areas of the chart. According to Sela and Auerbach-Shpak (2014), this segmentation unnecessarily complicated what should have been a simple calculation process. To reduce the risk of dose errors, radiation data was presented within a single table where the calculations to-be-performed were arranged in sequence. From a human factors perspective, the authors argued that this external representation would organise the calculation process, reduce users' memory loads, and facilitate easy checking. The potential benefits associated with an externally represented calculation process may be an argument for including individual vital sign scoring-rows (regardless of whether they are grouped or separate) on observation charts with combination and aggregate weighted scoring systems.

Hypotheses and behavioural experiment (Christofidis, Hill, Horswill, & Watson, 2015)

As described earlier, separate scoring-rows may help chart-users determine individual vital sign scores, while grouped rows may be of greater benefit when users add these scores. However, in anticipation of significantly more adding errors than scoring errors, it is hypothesised that users will determine total early-warning scores more accurately when scoring-rows are grouped. It is also hypothesised that, in the absence of scoring rows for individual vital signs, determining individual scores will prolong the storage period for intermediate values, increasing the rate of errors. However, it is anticipated that the absence of these scoring rows may yield a speed-accuracy trade-off in which users determine early-warning scores faster than when using charts with rows. Chapter 5 presents an experiment designed to test these hypotheses and hence resolve this contentious design problem. Although Chapter 4 already addresses the placement of scoring-rows in the context of identifying abnormal observations, the usability of this design feature may be more critical when

users engage with the scoring system itself (e.g., calculating and summing individual vital sign scores), rather than when simply detecting deterioration on charts where scores have already been computed.

Approach of the thesis

In the absence of expert consensus, this thesis proposes that we must turn to scientific experimentation to resolve controversies in patient observation chart design and evaluate best practice. Chapters 2 to 5 describe four behavioural experiments that address each aforementioned point of contention in turn. Chapter 6 discusses the implications of the empirical findings and the limitations of the project, as well as presenting suggestions for future research. Two dependent variables, error rate and response time, were used as the performance measures across each experiment for two key reasons. First, accuracy and efficiency are critical in the real-world task of recording and monitoring vital signs. From a usability perspective, observation charts should yield low errors rates from health professionals as well as efficient engagement for optimal productivity. Second, the reciprocity that can occur between errors and time mean that speed-accuracy trade-offs have the potential to explain certain findings. Sometimes, the speed-accuracy trade-off between systems differ because one design may induce more careful but slower behaviour, and the other faster but less precise behaviour (Wickens et al., 2004).

Chapter 2

Christofidis, M.J., Hill, A., Horswill, M.S., & Watson, M.O. (2013). A human factors approach to observation chart design can trump health professionals' prior chart experience. *Resuscitation, 84*(5), 657-665.

Table 1. Manuscript revision history for "A human factors approach to observation chart design can trump health professionals' prior chart experience"

Date	Detail
29 June 2012	Submitted to *Resuscitation*
31 August 2012	Article revised
19 September 2012	Article accepted for publication
7 May 2013	Published in print

Hypotheses

Chart-users will make fewer errors and respond more quickly when using the ADDS chart, compared to chart designs with which they are highly experienced. Or alternatively, chart-users will perform better when presented with the chart that they regularly use in their occupational role.

1. Introduction

Paper-based observation charts are typically designed at individual institutions, or at the level of the individual area health service, by clinical staff perceived as having some knowledge or experience of chart design.[1,2] This results in considerable variation in the design of charts between, and even within, hospitals.[2] Consequently, the type of chart designs that health professionals are experienced in using to record and monitor physiological variables can vary dramatically. Across Australia, for example, chart design can vary according to (a) the selection, order and display format (e.g., numerical vs. graphical) of vital signs that can be monitored; (b) whether or not track-and-trigger systems or emergency call criteria are used; (c) the orientation of data series and pages (i.e., landscape vs. portrait); and (d) the use of abbreviated terminology.[3]

In the context of this lack of standardization, health professionals have tended to rely on their own subjective judgments, and those of their peers, to assess the efficacy of individual chart designs.[1,2] Recently, however, there have been several efforts to guide the design of observation charts using evidence-based approaches.[1,2,4] For instance, Horswill et al.,[5] in consultation with clinicians, designed two versions of a chart (the Adult Deterioration Detection System, or ADDS, chart), which were developed from a human factors perspective to facilitate the detection of patient deterioration in a user-friendly manner. In a subsequent experimental study by Preece et al.,[1] experienced health professionals and novices were shown realistic patient data presented on the two ADDS charts and four pre-existing chart designs, and judged whether or not any of the vital signs were abnormal. Both groups made fewer errors and responded more quickly when using the user-friendly ADDS charts compared with the other designs, suggesting that variability in chart design quality can have a considerable effect on the performance of both experienced and novice observation chart users.

One limitation of the study, however, was that previous experience with particular chart designs was not controlled for.[1] On average, the health professionals, who were recruited from a tertiary referral teaching hospital,[6] reported having previously used two of the charts presented during the experiment (or very similar chart designs; see Table 1).[1] Given recent Australian government initiatives to develop an evidence-based general adult observation chart[7] with the potential for state- or nation-wide standardisation, and the possibility that other governments will follow suit, it is also crucial to assess whether health professionals' prior chart experience affects their ability to detect patient deterioration on the new, user-friendly designs. Although a widely used standardised chart could plausibly lead to efficiency gains for staff working in multiple facilities (either concurrently or over time), there are also several reasons why its initial implementation could potentially be problematic.

One possibility is that, if chart users have extensive experience with one particular chart, then this familiarity might result in superior performance using that chart as opposed to a better-designed replacement. This may be the case even if the replacement chart has been designed from a human factors perspective and can be demonstrated to be a superior choice for novice users or clinicians with experience using a variety of other charts (as with the ADDS[1]). Hence, it is not a foregone conclusion that the best-designed chart will immediately yield the best performance in all user-groups, irrespective of their prior experience. In addition, health professionals who are highly experienced in using a particular chart may be more resistant to the implementation of an alternative design, as familiarity and perceived satisfaction can strongly influence users' preference for a specific system.[8] Post implementation, health professionals may be less likely to comply with chart-related protocols if they falsely believe that a poorly designed chart that they are experienced in using is not problematic (and that a new best-practice chart is).[9]

Given these usability risks, the present study empirically evaluated the effect of observation chart design on the ability of health professionals, highly familiar with and experienced in using a specific chart favoured by their institution, to recognise abnormal vital sign observations on a range of chart designs. The six designs selected for comparison, and the patient cases recorded on them, were those used in Preece et al.'s study[1] (i.e., both versions of the ADDS Chart, and four pre-existing Australian designs). Two groups of participants were selected for their extensive experience with one or other of the pre-existing charts (or a very similar design). They were asked to judge whether observations recorded on the charts were physiologically abnormal or normal.

Table 1
Number and percentage of health professionals in Preece et al's study, and each experience group in the present study, who reported having experience with charts very similar to those used in the experiment (participants could select more than one chart).

Chart used in the experiments	**Health professional group**		
	Preece et al.[1] **(*n* = 45)**	**The present study**	
		Multiple parameter track-and-trigger chart experienced (*n* = 64)	**No track-and-trigger graphical chart experienced (*n* = 37)**
No track-and-trigger numerical	23 (51.11%)	3 (4.69%)	-
No track-and-trigger graphical	19 (42.22%)	1 (1.56%)	37 (100.00%)
Single parameter track-and-trigger	23 (51.11%)	1 (1.56%)	-
Multiple parameter track-and-trigger	8 (17.78%)	64 (100.00%)	-
ADDS chart with systolic blood pressure table	8 (17.78%)	-	-
ADDS chart without systolic blood pressure table	8 (17.78%)	-	-

Two competing hypotheses were proposed. *Hypothesis 1*: Prior experience will outweigh design, such that chart users will be most accurate and fastest when presented with the chart that they regularly use in their occupational role (or a very similar chart). *Hypothesis 2:* Alternatively, user-friendly design will outweigh prior experience, such that each experience group will make fewer errors and respond more quickly when using the two charts developed from a human factors perspective, compared with the pre-existing charts (including the design with which they are highly experienced).

2. Methods

2.1. Participants

Participants were two groups of doctors and nurses recruited and tested between September 2010 and April 2011. Participants experienced in using the multiple parameter track-and-trigger chart that was included in the study materials (n = 64) were recruited from The Canberra Hospital (Garran, Australian Capital Territory, Australia). Participants experienced with a no track-and-trigger graphical chart similar to the one included in the study materials (n = 37), were recruited from Mt Isa Base Hospital (Mt Isa, Queensland, Australia) and Logan Hospital (Meadowbrook, Queensland, Australia). An additional four health professionals participated in the study but were excluded from the analyses: one participant from The Canberra Hospital who reported not having used the multiple parameter track-and-trigger chart in their occupational role, and three from Mt Isa Base Hospital and Logan Hospital who reported not having used a no track-and-trigger graphical chart. All participants gave informed consent and were compensated AUD100 for their time. Each hospital's ethics committee approved the study.

2.2. Patient data

The forty-eight cases of genuine de-identified patient data used in the study by Preece et al.[1] were re-used in this study. Spanning 13 consecutive time-points, each case included data for the nine vital signs that were common to all six observation chart designs: respiratory rate, oxygen delivery, oxygen saturation, systolic and diastolic blood pressure, heart rate, temperature, consciousness and pain.

Twenty-four of the cases included an abnormal observation (i.e., a vital sign observation outside of the defined set of normal ranges provided by three of the observation charts used in this study; see Table 2 for the vital sign normal ranges), whilst the remaining twenty-four cases

contained only normal observations. The abnormal cases included derangements in oxygen saturation (6 hypoxic cases), systolic blood pressure (3 hypotensive and 3 hypertensive cases), heart rate (3 bradycardic and 3 tachycardic cases) and temperature (3 hypothermic and 3 febrile cases). Each set of patient data had been carefully hand-plotted onto each of the six chart designs tested in the study (48 cases × 6 charts = 288 charts). For additional details on these materials, see Preece et al.[1]

Table 2
Vital sign normal ranges used in the experiment (table reproduced from Preece et al.[1]).

Vital sign	Normal range
Respiratory rate	Between 9 – 20 breaths per minute
Oxygen delivery	Patient is receiving oxygen at ≤ 1 litre per minute
Oxygen saturation	Between 93 – 100%
Systolic blood pressure	Between 100 – 160 mmHg
Heart rate	Between 50 – 100 beats per minute
Temperature	Between 36.1 – 37.9 Celsius
Consciousness	Patient is classified as being alert
Pain	Patient is in no pain

2.3. Observation charts

Two versions of the Adult Deterioration Detection System (ADDS) chart were included in the study. These were developed from a human factors perspective in response to an evaluation of usability problems affecting 25 existing Australian and New Zealand observation charts.[3] Also included were charts that had been classified in that review as being either: (1) reasonably well-designed (i.e., the single and multiple parameter track-and-trigger charts); (2) of average design quality (i.e., the no track-and-trigger graphical chart); or (3) poorly designed (i.e., the no track-and-trigger numerical chart).

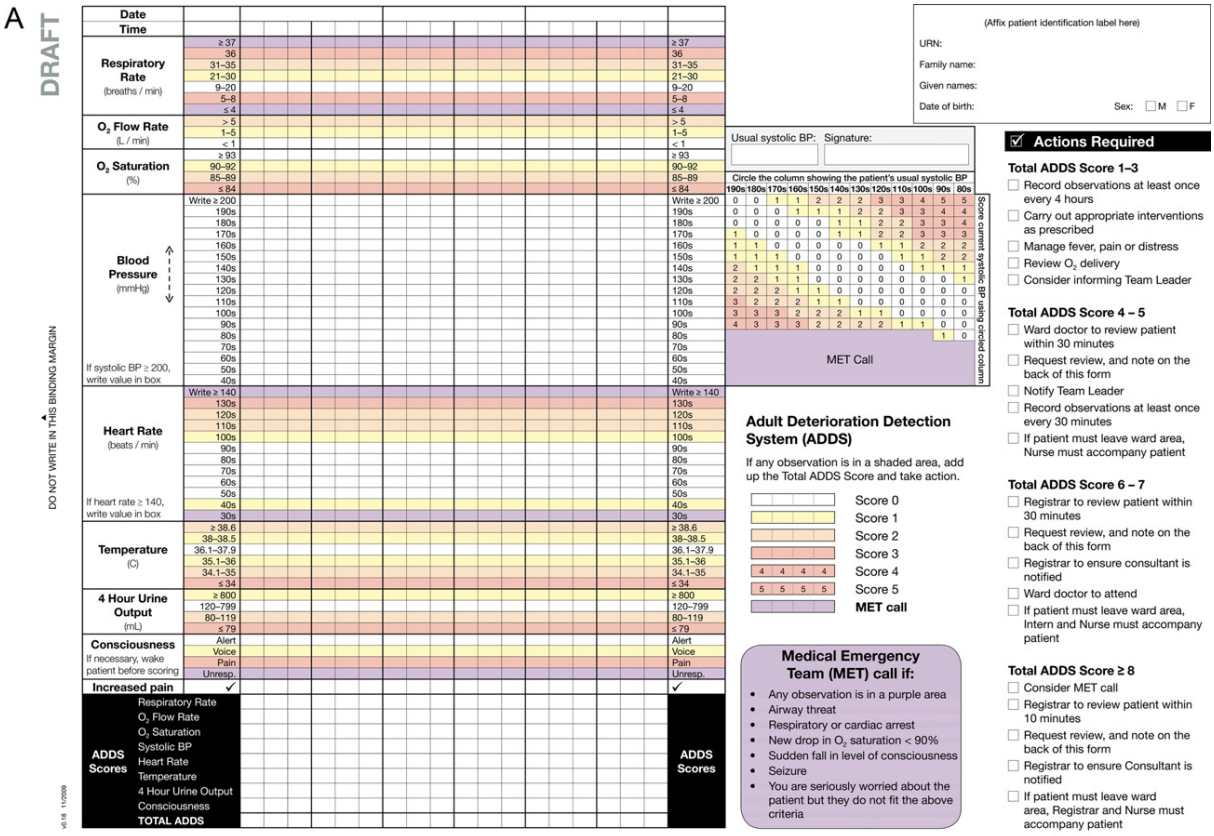
Table 3 outlines the key design characteristics of the six charts used in the study. For comparison, it also includes details of the chart habitually used by the no track-and-trigger graphical chart experience group. Figure 1 presents images of the two charts relevant to the experience groups (and *Hypothesis 1*), and the two charts developed from a human factors perspective (and relevant to *Hypothesis* 2).

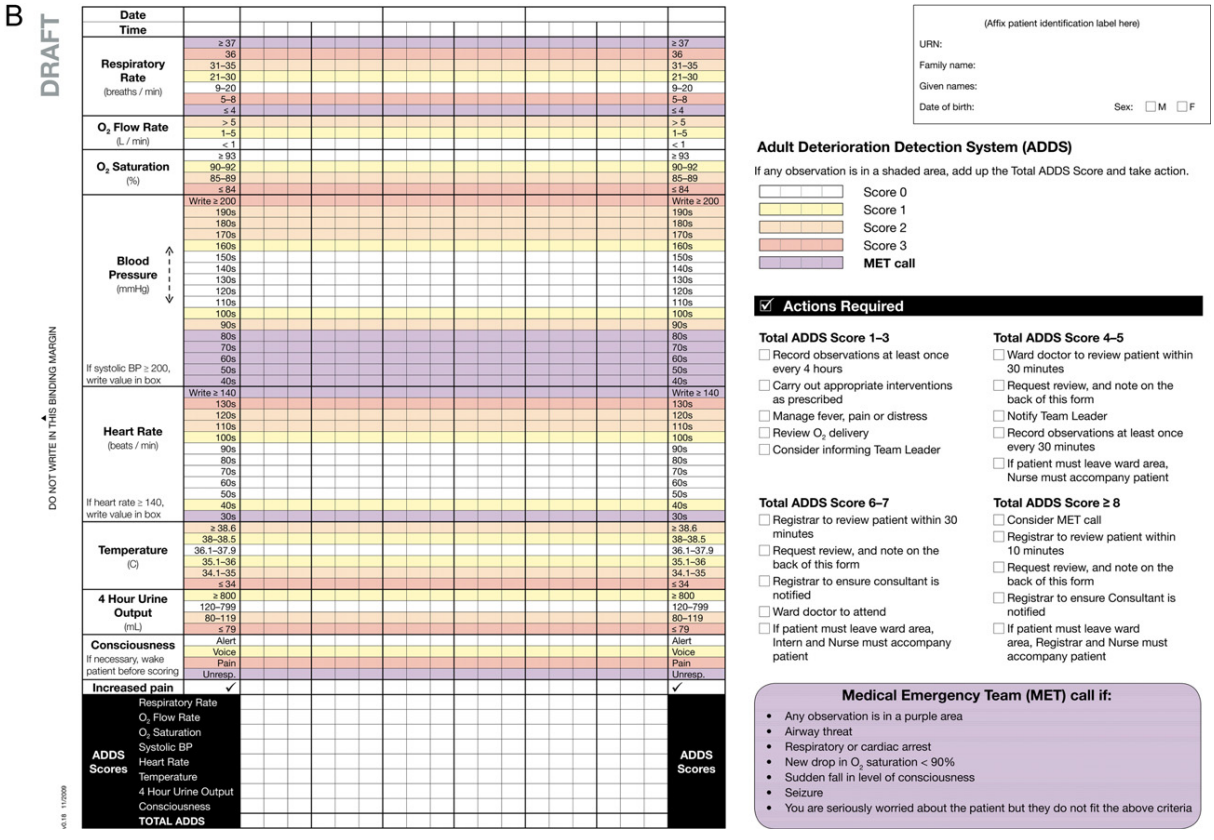
Table 3

Key design characteristics of the six charts used in the study, and the chart habitually used by the no track-and-trigger graphical chart experience group. (Note that the table has been rotated 90 degrees to fit the page.)

Key design characteristics	Adult Deterioration Detection System with systolic blood pressure table	Adult Deterioration Detection System without systolic blood pressure table	Multiple parameter track-and- trigger chart	Single parameter track-and-trigger chart	No track-and-trigger graphical chart	No track-and-trigger numerical chart	Chart used by the no track-and-trigger graphical chart experience group
Paper size and sidedness	A3, double-sided.	A3, double-sided.	A4, double-sided.	A3, double-sided.	A3, double-sided.	A4, single-sided.	A4, double-sided.
Page orientation	Landscape.	Landscape.	Portrait.	Landscape.	Landscape.	Portrait.	Portrait.
Display format of vital signs	All observations are recorded graphically, except pain.	All observations are recorded graphically, except pain.	All observations are recorded graphically, except oxygen delivery and pain.	All observations are recorded graphically except oxygen delivery and pain.	Temperature, blood pressure and heart rate are recorded graphically (all other vital signs are tabulated).	All observations are tabulated.	Temperature, blood pressure, heart rate and respirations are recorded graphically.
Use of track-and-trigger systems	Single and multiple parameter track-and-trigger systems integrated into the observations area.	Single and multiple parameter track-and-trigger systems integrated into the observations area.	Multiple parameter track-and-trigger system integrated into the observations area. Written emergency call criteria.	Single parameter track-and-trigger system integrated into the observations area.	No integrated track-and-trigger system. Written emergency call criteria.	No integrated track-and-trigger system or written emergency call criteria.	No integrated track-and-trigger system or written emergency call criteria.
Use of colour to signal abnormalities	Different bandings of colour (based on the severity of abnormality) are used to score vital signs. Purple bands are used indicate that an emergency call should be placed immediately.	Different bandings of colour (based on the severity of abnormality) are used to score vital signs. Purple bands are used indicate that an emergency call should be placed immediately.	Different bandings of colour (based on the severity of abnormality) are used to score vital signs.	Two bandings of colour (yellow and red) denote different levels of abnormality, indicating what action should be taken.	Colour is not used to signal abnormalities.	Colour is not used to signal abnormalities.	Colour is not used to signal abnormalities.
Presentation of blood pressure and heart rate	Presented as separate plots on separate graphs.	Presented as separate plots on separate graphs.	Presented as overlapping plots on the same graph using the same axes.	Presented as separate plots on separate graphs.	Presented as overlapping plots on the same graph using the same axes.	Presented as separate plots on separate graphs.	Presented as overlapping plots on the same graph using the same axes.
Scoring of blood pressure relative to the patient's usual systolic blood pressure	Systolic blood pressure scoring table located adjacent to the graph of blood pressure observations.	No individualised scoring of the patient's systolic blood pressure.	Systolic blood pressure scoring table located on the reverse of the page.	No individualised scoring of the patient's systolic blood pressure.	No individualised scoring of the patient's systolic blood pressure.	No individualised scoring of the patient's systolic blood pressure.	No individualised scoring of the patient's systolic blood pressure.

Fig. 1. Charts relevant to the hypotheses: the multiple parameter track-and-trigger chart (front and back; de-identified) (A); the no track-and-trigger graphical chart (inside pages only; de-identified) (B); the ADDS chart with systolic blood pressure table (inside pages only) (C); and the ADDS chart without systolic blood pressure (inside pages only) (D). See the web version of the article for colour images.





C

General Observation Chart

Modified Early Warning Scores	0	1	2	3	Observation chart number:
Date					
Time					
Resp Rate MET RR < 5 or > 36	≥ 36				
	31 – 35				
	21 – 30				
	9 – 20				
	≤ 8				
RR Score					
Oxygen mode (l/min)					
SpO_2	≥ 93				
	90 – 92				
	85 – 89				
	≤ 84				
SpO_2 Score					
Temp (°C)	≥ 39.6				
	38.6 – 39.5				
	38 – 38.5				
	37 – 37.9				
	36.1 – 36.9				
	35.1 – 36				
	34.1 – 35				
	≤ 34				
Temp Score					
Blood Pressure & Heart Rate ( ) MET criteria HR < 40 or > 140 SBP < 90 — **Usual SBP:**	200				
	190				
	180				
	170				
	160				
	150				
	140				
	130				
	120				
	110				
	100				
	90				
	80				
	70				
	60				
	50				
	40				
Score HR					
Score BP (see back)					
Sedation Score	0				
	1				
	2				
	3				
	4				
Sedation Score					
Urine for 4hrs	> 800				
	120 – 800				
	80 – 119				
	< 80				
Urine Score					
TOTAL MEWS					
Pain					
Bowels					
Initial					

General Observation Chart

60201(1007)

Date										
Time										
Weight										
DVT/PE Risk	H M L	H M L	H M L	H M L	H M L	H M L	H M L	H M L	H M L	H M L
Waterlow										

Urinalysis

Date										
Time										
Leucocytes										
Nitrate										
Urobilinogen										
Protein										
pH										
Blood										
Specific Gravity										
Ketones										
Bilirubin										
Glucose										

Height: **Weight on Admission:**

BMI **Pain Score** No pain – 0 Worst – 10

Sedation Score:
0 = awake & alert
1 = normally asleep, responds to stimuli
2 = mild, occasionally drowsy, easy to rouse
3 = moderate, frequently drowsy easy to rouse but unable to maintain wakeful state
4 = severe, somnolent, difficult to rouse

Other obs charts in use:
PCA
Epidural
Neurological
Blood transfusion
Underwater seal
Neurovascular
Peak flow
Alcohol withdrawal
Fluid balance
Bladder irrigation
Post angiogram
Post procedural
Stool chart
Blood glucose

Circle patient's usual BP

Additional Considerations

Usual SBP	190	180	170	160	150	140	130	120	110	100	90	80
Current Blood Pressure 200s	0	0	1	1	2	2	2	3	3	4	5	5
190s	0	0	0	1	1	1	2	2	3	3	4	4
180s	0	0	0	0	0	1	1	2	2	3	3	4
170s	1	0	0	0	0	1	1	2	2	3	3	3
160s	1	1	0	0	0	0	0	1	1	2	2	2
150s	1	1	1	0	0	0	0	0	1	1	2	2
140s	2	1	1	1	0	0	0	0	0	1	1	1
130s	2	2	1	1	0	0	0	0	0	0	0	1
120s	2	2	2	1	1	0	0	0	0	0	0	0
110s	3	2	2	2	1	1	0	0	0	0	0	0
100s	3	3	3	2	2	2	1	1	0	0	0	0
90s	4	3	3	3	2	2	2	2	1	1	0	0
80s	MET 4				MET 3			MET 2			1	0
70s											MET 1	

MET Criteria:

All respiratory & cardiac arrests

Threatened Airway, RR < 5 or > 36

Pulse < 40 or > 140

Systolic BP < 90

Sudden fall in level of consciousness, fall of GCS > 2, repeated or prolonged seizures

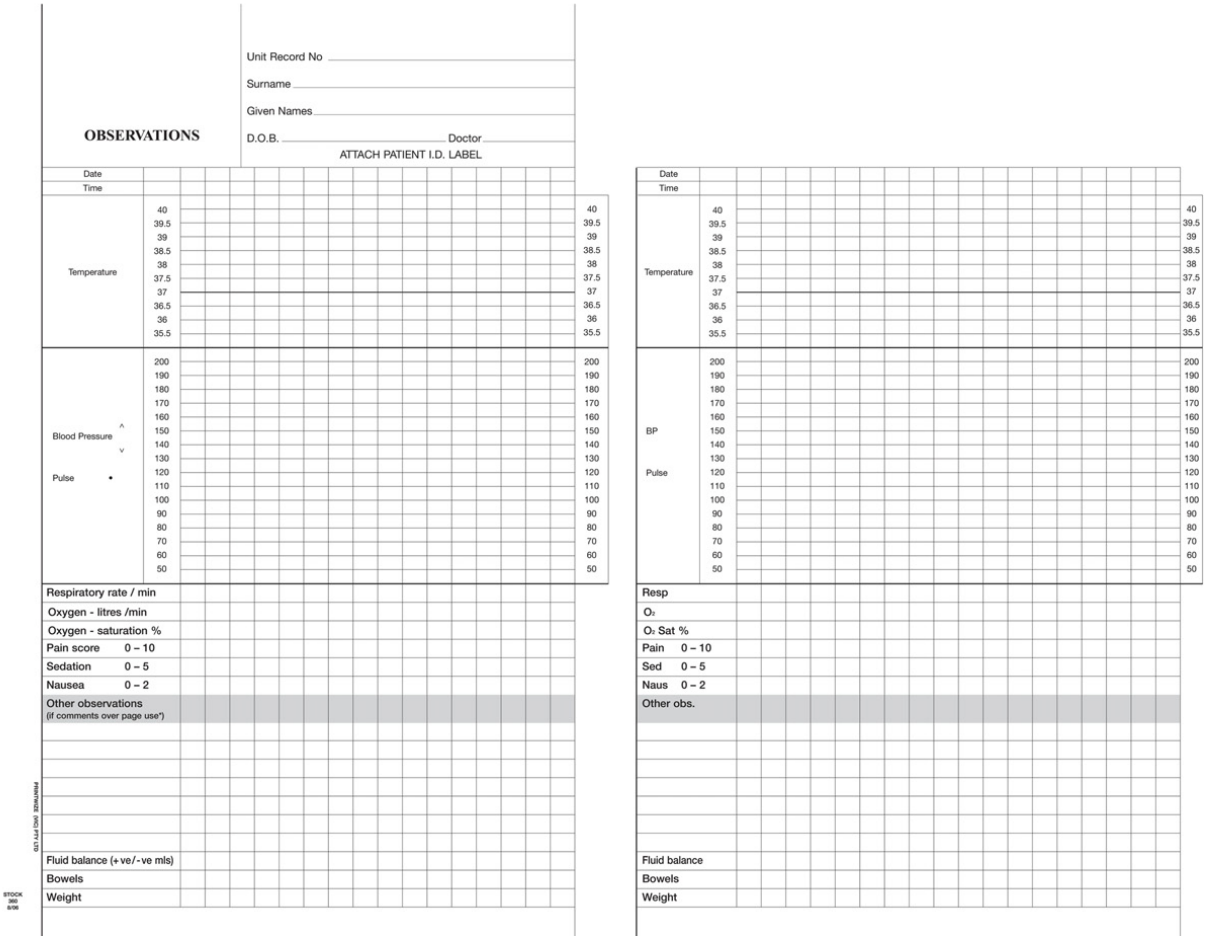
Any patient you are seriously worried about that does not fit the above criteria

Date/Time	Modification to MEWS	Signature

Action if MEWS 4

Date/Time	MEWS	Action	Signature

D



Unit Record No

Surname

Given Names

D.O.B. Doctor

ATTACH PATIENT I.D. LABEL

OBSERVATIONS

Date										
Time										
Temperature	40									40
	39.5									39.5
	39									39
	38.5									38.5
	38									38
	37.5									37.5
	37									37
	36.5									36.5
	36									36
	35.5									35.5
Blood Pressure ^ v / Pulse •	200									200
	190									190
	180									180
	170									170
	160									160
	150									150
	140									140
	130									130
	120									120
	110									110
	100									100
	90									90
	80									80
	70									70
	60									60
	50									50
Respiratory rate / min										
Oxygen - litres /min										
Oxygen - saturation %										
Pain score 0 – 10										
Sedation 0 – 5										
Nausea 0 – 2										
Other observations (if comments over page use*)										
Fluid balance (+ve/-ve mls)										
Bowels										
Weight										

Date										
Time										
Temperature	40									40
	39.5									39.5
	39									39
	38.5									38.5
	38									38
	37.5									37.5
	37									37
	36.5									36.5
	36									36
	35.5									35.5
BP / Pulse	200									200
	190									190
	180									180
	170									170
	160									160
	150									150
	140									140
	130									130
	120									120
	110									110
	100									100
	90									90
	80									80
	70									70
	60									60
	50									50
Resp										
O_2										
O_2 Sat %										
Pain 0 – 10										
Sed 0 – 5										
Naus 0 – 2										
Other obs.										
Fluid balance										
Bowels										
Weight										

PRINTWISE (NZ) PTY LTD

STOCK 360 8/08

2.4. Design and procedure

The study used a mixed design, with chart experience group (between-subjects: multiple parameter track-and-trigger chart experience vs. no track-and-trigger graphical chart experience) and chart type (within-subjects) as the independent variables.

Health professionals initially completed a questionnaire that assessed their demographic and clinical background. Participants then watched a training video that described: (a) the normal ranges for each of the nine vital signs; (b) track-and-trigger systems; and (c) how to use each observation chart (presented in a different random order for each participant). Next, participants' knowledge of the key information and normal ranges was tested with a 10-item multiple-choice examination (if an item was answered incorrectly, participants were required to study the normal ranges and retake the examination until they answered all items correctly). The experimental protocol was then described in a final video presentation.

Across 48 trials, each participant viewed each set of patient data once. The six charts were each used on eight trials, four times with abnormal data and four times with normal data. For each participant, cases of patient data were randomly assigned to chart designs with the constraint that, for each chart, derangements included oxygen saturation, systolic blood pressure, heart rate, and temperature. Trials were presented in a different random order for each participant to prevent order effects.

In each trial, the participant was presented with a chart and was asked to judge whether any of the observations were abnormal (and, if so, to specify which), or whether all of the observations were normal. Participants' responses and response times were recorded using a customized computer program.

Following the experiment, participants completed a questionnaire that assessed their prior chart experience.

2.5. Statistical Analyses

For each trial, a response was coded as 'correct' if the participant correctly singled-out an abnormal vital sign, or correctly identified a normal case as normal. Each participant's error rate (i.e., percentage of incorrect responses) and average response time were calculated for each chart as the outcome measures. Statistical analyses were performed using SPSS 20.0 (SPSS Inc, Chicago, Ill, USA). Statistical significance was set at $\alpha = 0.05$. Separate mixed-design (chart type x experience group) analyses of variance were conducted on error rates and response times, respectively. Because Mauchly's *W* was significant (indicating violation of the sphericity

assumption) for both analyses, the Greenhouse-Geisser correction was applied to the within-participants effects. For each significant omnibus test, η^2 was calculated as the measure of effect size.[10] Significant observation chart × experience group interactions were followed-up with pairwise comparisons between observation charts within each experience group. For these analyses, the Bonferroni-Holm correction for multiple comparisons was used to ensure that the familywise error rate did not exceed $p < .05$.[11] We also conducted simple effects tests to compare the experience groups' performance on each chart, using *Cohen's d* to quantify effect size.[12]

3. Results

3.1. Participant characteristics

Over 80% of participants reported using observation charts more than once a day as part of their current role (see Table 1 for participants' prior experience with each chart presented in the study). Table 4 presents detailed participant characteristics for both chart experience groups.

Table 4
Participant characteristics. Values are mean (*SD*) or percentage (*n*).

Variable		Experience group	
		Multiple parameter track-and-trigger chart experienced participants (*n* = 64)	**No track-and-trigger graphical chart experienced participants (*n* = 37)**
Age in years		40.17 (12.21)	37.35 (9.98)
Gender	Female	82.8% (53)	86.5% (32)
	Male	17.2% (11)	13.5% (5)
Years registered		15.37 (12.21)	13.46 (10.25)
Occupation	Doctor	1.6% (1)	5.4% (2)
	Nurse	98.4% (63)	94.6% (35)
Work area	Ward	59.4% (38)	40.5% (15)
	Emergency	4.7% (3)	13.5% (5)
	Theatre	3.1% (2)	-
	ICU	14.1% (9)	16.2% (6)
	Multiple areas	7.8% (5)	5.4% (2)
	Other	10.9% (7)	24.3% (9)
Frequency of observation chart use	More than once a day	82.8% (53)	75.7% (28)
	Once a day	3.1% (2)	13.5% (5)
	More than once a week, but less than once a day	10.9% (7)	5.4% (2)
	More than once a month, but less than once a week	-	2.7% (1)
	Less than once a month	3.1% (2)	2.7% (1)
Frequency of recording information in observation charts	More than once a day	75.0% (48)	78.4% (29)
	Once a day	6.3% (4)	-
	More than once a week, but less than once a day	4.7% (3)	2.7% (1)
	Once a week	3.1% (2)	2.7% (1)
	More than once a month, but less than once a week	-	5.4% (2)
	Less than once a month	4.7% (3)	5.4% (2)
	Not applicable	6.3% (4)	5.4% (2)
Training received in observation chart use †	None	-	16.2% (6)
	Read the instructions	25% (16)	24.3% (9)
	Informal (e.g., by co-worker)	34.4% (22)	62.2% (23)
	Formal (e.g., in-service or workshop)	89.1% (57)	29.7% (11)
	Other	1.6% (1)	-

† For this question, participants could select more than one form of training

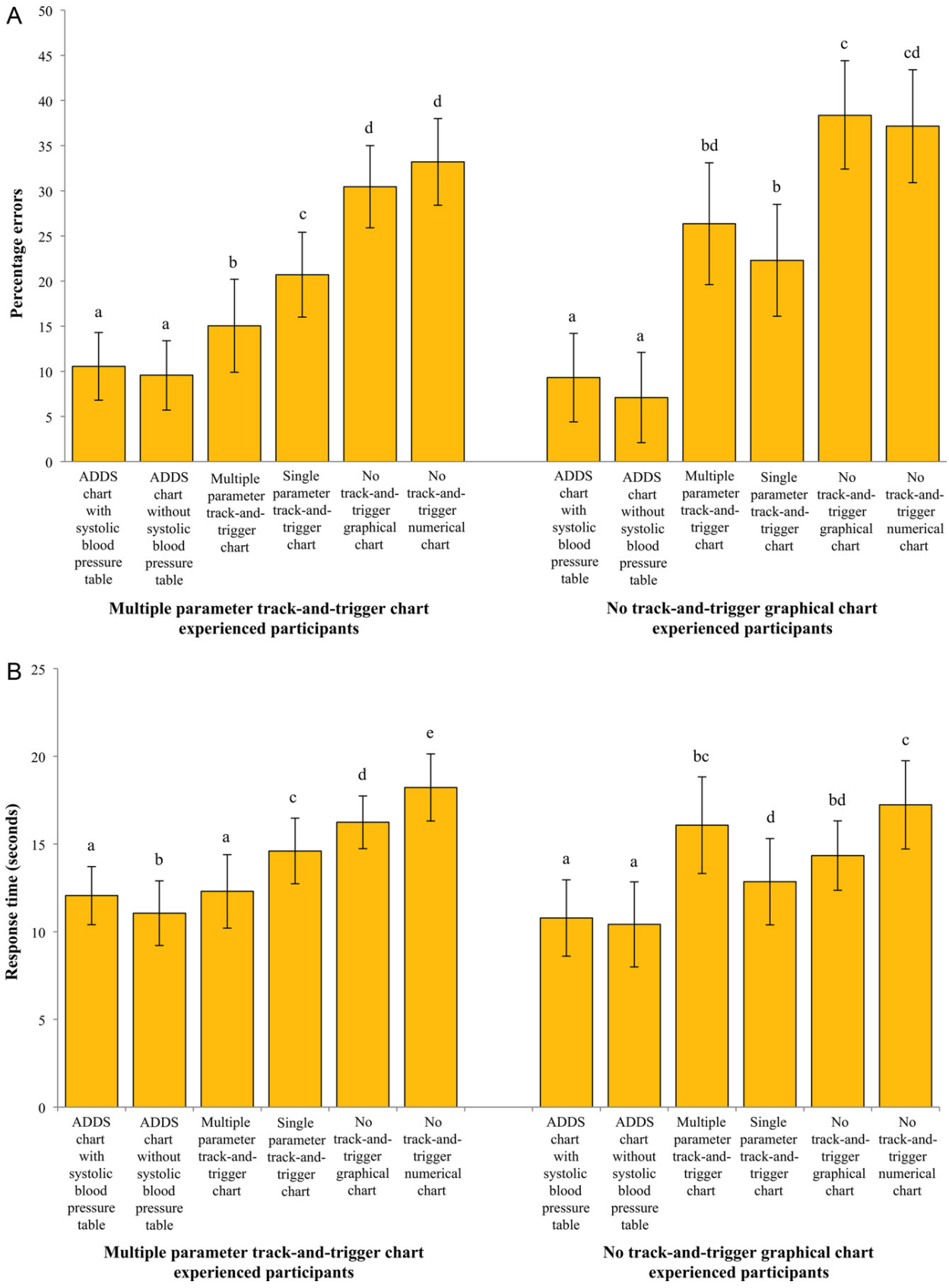
3.2. Error rate

Analysis of the error rate data revealed a significant main effect of chart type, $F(3.96, 391.81) = 63.16$, $p < 0.001$, $\eta^2 = 0.390$, no significant main effect of experience group, $F(1, 99) = 1.78$, $p = 0.186$, and a significant interaction between chart type and chart experience group, $F(3.96, 391.81) = 3.40$, $p = 0.010$, $\eta^2 = 0.030$ (see Figure 2 and online supplementary material for pairwise comparisons between charts). Simple effects tests revealed two significant differences between the experience groups. The multiple parameter track-and-trigger chart experienced participants made fewer errors than the no track-and-trigger graphical chart experienced participants on both their own chart, $t(99) = -2.63$, $p = 0.010$, Cohen's $d = -0.55$, and the no track-and-trigger graphical chart, $t(99) = -2.09$, $p = 0.040$, Cohen's $d = -0.43$.

3.3. Response time

For the response time data, there was a significant main effect of chart type, $F(2.83, 279.82) = 37.17$, $p < 0.001$, $\eta^2 = 0.270$, no significant effect of experience group, $F(1, 99) = 0.13$, $p = 0.723$, and a significant interaction between chart type and chart experience group, $F(2.83, 279.82) = 6.42$, $p < 0.001$, $\eta^2 = 0.060$ (see Figure 2 and online supplementary material for pairwise comparisons between charts). Simple effects tests revealed one significant difference between the groups. The multiple parameter track-and-trigger chart experienced participants responded faster than the no track-and-trigger graphical chart experienced participants on their own chart, $t(99) = -2.16$, $p = 0.033$, Cohen's $d = -0.45$.

Fig. 2. Error rates (A) and response times (B) for detecting abnormal observations on the six charts, arranged by experience group. Error bars indicate 95% confidence intervals. Within each group, different letters indicate significant differences at the 5% level using the Bonferroni-Holm correction.



4. Discussion

In line with the prediction that human factors design would outweigh health professionals' prior experience with a particular observation chart (*Hypothesis 2*), both chart experience groups made fewer errors and responded more quickly when using the ADDS rather than the other designs, including the charts that they were highly experienced in using. Since the ADDS out-performed all other charts on both metrics, a speed-accuracy trade-off cannot account for its success on either measure. Compared with the best-performing ADDS chart, the multiple parameter track-and-trigger chart yielded around 1.6 times as many errors by experienced users. Likewise, the no track-and-trigger graphical chart yielded around 5.4 times as many errors by participants experienced with a similar chart. These are large effects that, in practice, would be likely to influence the appropriate and timely detection of patient deterioration.

Nevertheless, the study also yielded some evidence of the benefits of experience with a particular chart. When using the multiple parameter track-and-trigger chart, participants with prior experience were both faster and more accurate than their counterparts in the no track-and-trigger graphical chart group. Unexpectedly, they were also more accurate in their ability to recognise abnormal patient observations on the no track-and-trigger graphical chart. It is hypothesised that this performance advantage may stem from The Canberra Hospital's interdisciplinary staff education program, which explicitly aims to enhance understanding of patient deterioration and the significance of abnormal observations.[13] Over 89% of participants experienced with the multiple-parameter track-and-trigger chart reported receiving this training which, in concert with the re-design of their observation chart (into its current form) and the implementation of a medical response system, appears to have improved the process of recognising clinical deterioration in their hospital.[4] In comparison, only 29% of no track-and-trigger graphical chart experienced participants reported receiving any type of formal chart training. However, future studies examining the effects of chart training on user performance would be required to test our hypothesis directly.

One limitation of the current study is that the design of the no track-and-trigger graphical chart differed slightly from that of the chart routinely used by participants from the Mt Isa Base and Logan hospitals. However, it was reasoned that the design differences between the two charts (e.g., paper size, presentation of respiratory rate, inclusion of written emergency call criteria on the reverse; see Table 3) were not substantial enough to significantly disadvantage these participants, and were trivial in comparison to the design differences between either chart and the ADDS charts.

The results of the current study suggest that the performance benefits associated with human factors designed observation charts can outweigh the potential negative effects of abandoning and replacing a chart that is highly familiar to health professionals in an institution. Rather than

disadvantage staff, the findings of Preece et al.[1] and the current study suggest that implementation of the Adult Deterioration Detection System charts may actually lead to performance improvements, even in health professionals whose prior chart experience is with a reasonably well-designed chart[3] (such as the multiple parameter track-and-trigger chart used in these studies). This finding is timely considering recent government initiatives to develop and implement evidence-based general adult observation charts.[7] When a new best-practice chart is introduced, one critical factor is whether it is accepted by the clinicians involved. If clinicians believe (even falsely so) that the new chart is inferior to the chart it is replacing, then this may lead to resistance to its use (resulting in problems with compliance, or even a failure to adopt the new chart at all). One potential driver of such resistance could be the assumption that staff will perform worse (at least, initially) on the new chart because of their extensive experience with the pre-existing chart. Given that the results of the present study suggest that this assumption may be unfounded (at least in the contexts tested), one solution might be to find a way of effectively communicating these findings to clinicians. For example, this information could be embedded in training materials accompanying the introduction of the new chart.[9]

Although this study demonstrates that careful consideration of observation chart design can improve user performance, it is not yet known which specific design elements are responsible for this benefit (because the charts examined in this study varied in a number of ways). Though indicative findings suggest that (1) integrated track-and-trigger systems, (2) graphical observations (especially on charts without track-and-trigger systems), and (3) grouped early warning scores may all be beneficial,[1,2,4] an objective and systematic evaluation of specific chart features is required to determine the unique contribution of these and other chart characteristics to performance.

5. Conclusion

In this study, health professionals performed better on novel well-designed charts, developed using human factors principles, than on the chart that they were experienced in using. Although there was some evidence that experience with a particular chart design can improve performance, the results also suggest that such performance increments do not adequately compensate for performance deficits attributable to chart designers' failure to effectively apply human factors principles to their designs. At least in the contexts examined, superior observation chart design appears to trump familiarity. Hence, hospitals motivated to improve the detection of patient deterioration should implement charts that have been designed from a human factors perspective and empirically evaluated through behavioural experimentation or alternative techniques that yield objective evidence.[9]

References

1. Preece MHW, Hill A, Horswill MS, Watson MO. Supporting the detection of patient deterioration: Observation chart design affects the recognition of abnormal vital signs. Resuscitation, in press.

2. Chatterjee MT, Moon JC, Murphy R, McCrea D. The "OBS" chart: An evidence based approach to re-design of the patient observation chart in a district general hospital setting. Postgrad Med J 2005;81:663-6.

3. Preece MHW, Horswill MS, Hill A, Karamatic R, Hewett D, Watson MO. Heuristic analysis of 25 Australian and New Zealand adult general observation charts. Sydney, New South Wales, Australia, Australian Commission on Safety and Quality in Health Care, 2009.

4. Mitchell IA, McKay H, Van Leuvan C, et al. A prospective controlled trial of the effect of a multi-faceted intervention on early recognition and intervention in deteriorating hospital patients. Resuscitation 2010;81:658-66.

5. Horswill MS, Preece MHW, Hill A, Christofidis MJ, Karamatic RM, Hewett DJ, Watson MO. Human factors research regarding observation charts: Research project overview. Sydney, New South Wales, Australia, Australian Commission on Safety and Quality in Health Care, 2010.

6. Royal Brisbane and Women's Hospital Metro North Health Service District. RBWH facts Queensland Government, 2009 (Accessed 13 March 2012, at http://www.health.qld.gov.au/rbwh/docs/rbwh_facts.pdf).

7. Australian Commission on Safety and Quality in Health Care. Recognising and responding to clinical deterioration: Use of observation charts to identify clinical deterioration. ACSQHC, 2009. (Accessed 13 March 2012, http://www.safetyandquality.gov.au/wp-content/uploads/2012/02/UsingObservationCharts-2009.pdf).

8. Andre AD, Wickens CD. When users want what's not best for them. Ergon Des 1995;(October):10-3.

9. Preece MHW, Hill A, Horswill MS, Karamatic R, Watson MO. Designing observation charts to optimise the detection of patient deterioration: Reliance on the subject preferences of healthcare professionals in not enough. Australian Critical Care, in press.

10. Howell DC. Statistical methods for psychology. Belmont, California, Duxbury Press, 1997.

11. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat 1979;6:65-70.

12. Cohen J. A power primer. Psychol Bull 1992;112:155-9.

13. Avard B, Slater N, McKay H, Daveson K, Lamberth P, Mitchell I. Compass "pointing you in the right direction". ACT Health, 2006. (Accessed 13 March 2012, at http://www.swarh2.com.au/assets/A/1230/Compass%20Training%20Manual.pdf)

Chapter 3

Christofidis, M.J., Hill, A., Horswill, M.S., & Watson, M.O. (2014). Observation charts with overlapping blood pressure and heart rate graphs do not yield the performance advantage that health professionals assume: an experimental study. *Journal of Advanced Nursing, 70*(3), 610-624.

Table 2. Manuscript revision history for "Observation charts with overlapping blood pressure and heart rate graphs do not yield the performance advantage that health professionals assume: an experimental study"

Date	Detail
26 January 2013	Submitted to *Journal of Advanced Nursing*
20 April 2013	Article revised
6 July 2013	Article accepted for publication
24 January 2014	Published in print

Hypotheses

Charts-users will perform significantly worse using charts with overlapping (vs. separate) plots, especially in the presence of an integrated colour-based track-and-trigger system. Chart-users who are trained to use the Seagull Sign will not perform faster nor make fewer errors when using charts with overlapping (vs. separate) plots.

INTRODUCTION

Vital sign observations can assist nurses in the detection of patient deterioration, since physiologically deranged observations may be present up to two days before an adverse event occurs (Franklin & Mathew 1994, Goldhill *et al.* 1999, Hillman *et al.* 2001, Endacott *et al.* 2007). Timely detection of abnormal vital signs is critical, as missed or delayed recognition of the deteriorating patient can lead to cardiac or respiratory arrest, emergency admission to the Intensive Care Unit, or unexpected death (Franklin & Mathew 1994; Goldhill *et al.* 1999; Hillman *et al.* 2001). In recent years, efforts have been made to improve the early detection of patient deterioration through observation chart design (Chatterjee *et al*. 2005, Preece *et al*. 2012a, Christofidis *et al*. 2013). However, some design-related issues remain contentious, none more so than the question of how blood pressure and heart rate observations should be arranged on the page (ACSQHC 2009). In Australia, for example, the most widely-endorsed observation charts can be divided into (a) those that include a separate graph for each of these vital signs (e.g., Horswill *et al.* 2010), and (b) those in which blood pressure and heart rate are plotted together on the same axes (e.g., Mitchell *et al.* 2010, ACT Health 2011a). In recent experimental studies, both clinicians and novice chart-users were consistently faster and more accurate at detecting abnormal observations on charts where these vital signs were graphed separately (Preece *et al*. 2012a, Christofidis *et al*. 2013). However, the charts also differed in other design features that are likely to have contributed to the overall performance differences. To date, no empirical study has directly assessed whether deranged blood pressure and heart rate observations can be detected more easily on separate or overlapping graphs.

Background

Until very recently, the vast majority of patient observation chart designs used in Australasian hospitals incorporated overlapping blood pressure and heart rate graphs (Preece *et al.* 2013). This being the case, it is perhaps unsurprising that Australian doctors and nurses whose opinions were surveyed in 2009 reported that they (a) preferred blood pressure and heart rate to be plotted together on the same axes, and (b) found it easier to detect patient deterioration when these vital signs were graphed together rather than separately (Preece *et al.* 2010). However, such subjective evidence is of limited practical value given that familiarity can lead people to prefer systems that actually hinder their performance (Andre & Wickens, 1995). In addition to familiarity, another key reason for this preference (Preece *et al.* 2010) may be that overlapping plots facilitate the use of a visual cue known as the ‘Seagull Sign’ (aka the ‘Portsmouth Sign’; Caballero *et al.*

2012, Morrice & Simpson 2007).

The Seagull Sign, which is widely used and endorsed by clinicians in several countries, including the United Kingdom and Australia (Darby *et al*. 2012), can only occur when systolic blood pressure and heart rate are graphed as overlapping plots on the same axes. Specifically, if a patient's heart rate (represented by a dot) is plotted higher than their systolic blood pressure (usually represented by a 'v' or an 'inverted v', according to local practice) at the same time-point, then this indicates a likely abnormality (Darby *et al.* 2012; see Figure 1A). Early detection of such vital sign derangements can lead to the initiation of appropriate clinical review and treatment, potentially reducing the risk of organ dysfunction and death (Darby *et al.* 2012). Physiologically, the Seagull Sign equates to a shock index score (i.e., heart rate ÷ systolic blood pressure; Rady *et al.* 1994, Cannon *et al.* 2009) of greater than one. There is evidence from both emergency (Rady *et al.* 1994, Cannon *et al.* 2009) and non-emergency settings (Kirkland *et al.* 2012, Sankaran *et al.* 2012) of a statistical relationship between shock index values (whether expressed as raw scores or the presence vs. absence of the Seagull Sign) and subsequent clinical deterioration (Darby *et al.* 2012). However, two recent studies found that: (a) compared with the Seagull Sign, modified early warning scores were a better predictor of unplanned ICU admissions (Ramrakha *et al.* 2012); and (b) modified shock index scores (i.e., heart rate ÷ mean arterial pressure), which have no consistent Seagull Sign equivalent, predicted emergency patient mortality in circumstances where standard shock index scores did not (Liu *et al.* 2012). Furthermore, there is no empirical evidence that the Seagull Sign itself – as a visual cue – actually assists nurses to detect deranged vital signs in practice. Rather, clinicians have merely assumed and asserted that the visual cue is quick and easy to identify (Darby *et al.* 2012), without ever testing this fundamental assumption.

Health professionals' perception of the Seagull Sign as a practically useful tool (Darby *et al.* 2012) may be partially explained by the memorable nature of the metaphor (Ortony 1993) that it represents; that is, just as it is abnormal for a patient's heart rate to be plotted above their systolic blood pressure, it is abnormal for a seagull (represented by the 'v' or 'inverted v') to defy gravity by defecating (represented by the dot) upwards. The role of metaphor in education is somewhat controversial. One the one hand, a metaphor can enable the transfer of understanding from something that is well-known to something less well-known in a vivid and memorable way, thereby enhancing efficient and effective learning. On the other hand, metaphors are not essential to a cognitive understanding of what is being taught and learned, and may encourage sloppy and misled thought (Petrie & Oshlag 1993). Nevertheless, the metaphor's utility as a teaching aid arguably contributes to some clinicians' assumption that the Seagull Sign is a readily identifiable visual cue, especially for novice chart-users (Darby *et al.* 2012).

Indeed, from a Gestalt psychology perspective (Zheng & Xue 2009), the isolated occurrence

of a heart rate observation plotted *above* a systolic blood pressure observation may be salient to chart-users because it appears visually dissimilar to the surrounding data (i.e., a series of heart rate observations plotted *below* corresponding systolic blood pressure observations). However, from a human factors perspective, a potentially more significant problem with the Seagull Sign is that the use of overlapping plots may lead to a visually cluttered display in which observations for one vital sign are difficult to separate perceptually from observations for the other (Wickens & Carswell 1995). Further, the associated processing demands may increase (Wickens & Carswell 1995) with the inclusion of an integrated colour-based alerting (or 'track-and-trigger') system, used in many observation charts to help chart-users recognise patient deterioration and respond appropriately (Preece *et al.* 2013). Therefore, the practical utility of the Seagull Sign – and the overlapping plots that it necessitates – cannot be assumed.

In light of recent government initiatives to develop and implement standardised evidence-based general adult observation charts (ACSCHC 2009), there is a pressing need to assess the efficacy of chart-related practices (Oliver *et al.* 2010, De Meester *et al.* 2012) because anecdotal information, despite its low ranking in the evidence hierarchy, can greatly influence clinical behaviour (Enkin & Jadad 1998). For instance, until recently, patient observation charts were typically designed by health professionals relying on their own experiences and subjective judgments – and those of their peers – to gauge the efficacy of their designs (Chatterjee *et al*. 2005, Preece *et al*. 2012a). Furthermore, some of the reaction to a recent effort to improve paper-based observation charts using evidence-based approaches (ACSQHC 2009) suggests that clinicians can become highly wedded to culturally-supported chart-related beliefs (Preece *et al*. 2012b). Consequently, they may resist changes to their favoured chart designs (a) without empirical support for their arguments and, more critically, (b) in the face of mounting evidence to the contrary.

THE STUDY

Aims

The present study aimed to provide the first direct empirical test of whether deranged blood pressure and heart rate observations can be detected more easily on separate or overlapping graphs. A secondary aim was to evaluate the practical utility of the Seagull Sign as a visual cue to assist in the detection of these observations. To address these aims, we tested the ability of chart-users – both 'Seagull-trained' and untrained – to recognise abnormal systolic blood pressure and heart rate observations on patient charts of varying design. A set of four chart design extracts was used in the experiment, which varied systematically in two ways: (1) the blood pressure and heart rate graphs

were either separate or overlapping; and (2) an integrated colour-based track-and-trigger system was either present or absent. We included both experienced nurse and novice groups in the sample to assess the generality of the results, since it is a practical necessity that observation charts and related practices should be effective for users with diverse levels of clinical expertise.

The study tested two hypotheses based on the human factors considerations outlined above. First, we predicted that overlapping blood pressure and heart rate plots would impede the detection of abnormal observations, such that chart users would take longer and make more errors when using charts with overlapping (vs. separate) plots, especially on charts with an integrated colour-based track-and-trigger system (*Hypothesis 1*). Second, we predicted that, even when viewing patient data that would yield the Seagull Sign if presented on overlapping plots, participants trained to use the Seagull Sign would perform no better when using charts with overlapping (vs. separate) plots (*Hypothesis 2*).

Design

The study comprised a 3x2x2 mixed factorial design experiment, with participant group (between-participants), graph format (separate vs. overlapping graphs, within-participants) and alerting system (integrated colour-based track-and-trigger system present vs. absent, within-participants) as the independent variables. We chose to vary both of the manipulated independent variables within-participants to maximize statistical power and to ensure that these factors could not be confounded by individual differences (e.g., level of expertise). The dependent measures were error rate and response time.

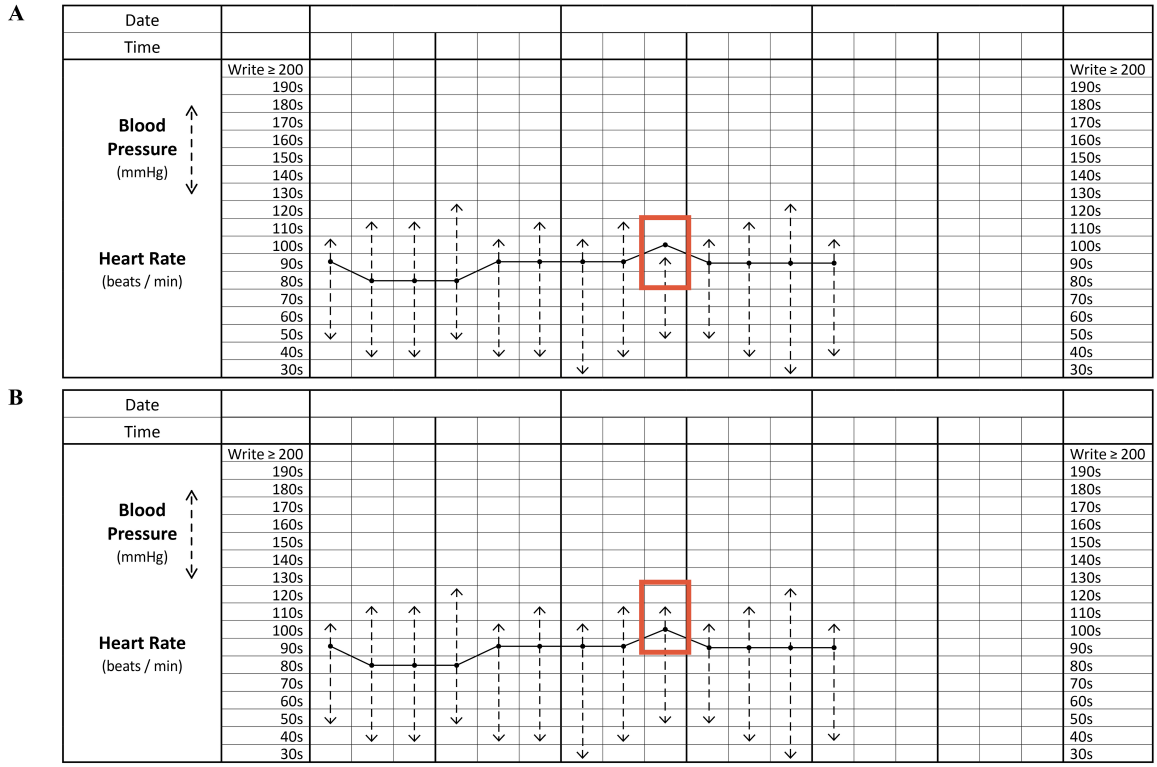
Patient data

Sixty-four cases of genuine de-identified patient data, collected from several Australian hospitals, were used in this study. Each case spanned 13 consecutive time-points and included data for the three vital signs relevant to the Seagull Sign: systolic blood pressure, diastolic blood pressure and heart rate. Half of these cases contained only normal observations (where the normal ranges were defined as systolic blood pressure from 90 to 139 mmHg, and heart rate from 50 to 99 beats per minute; ACT Health 2011b). The remainder comprised 16 hypotensive and 16 tachycardic cases, each containing one abnormal observation.

Two versions of each abnormal case were used in the study: (1) the original version, which would not yield a Seagull Sign on any chart; and (2) a slightly modified ‘Seagull Sign available’ version, which would yield a Seagull Sign when recorded on overlapping plots. This approach

ensured that cases with and without Seagull Signs were as comparable as possible. To create each modified case, one blood pressure and/or heart rate observation from the original case was shifted into one of the two adjacent range rows (see Figure 1). A Senior Medical Specialist reviewed each modified case and determined that all data were physiologically plausible. As detailed below, each participant saw only one version of each case in the experiment.

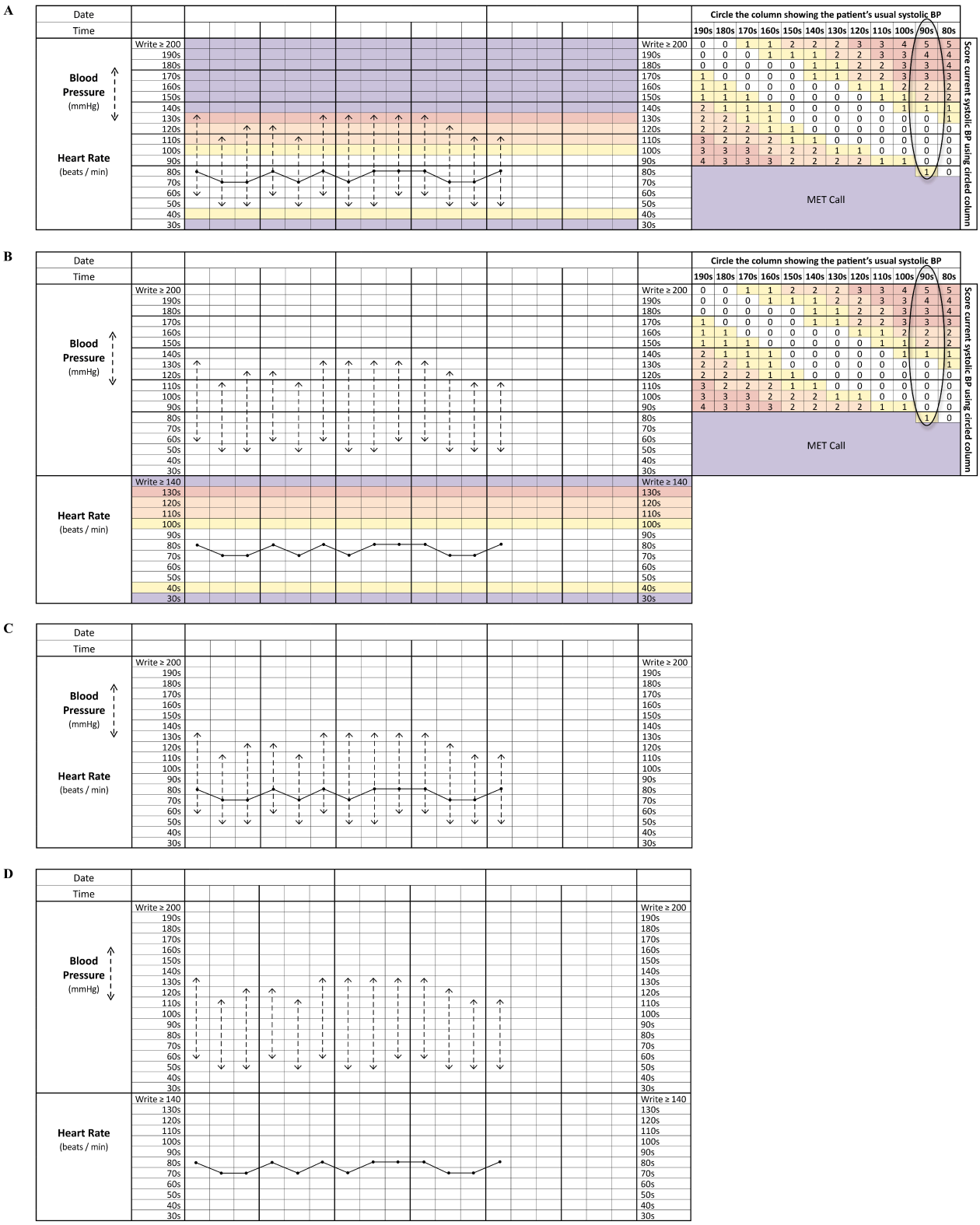
Fig. 1. An example of a tachichardic case with a Seagull Sign (A), and the equivalent case without a Seagull Sign (B), both plotted on an overlapping blood pressure and heart rate graph with no track-and-trigger system (see the *Patient data* section for details). Boxed areas highlight the Seagull Sign and the corresponding non-Seagull Sign data.



Observation chart designs

Four chart design extracts, based on observation chart designs currently used in Australia (Preece *et al.* 2013, Mitchell *et al.* 2010, Preece *et al.* 2012a), were created for use in this study (see Figure 2). Two incorporated an integrated colour-based track-and-trigger system, with either overlapping blood pressure/heart rate graphs (where the Seagull Sign could occur; Figure 2A), or separate blood pressure and heart rate graphs (Figure 2B). The others had no track-and-trigger system, and also featured either overlapping blood pressure/heart rate graphs (where the Seagull Sign could occur; Figure 2C), or separate graphs (Figure 2D).

Fig. 2. Chart design extracts used in the study: overlapping blood pressure and heart rate graphs with a track-and-trigger system (A); separate blood pressure and heart rate graphs with a track-and-trigger system (B); overlapping blood pressure and heart rate graphs with no track-and-trigger system (C); and separate blood pressure and heart rate graphs with no track-and-trigger system (D).



The chart design extracts with track-and-trigger systems included a systolic blood pressure scoring table designed to allow the patient's usual systolic blood pressure to be considered when deciding whether systolic blood pressure observations were normal or abnormal (Mitchell *et al.* 2010, ACT Health 2011a, Preece *et al.* 2012a). However, in the experiment, the patient's usual systolic blood pressure was always between 90 and 99 mmHg, which corresponded to the 90 to 139 mmHg normal range (ACT Health 2011b; see Figure 2A, 2B). This arrangement allowed colour-coding to be used for both blood pressure and heart rate observations, irrespective of whether their plots overlapped, eliminating a potential confound.

The four chart design extracts were created, and each set of patient data plotted onto each design, using Adobe InDesign CS5.5 (Adobe Systems Incorporated 2011).

Participants

A purposive sample of nurses (n = 41), who were compensated AUD75 for their time, were recruited from a tertiary hospital (ACT, Australia) via flyer advertisements. In this institution, the general observation chart incorporates overlapping blood pressure and heart rate graphs (Mitchell *et al.* 2010, ACT Health 2011a), and use of the 'Seagull Sign' in conjunction with the chart is encouraged by senior clinicians and taught as part of an interdisciplinary education program on the detection of patient deterioration available to all nursing staff (ACT Health 2011a).

Novice chart-users (n = 113) were a convenience sample of undergraduate psychology students from a Brisbane university (QLD, Australia), who received course credit for participating. The initial exclusion criterion for novices was any prior experience with a hospital observation chart. We deliberately chose to use a naïve sample for this group to ensure that particular design features could not be advantaged by participants' prior chart-related experiences or preferences. We also noted that in our previous observation chart experiments (Preece *et al.* 2012a, Christofidis *et al.* 2013), samples of health professionals and participants recruited via the psychology research participation scheme demonstrated the same (or very similar) patterns of results across charts. Hence, we reasoned that there would be no additional value in including a non-naïve novice group (e.g., nursing students or recent graduates).

In previous work using similar methods (Preece *et al.* 2012, Christofidis *et al.* 2013), a minimum sample size of approximately 40 participants per group was sufficient to yield statistically significant pairwise performance differences between alternative chart designs in every instance where the performance difference was deemed substantial enough to be of practical importance. In the present study, we therefore continued recruiting and testing until the number of participants in each group who were eligible for inclusion in the final sample exceeded this number.

Some participants were excluded from the analyses because either: (a) they scored less than 100% in the post-experiment multiple-choice examination, suggesting failure to retain some of the key background information; (b) they reported not using the Seagull Sign during the experiment despite having received 'Seagull training'; or, (c) their overall error rate exceeded 50% suggesting a lack of motivation for, or understanding of, the experimental task (see Figure 3). Excluding these participants ensured that, in the final sample, each group was comprised exclusively of individuals who understood their training, retained the key information, and complied with the task instructions. However, when the statistical analyses described below were repeated with these participants included, the overall patterns of results remained unchanged.

Data collection

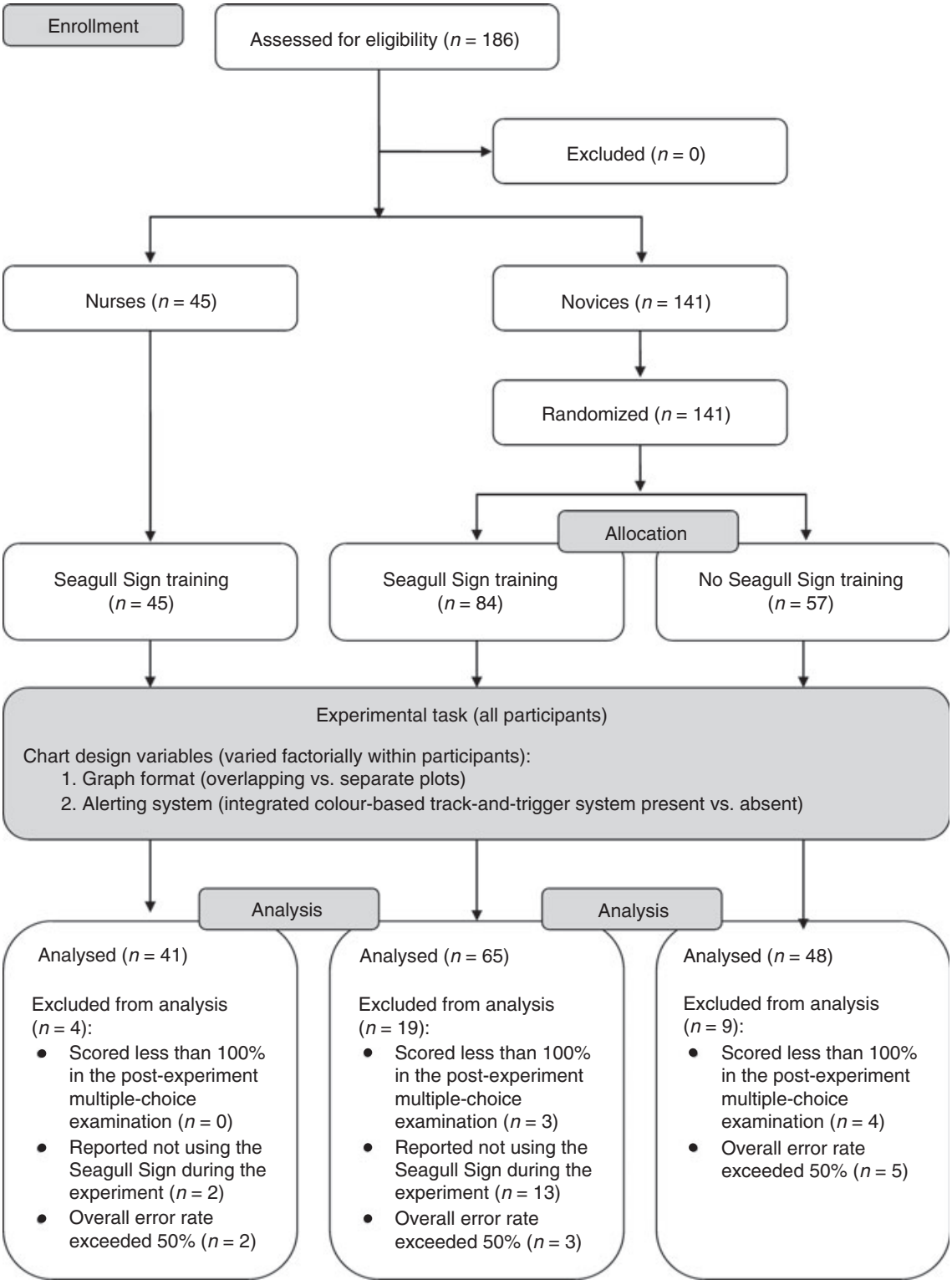
All experimental participants were recruited and tested between January and May 2011 and gave informed consent; however, they were not informed of the study's hypotheses or manipulations prior to participating. Each participant was trained and tested individually in a quiet room (i.e., a hospital training room or university laboratory). First, they completed a demographic questionnaire. Next, they watched training videos that explained: (a) systolic blood pressure and heart rate, and their normal ranges; (b) track-and-trigger systems; and (c) how to use each chart design (explained in a different random order for each participant). Novices were assigned to one of two conditions: 'Seagull-trained', or untrained (see Figure 3). Assignment was automated via an Excel spreadsheet (Microsoft Corporation 2010), created by the first author, which allocated each novice participant in turn to a training condition entirely at random. For 'Seagull-trained' novices and all nurses, the training video also explained: (d) the Seagull Sign; and (e) how to find it on charts with overlapping blood pressure/heart rate graphs.

Subsequently, a 5-item multiple-choice examination tested participants' mastery of the key background information required to participate in the study, including the normal ranges for systolic blood pressure (Q1) and heart rate (Q2), and the definitions of: cut-off scores (Q3); early warning scores (Q4); and either the Seagull Sign (for 'Seagull-trained' novices and nurses only), or observation charts (for untrained novices only) (Q5). If any item was answered incorrectly, the participant was required to study the background information from a summary sheet and retake the test until they scored 100%. Next, participants viewed a video that explained the experimental task.

Over the 64 experimental trials, each chart design appeared 16 times. For each participant, patient cases were randomly assigned to chart designs so that each case appeared only once. Constraints on randomisation ensured that, for each design, there were eight normal cases, four hypotensive cases (including two 'Seagull Sign available' cases), and four tachycardic cases

(including two 'Seagull Sign available' cases). Trials were presented in a different random order for each participant to prevent order effects.

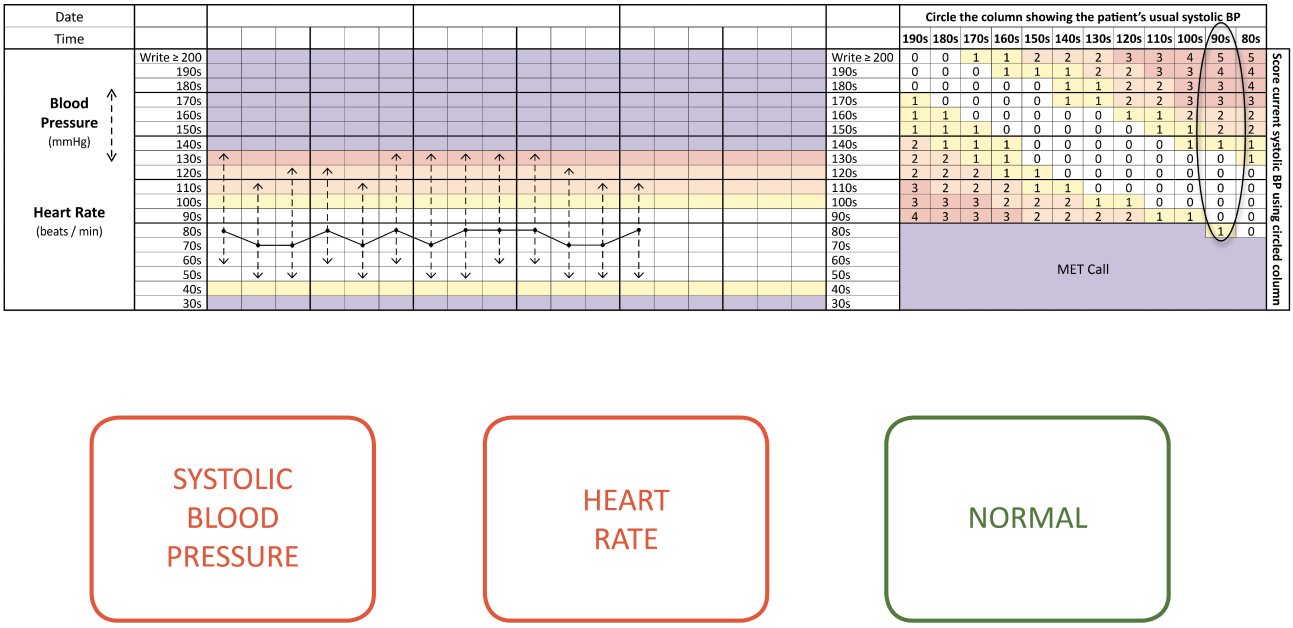
Fig. 3. Flow diagram illustrating the enrollment and allocation of participants, the within-participants experimental manipulations, and the exclusions made prior to analysis.



In each trial, the participant was presented with a chart design on a computer monitor, and responded by clicking on one of three buttons on the screen (see Figure 4): a green 'normal' button (to indicate that all observations were normal), or a red systolic blood pressure or heart rate button (to indicate an abnormality). For each trial, SuperLab experimental software (Cedrus Corporation 2007) was used to automate the process of presenting the images, and recording responses and response times (in milliseconds).

Following the experiment, participants re-sat the multiple-choice examination, so that individuals who had not retained the key background information could be excluded from the sample. This procedure ensured that failure to retain this information could not provide an alternative explanation for errors made in the experiment. In addition, 'Seagull-trained' novices and nurses reported how frequently they had utilised the Seagull Sign during the experiment.

Fig. 4. An example screen-shot from the experimental software as seen by participants, showing a chart design extract and the three response buttons.



Ethical considerations

This study was granted ethical approval in accordance with the review processes of the relevant hospital and university ethics committees.

Data Analysis

For each trial, a response was coded as 'correct' if the participant clicked on the appropriate

button. For each chart design, each participant's average response time and error rate (i.e., percentage incorrect) were calculated as the outcome measures, both overall and separately for 'Seagull Sign available' cases only.

Statistical analyses were performed using IBM SPSS 20.0 (IBM Corp., Armonk, NY: USA) with alpha set at 0.05. To test *Hypothesis 1,* separate mixed-design (participant group × graph format × alerting system) analyses of variance (ANOVAs) were conducted on overall response times and error rates, and η^2 was calculated for each significant omnibus test as the measure of effect size (η^2 indicates the proportion of within- or between-groups variance explained; Howell 1997). Significant interactions were followed up with simple effects tests, with Cohen's *d* as the effect size measure (Cohen's *d* is the difference between means in units of pooled standard deviation; Rosnow & Rosenthal 1996). To test *Hypothesis 2*, this process was repeated for each dependent measure in analyses confined to 'Seagull Sign available' cases and 'Seagull trained' nurses and novices.

RESULTS

Participant characteristics

Of the 'Seagull-trained' nurse participants in the final sample, 80.5% reported using the Seagull Sign in their current clinical role, and 92.7% had completed their institution's formal chart training program (which also included instruction on the use of the Seagull Sign). Table 1 presents detailed participant characteristics for the final sample, arranged by group.

Table 1
Participant characteristics. Except where ranges are specified, values are mean (*SD*) or percentage (*n*).

Variable		Participant group		
		'Seagull-trained' nurses (*n* = 41)	**'Seagull-trained' novices (*n* = 65)**	**Untrained novices (*n* = 48)**
Age in years		39.41 (12.50)	19.91 (3.84)	18.71 (2.16)
Age range in years		23 - 66	17 - 33	17 - 28
Gender	Female	78.0% (32/41)	73.8% (48/65)	79.2% (38/48)
	Male	22.0% (9/41)	26.2% (17/65)	20.8% (10/48)
Frequency of Seagull Sign use during experiment	All of the time	43.9% (18/41)	41.5% (27/65)	-
	Most of the time	34.1% (14/41)	38.5% (25/65)	-
	Some of the time	22.0% (9/41)	20.0% (13/65)	-
	None of the time	0.0% (0/41)	0.0% (0/65)	
Frequency of Seagull Sign use in occupational role	All of the time	43.9% (18/41)	-	-
	Most of the time	26.8% (11/41)	-	-
	Some of the time	9.8% (4/41)	-	-
	None of the time	19.5% (8/41)	-	-
Years registered		14.05 (11.06)	-	-
Work area	Ward	53.7% (22/41)	-	-
	Emergency	2.4% (1/41)	-	-
	Theatre	2.4% (1/41)	-	-
	ICU	24.4% (10/41)	-	-
	Other	17.1% (7/41)	-	-
Frequency of observation chart use in current role	More than once a day	87.8% (36/41)	-	-
	Once a day	2.4% (1/41)	-	-
	More than once a week, but less than once a day	4.9% (2/41)	-	-
	More than once a month, but less than once a week	2.4% (1/41)	-	-
	Less than once a month	2.4% (1/41)	-	-
Frequency of recording information in observation charts in current role	More than once a day	87.8% (36/41)	-	-
	Once a day	2.4% (1/41)	-	-
	Less than once a month	4.9% (2/41)	-	-
	Not applicable	4.9% (2/41)	-	-
Prior training received in observation chart use †	None	2.4% (1/41)	-	-
	Read the instructions	24.4% (10/41)	-	-
	Informal (e.g., trained by co-worker)	34.1% (14/41)	-	-
	Formal (e.g., in-service or workshop)	92.7% (38/41)	-	-

† For this question, participants could select more than one form of training.

Response time

The ANOVA on response time data for all cases revealed no significant main or interactive effect of participant group (all *p's* > 0.10; see online supplementary materials for individual group means). However, there were significant main effects of graph format, $F(1, 153) = 55.26$, $p < 0.001$, $\eta^2 = 0.27$, and alerting system, $F(1, 153) = 6.73$, $p = 0.01$, $\eta^2 = 0.05$, qualified by a significant graph format × alerting system interaction, $F(1, 153) = 9.91$, $p = 0.002$, $\eta^2 = 0.05$ [see Figure 5A(i)]. Simple effects tests revealed that participants responded faster using separate (vs. overlapping) graphs both on charts with a track-and-trigger system, $t(1, 153) = 7.10$, $p < 0.001$, Cohen's $d = 0.57$, and without, $t(1, 153) = 4.15$, $p < 0.001$, Cohen's $d = 0.33$. Separate graphs also yielded faster responses in the presence (vs. absence) of a track-and-trigger system, $t(1, 153) = -4.32$, $p < 0.001$, Cohen's $d = -0.35$.

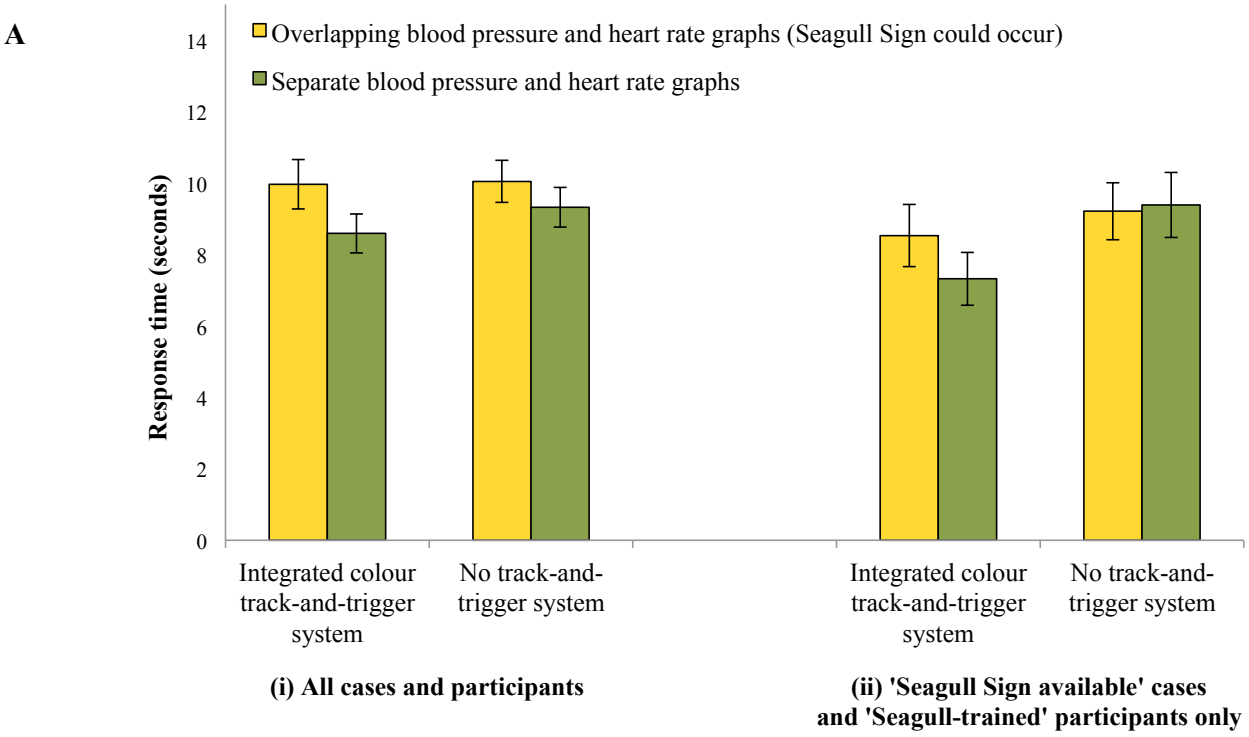
A second ANOVA, confined to 'Seagull Sign available' cases and 'Seagull-trained' nurses and novices, yielded a similar pattern of results. There were no significant main or interactive effects of participant group (all *p*'s > 0.70; see online supplementary materials for individual group means). However, there were significant main effects of graph format, $F(1, 105) = 4.74$, $p = 0.03$, $\eta^2 = 0.04$, and alerting system, $F(1, 105) = 25.41$, $p < 0.001$, $\eta^2 = 0.19$, qualified by a significant graph format × alerting system interaction, $F(1, 105) = 5.80$, $p = 0.02$, $\eta^2 = 0.05$ [see Figure 5A(ii)]. When a track-and-trigger system was present, participants responded faster using separate (vs. overlapping) graphs, $t(1, 105) = 3.25$, $p = 0.002$, Cohen's $d = 0.32$. For separate graphs, participants also responded faster using designs with (vs. without) a track-and-trigger system, $t(1, 105) = -5.66$, $p < 0.001$, Cohen's $d = -0.55$ (see Figure 5A).

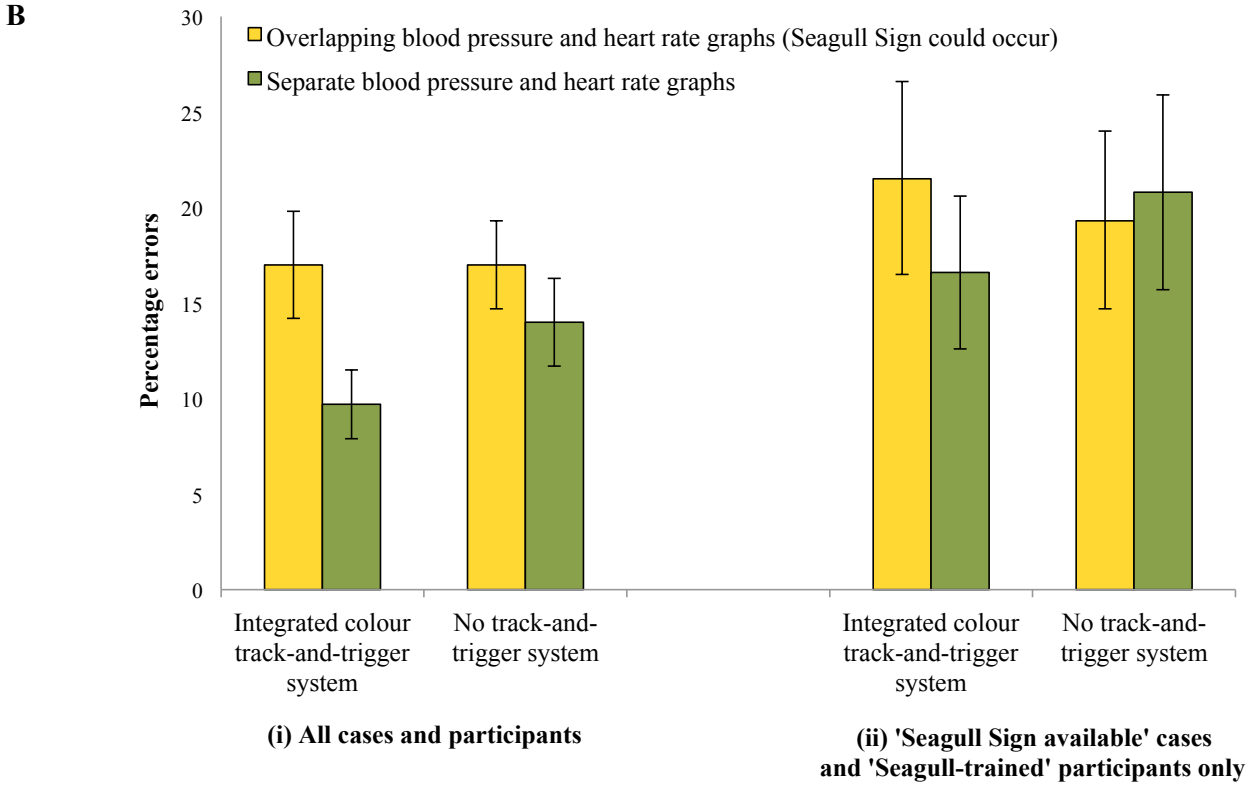
Error rate

The ANOVA on error rate data for all cases revealed no significant main or interactive effect of participant group (all *p's* > 0.10; see online supplementary materials for individual group means). However, there were significant main effects of graph format, $F(1, 153) = 60.50$, $p < 0.001$, $\eta^2 = 0.26$, and alerting system, $F(1, 153) = 7.20$, $p = 0.008$, $\eta^2 = 0.04$, qualified by a significant graph format × alerting system interaction, $F(1, 153) = 11.37$, $p = 0.001$, $\eta^2 = 0.06$ [see Figure 5B(i)]. Simple effects tests revealed that participants made fewer errors using separate (vs. overlapping) graphs, both on charts with a track-and-trigger system $t(1, 153) = 6.78$, $p < 0.001$, Cohen's $d = 0.55$, and without, $F(1, 153) = 4.05$, $p < 0.001$, Cohen's $d = 0.33$. Separate graphs also yielded fewer errors in the presence (vs. absence) of a track-and-trigger system, $t(1, 153) = -5.12$, $p < 0.001$, Cohen's $d = -0.41$.

Another ANOVA, restricted to ‘Seagull Sign available’ cases and ‘Seagull-trained’ novices and nurses, yielded one significant result (all other p’s > 0.05): a graph format × alerting system interaction, $F(1, 105) = 4.15$, $p = 0.04$, $\eta^2 = 0.04$ [see Figure 5B(ii)]. Specifically, when a track-and-trigger system was present, participants made fewer errors using separate (vs. overlapping) graphs, $t(1, 105) = 1.99$, $p = 0.049$, Cohen’s $d = 0.19$. Again, separate graphs also yielded fewer errors on designs with (vs. without) a track-and-trigger system, $t(1, 105) = -2.10$, $p = 0.04$, Cohen’s $d = -0.20$.

Fig. 5. Response times (A) and error rates (B) for detecting abnormal systolic blood pressure and heart rate observations, arranged by track-and-trigger system and graph format: for all cases and participants (i); and for 'Seagull Sign available' cases and 'Seagull-trained' participants only (ii). Error bars indicate 95% confidence intervals.





DISCUSSION

Empirical evidence from this carefully-controlled human performance experiment suggests that, irrespective of their level of clinical experience, chart users detect abnormal blood pressure and heart rate observations more accurately and efficiently when they are plotted on separate (rather than overlapping) graphs, especially on chart designs that incorporate an integrated colour-based track-and-trigger system. The results also conflict with a culturally-supported belief shared by many health professionals, namely the assumption that the observation chart 'Seagull Sign' is an easily identifiable visual cue that improves chart-users' practical ability to detect patient deterioration (Darby *et al.* 2012). Some clinicians have used the potential availability of the Seagull Sign as justification for endorsing chart designs that incorporate overlapping blood pressure and heart rate graphs (Darby *et al.* 2012). However, in the present sample, both novices (whether 'Seagull-trained' or not) and 'Seagull-trained' nurses responded faster and made fewer errors in identifying abnormal observations when vital signs were presented on separate graphs, performing best of all when a track-and-trigger system was also present. Even when only (a) participants who had received Seagull Sign training and (b) patient cases that could actually yield a Seagull Sign were considered, no advantage of overlapping plots was found. Rather, participants still performed best (in terms of both response time and accuracy) when the patient data appeared on separate graphs with a track-and-trigger system.

These findings suggest that the Seagull Sign does not yield the performance advantage that some health professionals assume (Darby *et al.* 2012), even under optimal conditions in which chart-users: (a) are familiar with the Seagull Sign; (b) are provided with specific Seagull Sign training immediately prior to testing; (c) are alerted to the Seagull Sign's likely presence during the experiment; (d) report having actively searched for the Seagull Sign during the experiment; and (e), in relation to the nurses: (i) are experienced in using an observation chart with overlapping blood pressure and heart rate graphs; (ii) work in an institution where use of the Seagull Sign has strong cultural support, and (iii), in 92.7% of cases, have previously completed a substantial formal chart education program that incorporated additional Seagull Sign training (ACT Health 2011a).

The results described above are consistent with our predictions that overlapping blood pressure and heart rate plots would hinder users' performance by producing a visually cluttered display in which observations for one vital sign were obscured by observations for the other (*Hypothesis 1*), and that use of the Seagull Sign would not countermand this disadvantage (*Hypothesis 2*). From a human factors perspective, it may have been difficult for chart-users to perceptually parse the systolic blood pressure and heart rate observations from one another (Wickens & Carswell 1995).

This is the first study to evaluate the effect of graph format and the Seagull Sign on the ability of nurses and novice chart-users to recognise derangements in systolic blood pressure and heart rate observations. The study had a number of significant methodological strengths. First, we used a standardised experimental paradigm that: (a) can be replicated precisely; (b) involves careful manipulation of key independent variables, allowing conclusions to be drawn about cause-and-effect; and (c) produces relatively clean data uncontaminated by extraneous factors that may also influence the detection of patient deterioration in the ward (e.g., distractions and interruptions). Second, we used genuine patient data to ensure that the results were as generalisable as possible given the laboratory-based nature of the testing. Third, we employed a range of careful experimental controls, including random assignment of novices to training groups, and the use of 'Seagull Sign available' and standard cases that were as equivalent to one another as possible. Finally, we measured both accuracy and response time, allowing us to check for trade-offs that could potentially have complicated interpretation of the results.

With regard to chart design, some health professionals have argued that blood pressure and heart rate should be recorded on the same axes because the Seagull Sign: (a) is quick and easy to detect, (b) does not require recall of trigger values, and (c) does not require mental calculations (e.g., summation of early warning scores) (Darby *et al.* 2012). However, our findings suggest that blood pressure and heart rate observations should be plotted separately, precluding the use of the Seagull Sign despite any predictive power it may have in a strictly statistical sense (e.g., Rady *et al.* 1994, Cannon *et al.* 2009, Darby *et al.* 2012). Further, we suggest that any ability of the Seagull Sign to alert users to deterioration is made redundant by the implementation of an effective early warning scoring system, such as a well-designed integrated colour-based track-and-trigger system. This is because the physiological values that yield a Seagull Sign would also activate the track-and-trigger system in almost all clinical scenarios. To illustrate this point, consider Figure 2B, which was based on one of the Adult Deterioration Detection System, or ADDS, charts developed by Horswill *et al.* 2010 for the Australian Commission on Safety and Quality in Healthcare, for nation-wide implementation. On this chart, the only exception would be a patient with a low usual systolic blood pressure of 80-89 mmHg, who presents with a heart rate of 90-99 beats per minute and a systolic blood pressure of 80-89 mmHg. Indeed, the Adult Deterioration Detection System charts, which adhere to both of our usability recommendations, have been shown to facilitate fast and accurate detection of patient deterioration among both novice chart-users and health professionals (Preece *et al.* 2012a), regardless of their prior chart experience (Christofidis *et al.* 2013). An additional advantage of early warning scoring systems is that they have been shown to empower nurses by giving them an unambiguous and concise means of communicating deterioration (Andrews & Waterman 2005).

Limitations

A limitation of the study is that we have not directly demonstrated that the results generalise to genuine clinical environments by, for example, conducting a multi-site clinical trial of alternative chart designs in conjunction with a randomised controlled trial of a Seagull Sign training intervention. However, given that the real-world conditions would almost certainly be less optimal for use of the Seagull Sign than those of the present study, it is arguably unlikely that overlapping blood pressure and heart rate plots would be demonstrably beneficial, even for those with Seagull Sign training.

Another limitation relates to the representativeness of our purposive nursing sample. If most of the individuals who volunteered to participate were especially motivated by an interest in patient deterioration, then the sample may have been above average in knowledge and diligence compared with the Australian nursing workforce in general (Preece *et al*. 2010). In contrast, if the novice participants were motivated primarily by the incentive offered, then they may have been less attentive on average than the nurses during the experimental task. However, given that (a) the graph format and alerting system variables were varied within-participants, and (b) the same patterns of results were obtained for nurses and novices (as in previous similar studies, e.g., Preece *et al.* 2012a), it is unlikely that sampling issues had a meaningful impact on our findings.

We also acknowledge that the findings of the present study may only apply to static domains, such as paper-based observation charts. Future studies would be required to evaluate the efficacy of the Seagull Sign in a dynamic display (e.g., an electronic vital sign monitor that presented blood pressure and heart rate data graphically), where overlapping plots could be potentially be made more discriminable by source differences such as distinct colours or differential motion (Wickens & Carswell 1995). Nevertheless, although hospitals will inevitably shift towards using electronic displays, paper-based observation charts are still likely to: (a) have a substantial shelf-life in developing countries; and (b) be retained as back-up for computer-based systems. In the latter case, it will become even more critical for paper-based charts to incorporate design features that support novice users in detecting and responding to physiological deterioration (e.g., in the context of this study, separate blood pressure and heart rate graphs). This is because, eventually, even highly experienced health professionals will not have had extensive practice using paper-based charts.

CONCLUSION

Like other recent work (Chatterjee *et al.* 2005, Preece *et al.* 2012a, Christofidis *et al.* 2013),

this study demonstrates that an evidence-based approach to chart design can improve the detection of patient deterioration. More generally, the results also illustrate the need for health professionals to assess the efficacy of chart-based practices through empirical evaluation, rather than relying on anecdotal information (Enkin & Jadad 1998).

References

ACT Health [Internet] (June 2011a) *Early recognition of deteriorating patient.* COMPASS. ACT Health. Available from: http://health.act.gov.au/professionals/general-information/compass/

ACT Health [Internet] (March 2011b) *Vital signs and early waning scores policy.* ACT Health. Available from: http://health.act.gov.au/professionals/general-information/compass/

Andre AD & Wickens CD (1995) When users want what's not best for them. *Ergonomics in Design* October 1995, 10-3.

Andrews T & Waterman H (2005) Packaging: a grounded theory of how to report physiological deterioration effectively. *Journal of Advanced Nursing* 52, 473-81. doi: 10.1111/j.1365-2648.2005.03615

Australian Commission on Safety and Quality in Health Care [Internet] (March 2009) *Recognising and responding to clinical deterioration: Use of observation charts to identify clinical deterioration.* Available from: http://www.safetyandquality.gov.au/wp-content/uploads/2012/02/UsingObservationCharts-2009.pdf.

Caballero C, Creed F, Gochmanski C & Lovegrove J (2012) *Nursing OSCEs: A Complete Guide to Exam Success.* Oxford, Oxford University Press.

Cannon CM, Braxton CC, Kling-Smith M, Mahnken JD, Carlton E & Moncure M (2009) Utility of the shock index in predicting mortality in traumatically injured patients. *The Journal of Trauma* 67, 1426-30. doi: 10.1097/TA.0b013e3181bbf728

Chatterjee MT, Moon JC, Murphy R & McCrea D (2005) The "OBS" chart: An evidence based approach to re-design of the patient observation chart in a district general hospital setting. *Postgraduate Medical Journal* 81, 663-6. doi: 10.1136/pgmj.2004.031872

Christofidis MJ, Horswill MS, Hill A & Watson MO (2013). A human factors approach to observation chart design can trump health professionals' prior chart experience. *Resuscitation* 84, 657-665. doi: 10.1016/j.resuscitation.2012.09.023

Darby B, Mitchell I, Van Leuvan C, Kingsbury A & McKay H (2012) *Seagulls could save lives.* In: Official 2012 Program of the 7th International Conference on Rapid Response Systems and Medical Emergency Teams; Sydney, New South Wales, Australia. p. 56.

De Meester K, Van Bogaert P, Clarke SP & Bossaert L, forthcoming 2012 July. In-hospital mortality after serious adverse events on medical and surgical nursing units: a mixed methods study. *Journal of Clinical Nursing,* in press. doi: 10.1111/j.1365-2702.2012.04154.

Endacott R, Kidd T, Chaboyer W & Edington J (2007). Recognition and communication of patient deterioration in a regional hospital: a multi-methods study. *Australian Critical Care* 20:100–5. doi: 10.1016/j.aucc.2007.05.002

Enkin MW & Jadad AR (1998) Using anecdotal information in evidence-based health care: Heresy or necessity? *Annals of Oncology* 8, 963-966.

Franklin C & Mathew J (1994). Developing strategies to prevent in hospital cardiac arrest: analyzing responses of physicians and nurses in the hours before the event. *Critical Care Medicine*

22:189–91.

Goldhill DR, White SA & Sumner A (1999). Physiological values and procedures in the 24 h before ICU admission from the ward. *Anaesthesia* 54:529–34. doi: 10.1046/j.1365-2044.1999.00837

Hillman KM, Bristow PJ, Chey T, Daffurn K, Jacques T, Norman SL, Bishop GF & Simmons G (2001). Antecedents to hospital deaths. Internal Medicine Journal 31:343–8. doi: 10.1046/j.1445-5994.2001.00077

Horswill MS, Preece MHW, Hill A, Christofidis MJ, Karamatic RM Hewett DJ & Watson MO (2010) *Human factors research regarding observation charts: Research project overview.* Sydney, New South Wales, Australia, Australian Commission on Safety and Quality in Health Care.

Howell DC (1997) *Statistical methods for psychology.* Belmont, California, Duxbury Press.

Kirkland LL, Malinchoc M, O'Bryne M, Benson JT, Kashiwagi DT, Burton MC, Varkey P & Morgenthaler TI (forthcoming 2012 July) A clinical deterioration prediction tool for internal medicine patients. *American Journal of Medical Quality.* doi:10.1177/1062860612450459

Liu Y, Liu J, Fang ZA, Shan G, Xu J, Qi Z, Zhu H, Wang Z & Yu X (2012) Modified shock index and mortality rate of emergency patients. *World Journal of Emergency Medicine* 3, 114-7. doi: 10.5847/ wjem.j.1920-8642.2012.02.006

Mitchell IA, McKay H, Van Leuvan C, Berry R, McCutcheon C, Avard B, Slater N, Neeman T & Lamberth P (2010) A prospective controlled trial of the effect of a multi-faceted intervention on early recognition and intervention in deteriorating hospital patients. *Resuscitation* 81, 658-66. doi: 10.1016/j.resuscitation.2010.03.001

Morrice A & Simpson HJ (2007) Identifying level one patients: A cross-sectional survey on an in-patient hospital population. *Intensive and Critical Care Nursing* 23, 23-32. doi:10.1016/j.iccn.2006.07.001

Oliver A, Powell C, Edwards D & Mason B (2010) Observations and monitoring: routine practices on the ward. *Paediatric Nursing*, 22, 28-32.

Ortony A (1993) *Metaphor and thought*. Cambridge, Cambridge University Press.

Petrie HG & Oshlag RS (1993) 'Metaphor and Learning', in A. Ortony (ed.) *Metaphor and Thought, 2nd edn,* pp. 579–609. Cambridge: Cambridge University Press

Preece MHW, Hill A, Horswill MS, Karamatic R, Hewett DG & Watson MO (2013) Applying heuristic evaluation to observation chart design to improve the detection of patient deterioration. *Applied Ergonomics*. doi:10.1016/j.apergo.2012.11.003

Preece MHW, Hill A, Horswill MS, Karamatic R & Watson MO (2012b) Designing observation charts to optimise the detection of patient deterioration: Reliance on the subject preferences of healthcare professionals in not enough. *Australian Critical Care.* doi:10.1016/j.aucc.2012.01.003

Preece MHW, Hill A, Horswill MS & Watson MO (2012a) Supporting the detection of patient deterioration: Observation chart design affects the recognition of abnormal vital signs. *Resuscitation*, 83, 1111-18. doi: 10.1016/j.resuscitation.2012.02.009

Preece MHW, Horswill MS, Hill A, Karamatic R & Watson MO (2010) *An online survey of health professionals' opinions regarding observation charts.* Sydney, New South Wales, Australia, Australian Commission on Safety and Quality in Health Care.

Rady MY, Smithline HA, Blake H & Nowak R (1994) A comparison of the shock index and conventional vital signs to identify acute, critical illness in the emergency department. *Annals of Emergency Medicine* 24, 685-90.

Ramrakha P, Slater N & Mitchell I (2012) *Comparing patient deterioration systems in unplanned intensive care admissions.* In: Official 2012 Program of the 7th International Conference on Rapid Response Systems and Medical Emergency Teams; Sydney, New South Wales, Australia. p. 79.

Rosnow RL & Rosenthal R (1996) Computing contrasts, effect sizes, and countermills on other people's published data: General procedures for research consumers. *Psychological Methods* 1, 331-40.

Sankaran P, Kamath AV, Tariq SM, Ruffell H, Smith AC Prentice P, Subramanian DN, Musonda P & Myint PK (2011) Are shock index and adjusted shock index useful in predicting mortality and length of stay in community-acquired pneumonia? *European Journal of Internal Medicine* 22, 282-5. doi: 10.1016/j.ejim.2010.12.009

Wickens CD & Carswell CM (1995) The proximity compatibility principle: Its psychological foundation and relevance to display design. *Human Factors* 37, 473. doi: 10.1518/001872095779049408

Zheng N & Xue J (2009) *Statistical learning and pattern analysis for image and video processing.* New York, London, Springer.

Chapter 4

Christofidis, M.J., Hill, A., Horswill, M.S., & Watson, M.O (2016). Observation chart design features affect the detection of patient deterioration: A systematic experimental evaluation. *Journal of Advanced Nursing, 72(1),* 158-172.

Table 3. Manuscript revision history for "Observation chart design features affect the detection of patient deterioration: A systematic experimental evaluation"

Date	Detail
13 January 2015	Submitted to the *Journal of Advanced Nursing*
13 August 2015	Article revised
25 August 2015	Article accepted for publication
9 November 2015	Published online

Hypotheses

Chart-users will make fewer errors and respond more quickly when using chart designs with drawn-dot observations, an integrated colour-based scoring-system and grouped scoring-rows, consistent with the apriori human factor principles that were adapted from the web and software domains.

INTRODUCTION

Inevitably, some patients will experience physiological deterioration while in hospital. Early recognition of the deteriorating patient is essential because delayed or missed recognition can result in adverse events including respiratory or cardiac arrest, unplanned admission to intensive care, and even unexpected death (Franklin & Mathew 1994, Goldhill *et al.* 1999, Hillman et al. 2001). Since deranged vital signs can signal deterioration as early as 48 hours before an adverse event (Franklin & Mathew 1994, Goldhill *et al.* 1999, Hillman *et al.* 2001, Endacott *et al.* 2007), one promising avenue for improving early recognition is to develop patient charts specifically designed to make abnormal observations easier for chart-users (including the least experienced nurses and doctors) to detect.

In recent years, clinicians and researchers in Australia and the UK have created new charts with this precise objective in mind, and have employed several techniques to examine the effects of chart design on the detection of patient deterioration, including: prospective before-and-after controlled intervention trials (Mitchell *et al.* 2010); comparative clinical evaluations (Chatterjee *et al.* 2005; Elliott *et al.* 2014); and behavioural experiments (Preece *et al.* 2012a, Christofidis *et al.* 2013; Fung *et al.* 2014). In each of these studies, the performance of chart-users (including nurses) was compared across two or more charts, and, in almost all cases, designs that included early-warning scoring-systems yielded the best results. On charts of this type, each value in a set of observations can be scored according to its degree of deviation from the normal range, and these scores totalled to obtain an "early-warning score" that summarizes the patient's overall physical condition and can be used to trigger appropriate clinical actions (Prytherch *et al.* 2005, Lawson & Peate 2009).

In experimental studies that compared multiple charts with early-warning scoring-systems, two designs consistently yielded the fastest and most accurate identification of abnormal observations (Preece *et al.* 2012a, Christofidis *et al.* 2013), and both were versions of Horswill *et al.*'s (2010) Adult Deterioration Detection System (or ADDS) chart. This chart was designed by a multi-disciplinary team of human factors specialists and clinicians who took into account a wide range of usability considerations (Horswill *et al.* 2010, Preece *et al.* 2013). However, the precise reasons for its superior performance (and potential avenues for further improvement) remain unclear, because several design features varied unsystematically between it and the charts with which it was compared. For example, unlike some other charts with early-warning scoring-systems, the ADDS incorporates separate (vs. overlapping) blood pressure and heart rate graphs, drawn-dot observations (vs. written numbers), an integrated colour-based scoring-system (vs. a non-integrated tabular system), and scoring rows grouped together at the bottom of the page (vs. presented

separately, immediately below the corresponding vital sign data). A more recent experimental study has demonstrated that abnormal blood pressure and heart rate observations can be detected more quickly and accurately when these two vital signs are plotted separately, especially on charts with an integrated colour-based early-warning scoring-system (Christofidis *et al.* 2014). However, no empirical study to date has directly assessed the effects of other individual observation chart design features on the detection of patient deterioration.

Background

Before supplying a new medical device to the market, manufacturers must obtain empirical data to support their claims about its safety and performance (TGA 2011). However, when paper-based observation charts are designed (or re-designed), this level of evidence-based accountability is seldom demanded despite comparable potential risks to patient safety. Instead, the efficacy of patient charts is typically assessed only via subjective judgements made by the health professionals who designed them, and their colleagues (Chatterjee *et al.* 2005, Preece *et al.* 2012a). Consequently, observation chart designs (Preece *et al.* 2013), and health professionals' perceptions of good design (Preece *et al.* 2012b), can vary considerably between locations. In the absence of objective evidence, however, there is no way to ascertain which design options represent best practice (Preece *et al.* 2012b).

The traditional subjective approach to chart development is inherently risky, as mounting research evidence suggests that health professionals' preferences for particular chart features are not always consistent with objective performance data (Preece *et al.* 2012b). For instance, in a recent survey study, most health professionals reported that they preferred, and found it easier to detect patient deterioration, when blood pressure and heart rate were plotted together on the same graph (Preece *et al.* 2010). However, these opinions are at odds with more recent objective data. In Christofidis *et al.*'s (2014) experiment, overlapping blood pressure and heart rate plots actually impeded recognition of abnormal vital signs by experienced nurses and novice chart-users alike, slowing them down and increasing their error rates. It has been suggested that performance-preference dissociations like this arise due to the inordinate influence of extraneous factors, such as familiarity and aestheics, on people's judgements and preferences (Andre & Wickens 1995). Given that such dissociations occur, it is possible that charts designed and endorsed on the basis of subjective judgements have contributed to documented failures (Franklin & Mathew 1994, Goldhill *et al.* 1999, Endacott *et al.* 2007) by hospital staff to record observations correctly and to detect or anticipate deterioration.

Indeed, the results of another two experimental studies suggest that poor design decisions

may have potentially catastrophic consequences (Preece *et al.* 2012b, Christofidis *et al.* 2013). In these studies, participants were asked to detect abnormalities among vital sign observations presented on six observation charts of varying design quality, including four charts used in Australian hospitals and two versions of the ADDS chart, which had been designed as a more 'user friendly' alternative (Horswill *et al.* 2010, Preece *et al.* 2013). Both novice chart-users (Preece *et al.* 2012a) and health professionals (Preece *et al.* 2012a, Christofidis *et al.* 2013) made the least errors and responded fastest when using ADDS charts. These effects even held for clinicians who had prior clinical experience with one of the other charts used in the experiment, or a similar design (Christofidis *et al.* 2013). In fact, compared with the ADDS charts, the worst-performing design yielded up to 5.4 times as many errors by nurses and doctors who were experienced with a similar chart (Christofidis *et al.* 2013). As well as illustrating the dangers of poor design, these findings suggest that clinical experience alone may not be enough to overcome design deficiencies.

Given that improved observation charts could potentially deliver substantial patient safety gains, it is crucial that we develop a clear and thorough understanding of how precisely their design can be optimized. However, in past studies comparing the detection of deterioration across two or more charts (e.g. Chatterjee *et al.* 2005, Mitchell *et al.* 2010, Preece *et al.* 2012a, Christofidis *et al.* 2013), the unique contributions of specific design features to the outcomes were unclear, because the charts varied unsystematically on more than one dimension. For instance, we cannot infer that every design feature included in the ADDS chart positively contributed to its superior performance (Preece *et al.* 2012a, Christofidis *et al.* 2013). Rather, there may be room for further improvement and, in some cases, health professionals' subjective preferences might still lead to better detection of patient deterioration. After all, even human-factors based chart design involves opinion-based compromises between competing design considerations (Preece *et al.* 2013). Hence, without systematic and objective comparisons, the efficacy of individual design features cannot be determined.

THE STUDY

Aims

This study aimed to systematically evaluate three design features that vary across Australasian charts with early-warning scoring-systems (Preece *et al.* 2013). Specifically, we manipulated data-recording format (drawn dots vs. written numbers), scoring-system integration (integrated colour-based system vs. non-integrated tabular system) and scoring-row placement (grouped vs. separate). For each of these design features, the first listed alternative had been incorporated into the ADDS chart (Horswill *et al.* 2010, Preece *et al.* 2013), which was designed as

part of a national initiative to develop a standardised adult general observation form (ACSQHC 2009). Using a similar methodology to prior experimental studies (Preece *et al.* 2012a, Christofidis *et al.* 2013, Christofidis *et al.* 2014), we evaluated each feature by testing charts-users' ability to recognise abnormal observations on eight chart designs representing a factorial combination of these alternatives. In line with recent indicative findings (Chatterjee *et al.* 2005, Mitchell *et al.* 2010, Preece *et al.* 2012a, Christofidis *et al.* 2013; Fung *et al.* 2014) and the human-factors-based design choices made in the development of the ADDS chart (Horswill *et al.* 2010, Preece *et al.* 2013), we predicted that chart-users would be faster and more accurate when using chart designs with: drawn-dot observations (Hypothesis 1); an integrated colour-based scoring-system (Hypothesis 2); and grouped scoring-rows (Hypothesis 3).

Design

The study employed a 2x2x2x2 mixed factorial design with data-recording format, scoring-system integration, and scoring-row placement varied within-participants. In addition, the presence vs. absence of scores (i.e. overall early-warning scores, and the scores for individual vital signs from which they are derived) was varied between-participants (see *Scores* for details and rationale). The dependent measures were response time and error rate.

Patient data

To ensure content validity, sixty-four cases of genuine de-identified patient data, each spanning 13 consecutive time-points, were used in the study. The cases, which were collected from several Australian hospitals, included data for the ten parameters included in the ADDS chart (Horswill *et al.* 2010): respiratory rate, oxygen delivery, oxygen saturation, systolic and diastolic blood pressure, heart rate, temperature, four hour urine output, consciousness and pain. Half of the cases contained only normal observations (see Table 1 for normal ranges: ACT Health 2011), and the others each included one abnormal observation: a derangement in oxygen saturation (8 hypoxic cases), systolic blood pressure (4 hypotensive and 4 hypertensive cases), heart rate (4 bradycardic and 4 tachycardic cases) or temperature (4 hypothermic and 4 febrile cases).

The original data were only modified if either: (a) a vital sign remained abnormal for more than one time-point (excess abnormal data-points were shifted into the normal range); or (b) a data-point was missing (a plausible value was extrapolated or interpolated). Most of the cases (75%) had been used in prior studies employing a similar experimental paradigm (Preece *et al.* 2012a, Christofidis *et al.* 2013).

Table 1
Vital sign normal ranges used in the experiment (table adapted from Preece *et al.* 2012a).

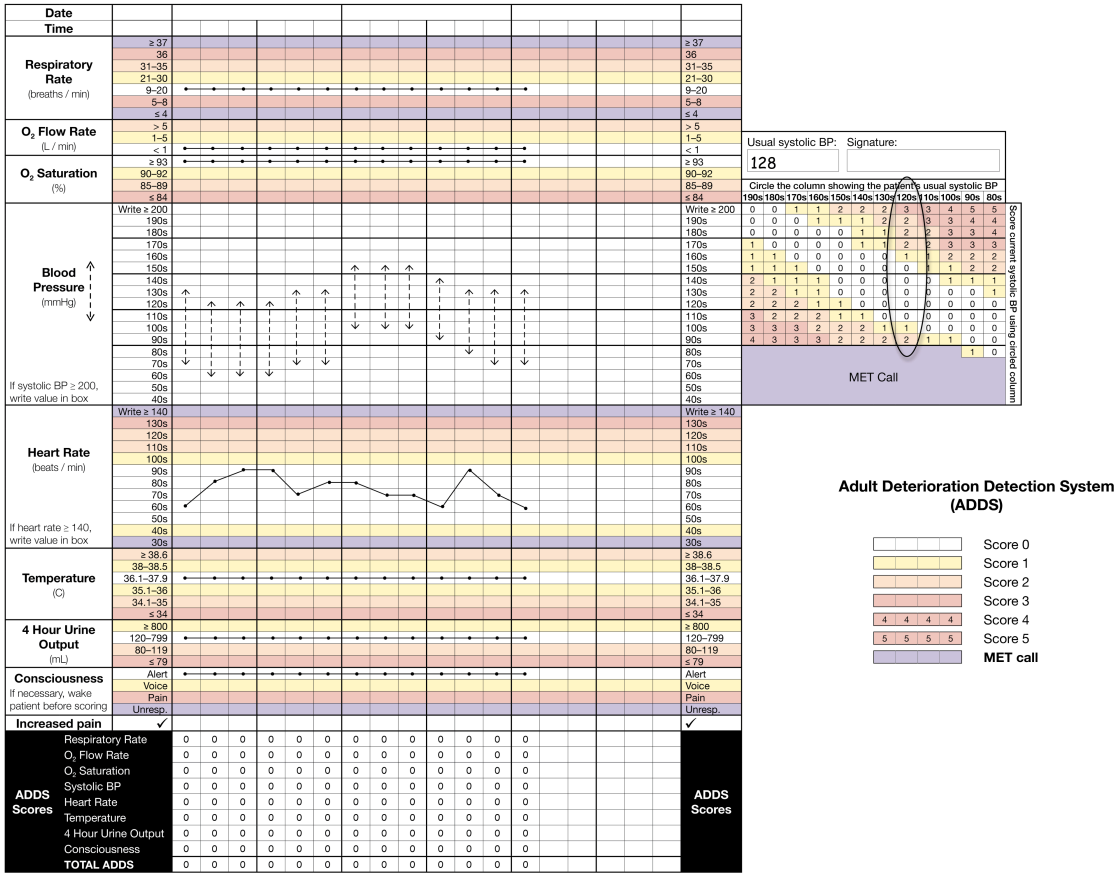
Vital sign	Normal range
Respiratory rate	9 – 20 breaths per minute
Oxygen delivery	Patient is receiving oxygen at ≤ 1 litre per minute
Oxygen saturation	93 – 100%
Systolic blood pressure	100 – 160 mmHg
Heart rate	50 – 100 beats per minute
Temperature	36.1 – 37.9 Celsius
Four hour urine output	120 – 799 mL
Consciousness	Patient is classified as being alert
Pain	Patient is in no pain

Observation chart designs

The eight observation charts created for this study, which were based on a version of the ADDS chart (Horswill *et al.* 2010), represented a factorial combination of two options for each of three design features, namely: (1) data-recording format (drawn dots vs. written numbers); (2) scoring-system integration (integrated colour-based system vs. non-integrated tabular system); and (3) scoring-row placement (grouped vs. separate) (Figure 1). Apart from these manipulations, the charts were identical. The designs were created, and each set of patient data plotted onto each design, using Adobe InDesign CS5.5 (Adobe Systems Incorporated, 2011).

Fig. 1. Four examples of chart designs used in the study, with: (a) an integrated colour-based scoring-system and grouped scoring-rows; (b) an integrated colour-based scoring-system and separate scoring-rows; (c) a non-integrated tabular scoring-system and grouped scoring-rows; and (d) a non-integrated tabular scoring-system and separate scoring-rows. Each example includes either drawn-dot (a and d) or written-number (b and c) observations. The remaining four designs were identical, except that each used the alternative data-recording format option.

(a)



(b)

Date																			
Time																			
Respiratory Rate (breaths / min)	≥ 37																		≥ 37
	36																		36
	31–35																		31–35
	21–30																		21–30
	9–20	20	20	18	18	19	18	18	20	20	20	20	20	16					9–20
	5–8																		5–8
	≤ 4																		≤ 4
		0	0	0	0	0	0	0	0	0	0	0	0	0					
O_2 Flow Rate (L / min)	> 5																		> 5
	1–5																		1–5
	< 1	0	0	0	0	0	0	0	0	0	0	0	0	0					< 1
		0	0	0	0	0	0	0	0	0	0	0	0	0					
O_2 Saturation (%)	≥ 93	98	98	98	98	98	98	97	98	98	98	98	97	94					≥ 93
	90–92																		90–92
	85–89																		85–89
	≤ 84																		≤ 84
		0	0	0	0	0	0	0	0	0	0	0	0	0					
Blood Pressure (mmHg) ↑ ↓	Write ≥ 200																		Write ≥ 200
	190s																		190s
	180s																		180s
	170s																		170s
	160s																		160s
	150s							158	159	150									150s
	140s										140								140s
	130s	130				132	137					130	135	132					130s
	120s		128	126	124														120s
	110s																		110s
	100s							102	108	101									100s
	90s										97								90s
	80s											80							80s
	70s	72				72	70							74	72				70s
	60s		68	68	68														60s
If systolic BP ≥ 200, write value in box	50s																		50s
	40s																		40s
		0	0	0	0	0	0	0	0	0	0	0	0	0					
Heart Rate (beats / min)	Write ≥ 140																		Write ≥ 140
	130s																		130s
	120s																		120s
	110s																		110s
	100s																		100s
	90s			91	99							98							90s
	80s		80				80	80											80s
	70s					78			70	70				78					70s
	60s	65									60			63					60s
If heart rate ≥ 140, write value in box	50s																		50s
	40s																		40s
	30s																		30s
		0	0	0	0	0	0	0	0	0	0	0	0	0					
Temperature (C)	≥ 38.6																		≥ 38.6
	38–38.5																		38–38.5
	36.1–37.9	36.3	36.3	36.5	36.1	36.1	36.6	36.1	36.1	36.1	36.4	36.8	36.2	36.5					36.1–37.9
	35.1–36																		35.1–36
	34.1–35																		34.1–35
	≤ 34																		≤ 34
		0	0	0	0	0	0	0	0	0	0	0	0	0					
4 Hour Urine Output (mL)	≥ 800																		≥ 800
	120–799	500	513	178	364	700	229	683	478	600	566	412	423	420					120–799
	80–119																		80–119
	≤ 79																		≤ 79
		0	0	0	0	0	0	0	0	0	0	0	0	0					
Consciousness If necessary, wake patient before scoring	Alert	A	A	A	A	A	A	A	A	A	A	A	A	A					Alert
	Voice																		Voice
	Pain																		Pain
	Unresp.																		Unresp.
		0	0	0	0	0	0	0	0	0	0	0	0	0					
Increased pain	✓																		✓
TOTAL ADDS		0	0	0	0	0	0	0	0	0	0	0	0	0					

Usual systolic BP: 128 Signature:

Circle the column showing the patient's usual systolic BP

	190s	180s	170s	160s	150s	140s	130s	120s	110s	100s	90s	80s
Write ≥ 200	0	0	1	1	2	2	2	3	3	4	5	5
190s	0	0	0	1	1	1	2	2	3	3	4	4
180s	0	0	0	0	0	1	1	2	2	3	3	4
170s	1	0	0	0	0	1	1	2	2	3	3	3
160s	1	1	0	0	0	0	0	1	1	2	2	2
150s	1	1	1	0	0	0	0	0	1	1	2	2
140s	2	1	1	1	0	0	0	0	0	1	1	1
130s	2	2	1	1	0	0	0	0	0	0	0	1
120s	2	2	2	1	1	0	0	0	0	0	0	0
110s	3	2	2	2	1	1	0	0	0	0	0	0
100s	3	3	3	2	2	2	1	1	0	0	0	0
90s	4	3	3	3	2	2	2	2	1	1	0	0
80s											1	0

Score current systolic BP using circled column

MET Call

Adult Deterioration Detection System (ADDS)

	Score 0
	Score 1
	Score 2
	Score 3
4 4 4 4	Score 4
5 5 5 5	Score 5
	MET call

(c)

Date																			
Time																			
Respiratory Rate (breaths / min)	≥ 37																		≥ 37
	36																		36
	31–35																		31–35
	21–30																		21–30
	9–20	20	20	18	18	19	18	18	20	20	20	20	20	16					9–20
	5–8																		5–8
	≤ 4																		≤ 4
O_2 Flow Rate (L / min)	> 5																		> 5
	1–5																		1–5
	< 1	0	0	0	0	0	0	0	0	0	0	0	0	0					< 1
O_2 Saturation (%)	≥ 93	98	98	98	98	98	98	97	98	98	98	98	97	94					≥ 93
	90–92																		90–92
	85–89																		85–89
	≤ 84																		≤ 84
Blood Pressure (mmHg) ↑ ↓	Write ≥ 200																		Write ≥ 200
	190s																		190s
	180s																		180s
	170s																		170s
	160s																		160s
	150s							158	159	150									150s
	140s										140								140s
	130s	130				132	137					130	135	132					130s
	120s		128	126	124														120s
	110s																		110s
	100s							102	108	101									100s
	90s										97								90s
	80s											80							80s
	70s	72				72	70							74	72				70s
	60s		68	68	68														60s
If systolic BP ≥ 200, write value in box	50s																		50s
	40s																		40s
Heart Rate (beats / min)	Write ≥ 140																		Write ≥ 140
	130s																		130s
	120s																		120s
	110s																		110s
	100s																		100s
	90s			91	99							98							90s
	80s		80				80	80											80s
	70s					78			70	70				78					70s
	60s	65									60			63					60s
If heart rate ≥ 140, write value in box	50s																		50s
	40s																		40s
	30s																		30s
Temperature (C)	≥ 38.6																		≥ 38.6
	38–38.5																		38–38.5
	36.1–37.9	36.3	36.3	36.5	36.1	36.1	36.6	36.1	36.1	36.1	36.4	36.8	36.2	36.5					36.1–37.9
	35.1–36																		35.1–36
	34.1–35																		34.1–35
	≤ 34																		≤ 34
4 Hour Urine Output (mL)	≥ 800																		≥ 800
	120–799	500	513	178	364	700	229	683	478	600	566	412	423	420					120–799
	80–119																		80–119
	≤ 79																		≤ 79
Consciousness If necessary, wake patient before scoring	Alert	A	A	A	A	A	A	A	A	A	A	A	A	A					Alert
	Voice																		Voice
	Pain																		Pain
	Unresp.																		Unresp.
Increased pain	✓																		✓
ADDS Scores	Respiratory Rate	0	0	0	0	0	0	0	0	0	0	0	0	0					**ADDS Scores**
	O_2 Flow Rate	0	0	0	0	0	0	0	0	0	0	0	0	0					
	O_2 Saturation	0	0	0	0	0	0	0	0	0	0	0	0	0					
	Systolic BP	0	0	0	0	0	0	0	0	0	0	0	0	0					
	Heart Rate	0	0	0	0	0	0	0	0	0	0	0	0	0					
	Temperature	0	0	0	0	0	0	0	0	0	0	0	0	0					
	4 Hour Urine Output	0	0	0	0	0	0	0	0	0	0	0	0	0					
	Consciousness	0	0	0	0	0	0	0	0	0	0	0	0	0					
	TOTAL ADDS	0	0	0	0	0	0	0	0	0	0	0	0	0					

Usual systolic BP: 128 Signature:

Circle the column showing the patient's usual systolic BP

	190s	180s	170s	160s	150s	140s	130s	120s	110s	100s	90s	80s
Write ≥ 200	0	0	1	1	2	2	2	3	3	4	5	5
190s	0	0	0	1	1	1	2	2	3	3	4	4
180s	0	0	0	0	0	1	1	2	2	3	3	4
170s	1	0	0	0	0	1	1	2	2	3	3	3
160s	1	1	0	0	0	0	0	1	1	2	2	2
150s	1	1	1	0	0	0	0	0	1	1	2	2
140s	2	1	1	1	0	0	0	0	0	1	1	1
130s	2	2	1	1	0	0	0	0	0	0	0	1
120s	2	2	2	1	1	0	0	0	0	0	0	0
110s	3	2	2	2	1	1	0	0	0	0	0	0
100s	3	3	3	2	2	2	1	1	0	0	0	0
90s	4	3	3	3	2	2	2	2	1	1	0	0
80s											1	0

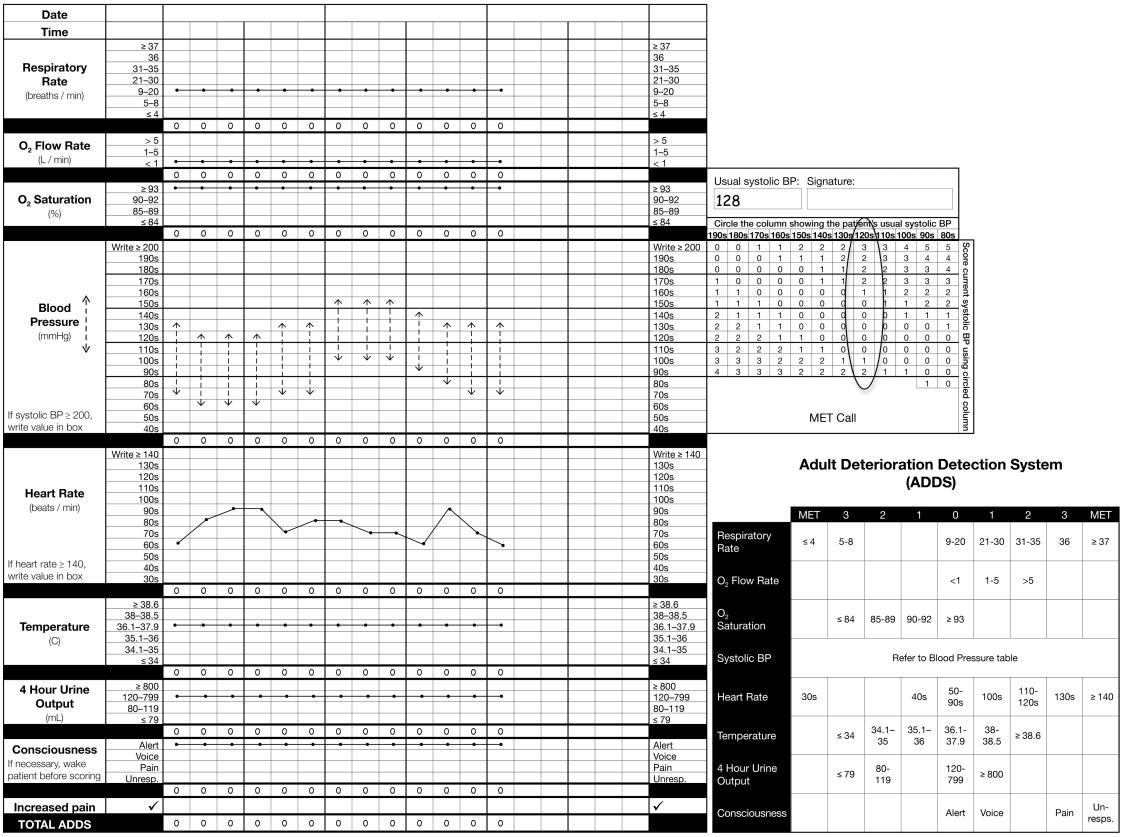
Score current systolic BP using circled column

MET Call

Adult Deterioration Detection System (ADDS)

	MET	3	2	1	0	1	2	3	MET
Respiratory Rate	≤ 4	5-8			9-20	21-30	31-35	36	≥ 37
O_2 Flow Rate					<1	1-5	>5		
O_2 Saturation		≤ 84	85-89	90-92	≥ 93				
Systolic BP	Refer to Blood Pressure table								
Heart Rate	30s			40s	50-90s	100s	110-120s	130s	≥ 140
Temperature		≤ 34	34.1–35	35.1–36	36.1-37.9	38-38.5	≥ 38.6		
4 Hour Urine Output		≤ 79	80-119		120-799	≥ 800			
Consciousness					Alert	Voice		Pain	Un-resps.

(d)



Adult Deterioration Detection System (ADDS)

	MET	3	2	1	0	1	2	3	MET
Respiratory Rate	≤ 4	5-8			9-20	21-30	31-35	36	≥ 37
O_2 Flow Rate					<1	1-5	>5		
O_2 Saturation		≤ 84	85-89	90-92	≥ 93				
Systolic BP	Refer to Blood Pressure table								
Heart Rate	30s			40s	50-90s	100s	110-120s	130s	≥ 140
Temperature		≤ 34	34.1–35	35.1–36	36.1-37.9	38-38.5	≥ 38.6		
4 Hour Urine Output		≤ 79	80-119		120-799	≥ 800			
Consciousness					Alert	Voice		Pain	Un-resps.

Scores

In real-world clinical situations, chart-users interpret observation charts that are in different states of completion. Sometimes, all vital sign data, individual vital sign scores, and early-warning scores to date will already be present before a particular clinician picks up the chart. In other cases, some or all of the scores will be missing, either because compliance with the scoring-system is less than 100% (Odell *et al.* 2009), or because the nurse is in the process of recording the vital signs and has yet to complete the scoring. It is not necessarily the case that the same design options would be beneficial in all circumstances. Therefore, to obtain results generalizable to a broader range of real-world clinical situations, we manipulated whether or not scores were provided to participants.

Prior to testing, participants were assigned to one of two conditions using a random sequence generated by Microsoft Excel 2011: (1) *scores present*, where all charts had real scores recorded on them (n = 102); or (2) *scores absent*, where all charts contained uninformative fillers (the letter 'U') in place of the real scores (n = 103) (Figure 2). These fillers prevented the presence vs. absence of scores from being confounded with the absence vs. presence of blank scoring-rows. To account for this manipulation, the task instructions (see below) informed participants in the scores absent condition that 'U' was an abbreviation for 'unrecorded'.

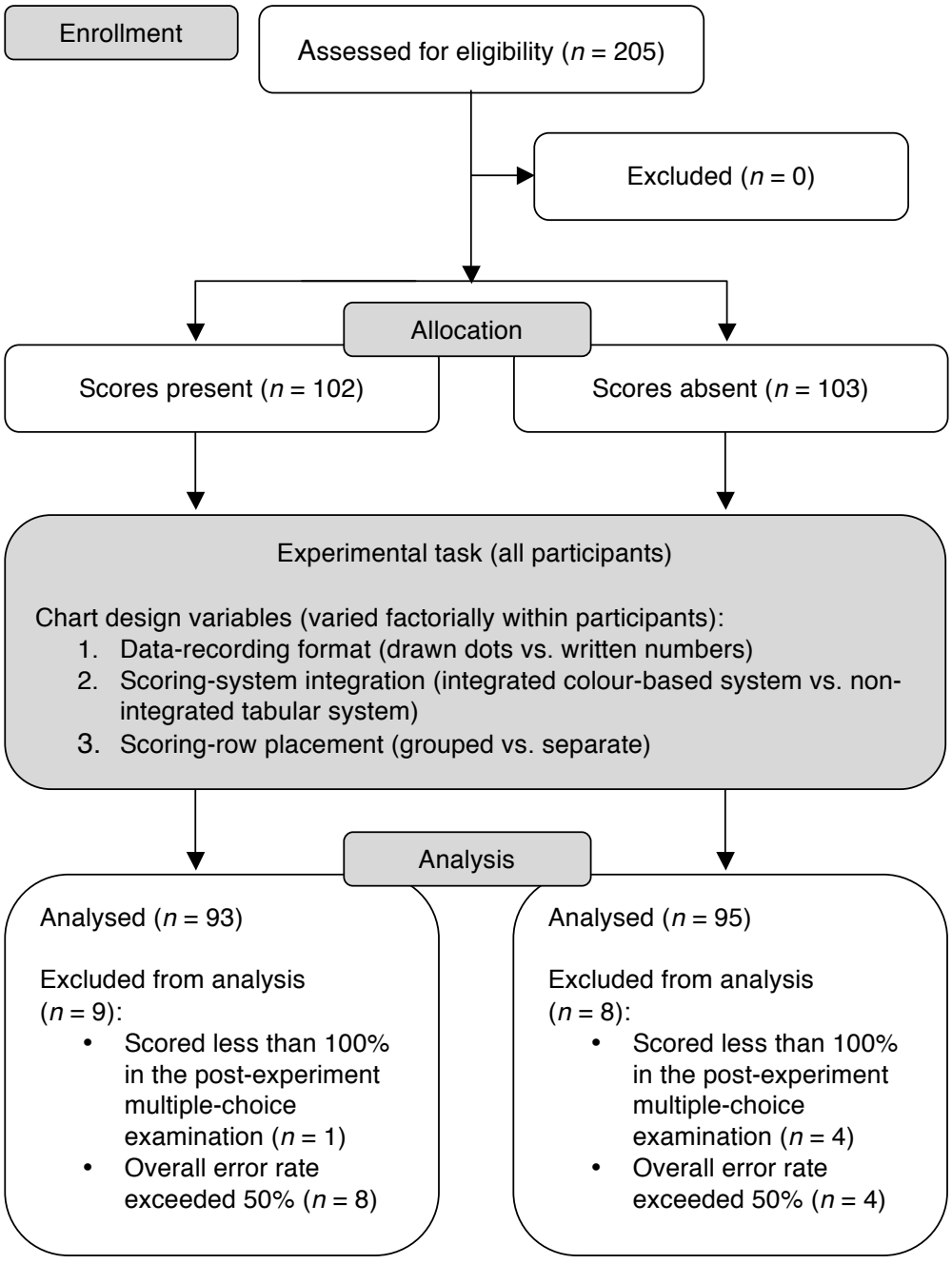
Participants

Given that initial decisions about deteriorating patients are often made by relatively inexperienced nurses and doctors (Endacott *et al.* 2010), the present study focussed on novice performance. Power analysis (G*Power 3.1.9.2: Faul et al., 2007) indicated that a minimum sample of 180 participants was necessary to detect medium-sized effects with 95% power and alpha set at 0.05. A convenience sample of 205 novice chart-users, recruited from a Brisbane university (QLD, Australia), received psychology course credit for participating. Only individuals with no prior hospital chart experience were eligible, to ensure that no particular design option could be advantaged by participants' previous chart-related preferences or experiences. In our prior experiments addressing observation chart design (Preece *et al.* 2012a, Christofidis *et al.* 2013, Christofidis *et al.* 2014), samples of naïve participants (recruited through the psychology research participation scheme) and health professionals consistently yielded very similar patterns of results across charts. Therefore, we reasoned that there would be no additional value in including a group of non-naïve novices, such as medical or nursing students.

After participating in the experiment, participants were excluded if they answered one or more items incorrectly in the post-experiment multiple-choice examination (see *Data collection*) or

their overall error rate exceeded 50% (Figure 2). This was to ensure that, in the final sample, failure to understand the training instructions or retain the key information could not provide an alternative explanation for the results. Nevertheless, the overall patterns of results reported below remained unchanged when statistical analyses were re-run with these participants included.

Fig. 2. Flow diagram illustrating the enrollment, allocation and analysis of participants.



Data collection

Participants were recruited and tested between March 2011 – March 2014. Each participant was trained and tested individually in a quiet room. After completing a demographic questionnaire, participants watched a series of training videos that explained: (a) the ten vital signs included in the chart and their normal ranges; (b) track-and-trigger systems; and (c) how to use each chart design (presented in a different random order for each participant).

Next, the key concepts and vital sign normal ranges were tested with a 10-item multiple-choice examination. Participants scoring below 100% studied a summary and retook the examination until they answered everything correctly. A final video explained the experiment, and indicated that responses and response times would be recorded.

Using a similar methodology to previous studies (Preece *et al.* 2012a, Christofidis *et al.* 2013), participants completed 64 experimental trials where they were presented with a patient chart containing a different case of patient data. For each participant, cases were randomly assigned to charts with the constraint that each design was assigned four normal and four abnormal cases (comprising derangements in oxygen saturation, systolic blood pressure, heart rate, and temperature). To prevent order effects, trials were presented in a different random order for each participant.

In each trial, a chart appeared on a computer monitor, and the participant responded by clicking on a green 'normal' button at the bottom of the screen (to indicate that all observations were normal) or on the appropriate vital sign graphing area (to indicate an abnormality). SuperLab experimental software (Cedrus Corporation, 2007) was used to present the images, and to record the responses and response times (in milliseconds) for each trial. After completing all 64 trials, participants re-sat the multiple-choice examination.

Ethical considerations

This study was granted ethical approval in accordance with the review processes of the university ethics committees.

Data Analysis

For each trial, the response was coded as 'correct' if the participant clicked on the appropriate area of the screen, identifying an abnormal vital sign or classifying a normal case as normal. Each participant's average response time and error rate (i.e. percentage of incorrect

responses) were calculated for each design. Statistical analyses were performed using IBM SPSS 21.0 (IBM Corp., Armonk, NY: USA) with statistical significance set at $\alpha = 0.05$. Separate mixed-design (data-recording format × scoring-system integration × scoring-row placement × scores) analyses of variance (ANOVAs) were conducted on response times and error rates, with η^2 as the measure of effect size (Howell 1997). T-tests were used to follow-up significant interactions, with Cohen's *d* as the effect size measure (Cohen 1992).

RESULTS

Participant characteristics

Table 2 presents participant characteristics for the final sample of 188.

Table 2
Participant characteristics, including *p*-values for comparisons between conditions on age (*t*-test) and gender (chi-squared test). Values are mean (*SD*) or percentage (*n*).

Variable		Experimental condition		*p*-value
		Real early-warning scores (*N* = 93)	Filler early-warning scores (*N* = 95)	
Age in years		20.03 *(5.39)*	19.48 *(4.54)*	0.409
Gender	Female	69.89% (*65*)	63.16% (*60*)	0.328
	Male	30.11% (*28*)	36.84% (*35*)	0.328

Response time

Analysis of the response time data revealed a significant main effect of data-recording format, $F(1, 186) = 82.05$, $p < 0.001$, $\eta^2 = 0.27$, qualified by a significant data-recording format × scores interaction, $F(1, 186) = 38.56$, $p < 0.001$, $\eta^2 = 0.13$ (Figure 3a). Participants for whom scores were absent responded 2.24 seconds faster (CI 1.76-2.72) using drawn-dot (vs. written-number) observations, $t(1,94) = -9.21$, $p < 0.001$, Cohen's $d = -0.55$, and participants with access to scores responded 0.42 seconds faster (CI 0.10-0.74), $t(1, 92) = -2.58$, $p < 0.05$, Cohen's $d = -0.13$.

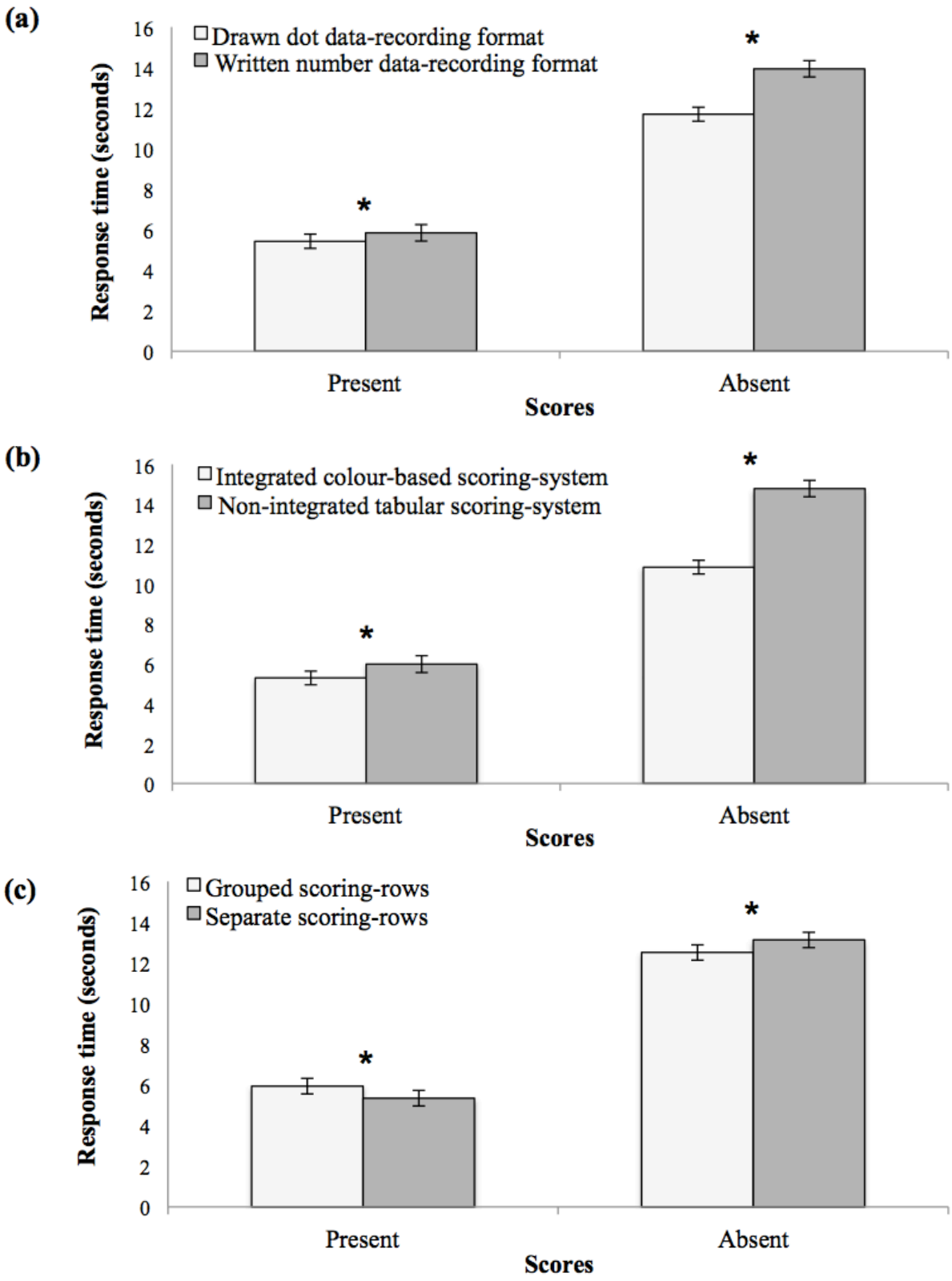
We also found a significant main effect of scoring-system integration, $F(1, 186) = 195.83$, $p < 0.001$, $\eta^2 = 0.41$, qualified by a significant interaction with scores, $F(1, 186) = 96.90$, $p < 0.001$, $\eta^2 = 0.20$ (Figure 3b). Participants for whom scores were absent responded 3.94 seconds faster (CI 3.40-4.48) using an integrated colour-based (vs. tabular) system, $t(1, 94) = -14.52$, $p < 0.001$, Cohen's $d = -0.95$, and participants with access to scores responded 0.69 seconds faster (CI 0.32-1.06), $t(1, 92) = -3.68$, $p < 0.001$, Cohen's $d = -0.22$.

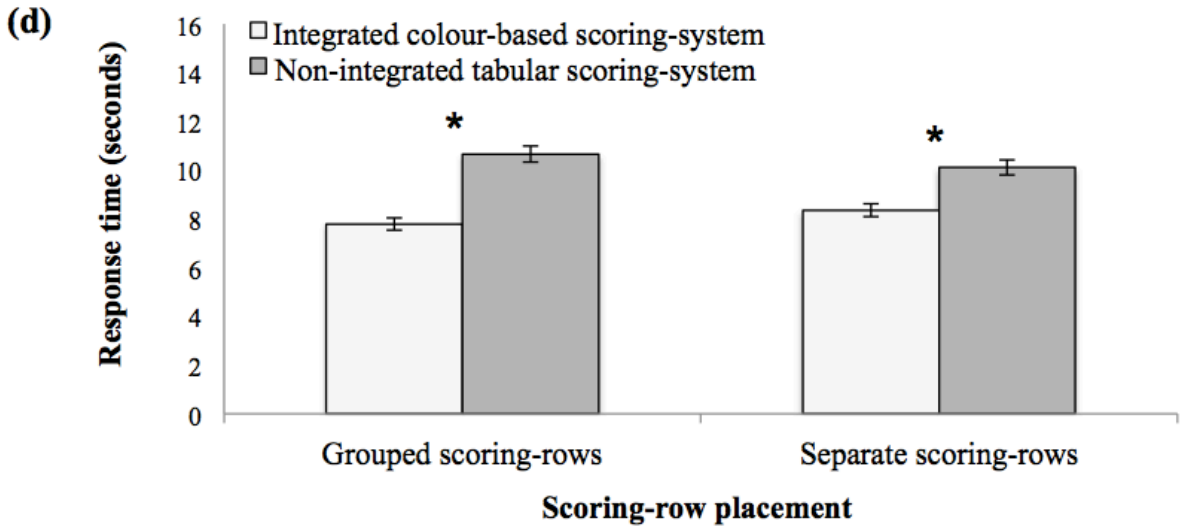
Although there was no significant main effect of scoring-row placement, $F(1, 186) = 0.01$, $p = 0.941$, there was a significant interaction with scores, $F(1, 186) = 13.60$, $p < 0.001$, $\eta^2 = 0.07$ (Figure 3c). Participants for whom scores were absent responded 0.62 seconds faster (CI 0.14-1.09) using grouped (vs. separate) scoring rows, $t(1, 94) = -2.58$, $p < 0.05$, Cohen's $d = -0.15$. However, participants with access to scores responded 0.59 seconds faster (CI 0.15-1.04) using separate (vs. grouped) scoring-rows, $t(1, 92) = 2.64$, $p < 0.05$, Cohen's $d = 0.18$.

Further, there was a significant scoring-system integration × scoring-row placement interaction, $F(1, 186) = 16.82$, $p < 0.001$, $\eta^2 = 0.08$ (Figure 3d). Participants responded 2.89 seconds faster (CI 2.38-3.39) using an integrated colour-based (vs. tabular) system when scoring-rows were grouped $t(1, 187) = -11.23$, $p < 0.001$, Cohen's $d = -0.55$, and 1.78 seconds faster (CI 1.32-2.23) when scoring-rows were separate , $t(1, 187) = -7.70$, $p < 0.001$, Cohen's $d = -0.32$.

Additionally, there was a main effect of scores, indicating that participants for whom scores were present (vs. absent) responded faster overall, $F(1, 186) = 194.80$, $p < 0.001$, $\eta^2 = 0.52$. However, this effect was also qualified by the interactions with data-recording format, scoring-system integration, and scoring-row placement outlined above.

Fig. 3. Response times for detecting abnormal observations, arranged by: (a) data-recording format and scores; (b) scoring-system integration and scores; (c) scoring-row placement and scores; and (d) scoring-system integration and scoring-row placement. Error bars indicate standard errors. Significant differences between adjacent bars are marked with an asterisk.





Error rate

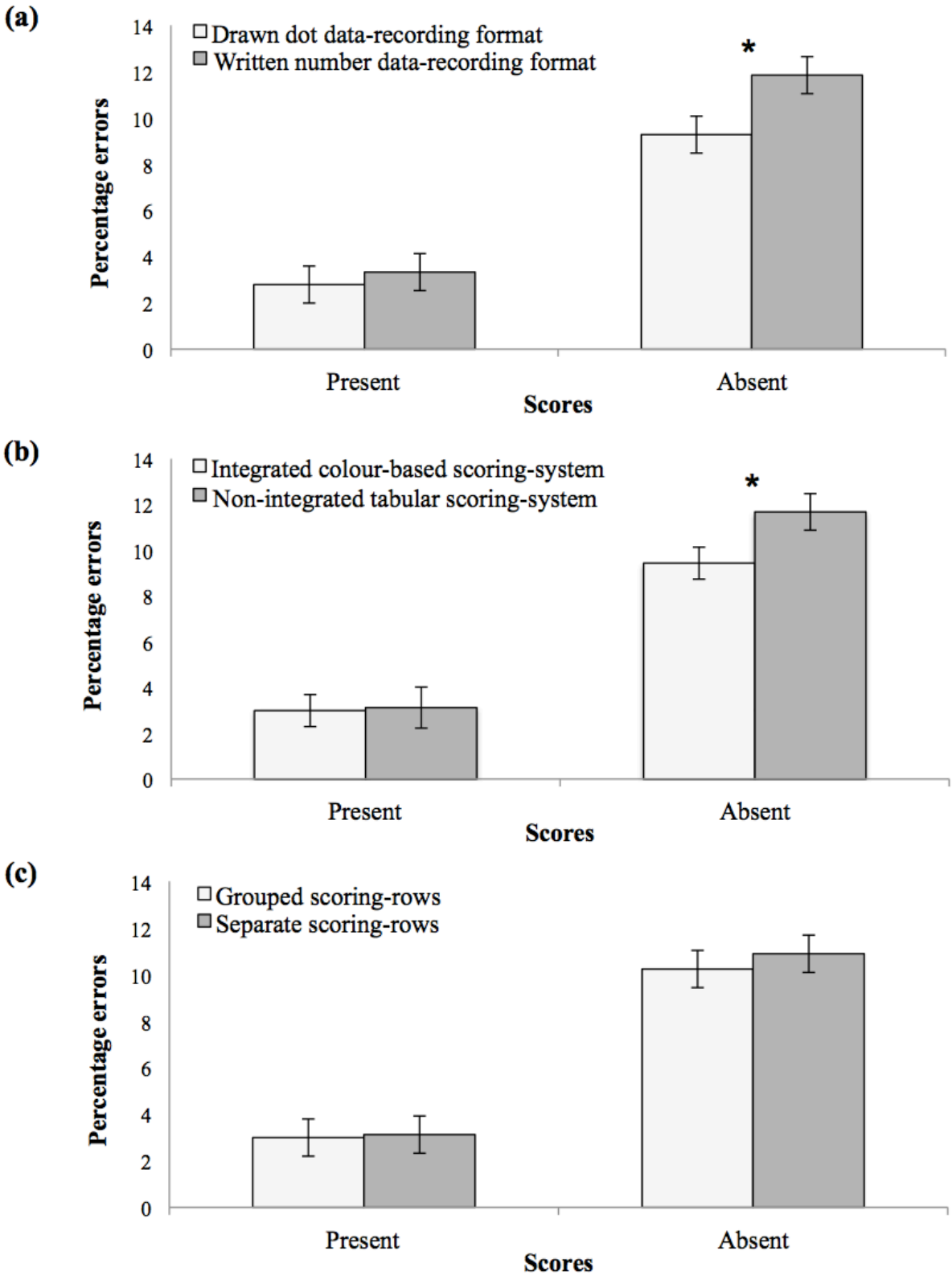
The ANOVA on error rate data revealed a significant main effect of data-recording format, $F(1, 186) = 14.88$, $p < 0.001$, $\eta^2 = 0.07$, again qualified by a significant data-recording format × scores interaction, $F(1, 186) = 6.36$, $p < 0.05$, $\eta^2 = 0.03$ (Figure 4a). Participants for whom scores were absent made 2.57% fewer errors (CI 1.19-3.94) using drawn dots (vs. written numbers), $t(1, 94) = -3.70$ $p = < 0.001$, Cohen's $d = -0.27$. However, for participants with access to scores, there was no effect of data-recording format ($p > 0.05$).

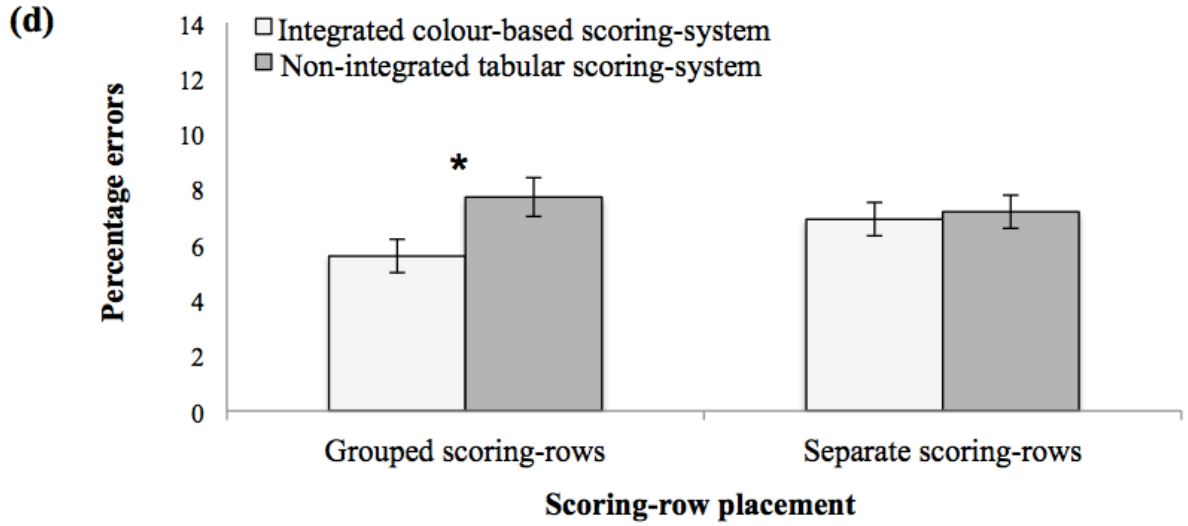
Once again, there was a significant main effect of scoring-system integration, $F(1, 186) = 7.66$, $p < 0.05$, $\eta^2 = 0.04$, qualified by a significant interaction with scores, $F(1, 186) = 6.02$, $p < 0.05$, $\eta^2 = 0.03$ (Figure 4b). Participants for whom scores were absent made 2.24% fewer errors (CI 0.75-3.73) using an integrated colour-based (vs. tabular) system, $t(1, 94) = -2.98$, $p < 0.05$, Cohen's $d = -0.23$. However, this effect was not significant for participants with access to scores ($p > 0.05$).

For error rate, scoring-row placement yielded no significant main effect or interaction with scores (p's > .05; Figure 4c). However, as with response time, there was a significant scoring-system integration × scoring-row placement interaction, $F(1, 186) = 5.29$, $p < 0.05$, $\eta^2 = 0.03$ (Figure 4d). Participants made 2.13% fewer errors (CI 1.01-3.25) using an integrated colour-based (vs. tabular) system when scoring-rows were grouped, $t(1, 187) = -3.75$, $p < 0.001$, Cohen's $d = -0.22$. However, this effect was not significant when scoring-rows were separate ($p > 0.05$).

Again, there was a main effect of scores: participants for whom scores were present (vs. absent) made fewer errors overall, $F(1, 186) = 51.99$, $p < 0.001$, $\eta^2 = 0.22$. However, this effect was also qualified by the interactions with data-recording format and scoring-system integration reported above.

Fig. 4. Error rates for detecting abnormal observations, arranged by: (a) data-recording format and scores; (b) scoring-system integration and scores; (c) scoring-row placement and scores; and (d) scoring-system integration and scoring-row placement. Error bars indicate standard errors. Significant differences between adjacent bars are marked with an asterisk.





DISCUSSION

This is the first study to systematically evaluate several observation chart design features to assess their contributions to the detection of patient deterioration. In support of Hypothesis 1, participants responded significantly faster (whether scores were present or absent on their charts) and made significantly fewer errors (if scores were absent) using drawn-dot (vs. written-number) observations. This suggests that, in a range of real-world clinical contexts, drawn-dot observations may yield faster detection of abnormal vital signs, and they may also prevent errors in some circumstances. In contrast, we found no evidence of any advantage for written-number observations, supporting existing indicative findings (Chatterjee *et al.* 2005; Fung *et al.* 2014). Considering this pattern of results, we argue that paper-based observation charts should utilise drawn-dot observations.

Our findings are consistent with the proposal that drawn-dot vital sign observations eliminated a potential source of unwanted workload, and therefore freed chart-users' cognitive resources for higher-level tasks (Gerhardt-Powals 1996), by preventing the mental processing that numerical observations might have triggered (e.g. automatically reading the numbers and/or comparing them with clinical criteria stored in memory). In so doing, the use of drawn-dots also ensured that the task of searching for abnormal observations was not unnecessarily data-driven, potentially reducing the time that chart-users spent assimilating raw vital sign data (Gerhardt-Powals 1996). This interpretation is even more compelling when one considers that, just like the drawn-dot observations, the written numbers used in the experiment were presented graphically (as 'quasi-graphs'; Preece *et al.* 2009). Hence, participants did not *need* to read the numbers to determine whether or not any particular observation was normal or abnormal, or to observe trends in the data. This contrasts with the many charts in clinical use that present observations as numbers

written in a single row or column for each vital sign (Preece *et al.* 2009), forcing users to mentally visualise them in a graph-like format to interpret trends (Preece *et al*. 2013). Indeed, charts featuring tabulated observations have yielded markedly slower response times and higher error rates (by both experienced clinicians and novice chart-users) in similar experimental studies (Preece *et al.* 2012a, Christofidis *et al.* 2013). Further, written-number observations are arguably even more redundant when one considers that measurement error and transient variability (due to perturbations, natural steady-state variability, or clinicians' technique; Reisner *et al.* 2012) can cause vital signs to fluctuate substantially over time.

Consistent with Hypothesis 2, participants were significantly faster (whether scores were present or absent) and significantly more accurate (absent scores only) when using an integrated colour-based (vs. non-integrated tabular) scoring-system. This finding has practical implications for chart-users' efficiency: regardless of whether scores are recorded or not, an integrated colour-based system should lead to faster recognition of patient deterioration and, in some circumstances, fewer errors. Further, the study yielded no evidence of any circumstance where a non-integrated system would be advantageous. Given these results, we suggest that charts should also utilise colour-based, rather than tabular, scoring-systems. We propose that the presence of an integrated colour-based scoring-system automated chart-users' unwanted workload by reducing the need for mental comparisons and unnecessary thinking (Gerhardt-Powals 1996). That is, participants did not need to consider normal ranges listed in a look-up table or held in memory. Instead, they could use the colour cues embedded in the graphs to identify criterion breaches rapidly; hence, the system also eliminated any need for the detection of abnormal observations to be a time-consuming, data-driven task (Gerhardt-Powals 1996).

The results relating to Hypothesis 3 were more mixed. The effect of scoring-row placement was confined to the response time data, and differed in direction depending on whether scores were present or absent. Participants without access to scores were significantly faster using charts that had scoring-rows grouped together at the bottom of the page rather than separate scoring-rows, as predicted. However, when scores were present, charts with separate scoring-rows outperformed those with grouped rows. These findings should be read in conjunction with the results of a recent experimental study which found that participants were faster at determining and recording early-warning scores when the scoring-rows were separate, rather than grouped (Christofidis *et al.* 2015). If chart-related protocols are adhered to and all observations are scored, then the results of the present study also suggest, contrary to Hypothesis 3, that separate scoring rows may be preferable. Hence, the optimal arrangement of scoring-rows may depend on the clinical context and compliance culture, and we can make no overarching recommendation.

Interestingly, there was a significant interaction between scoring-system integration and

scoring-row placement for *both* response time and error rate. Deconstruction of these interactions indicated that, irrespective of whether the chart design featured grouped or separate scoring rows, participants performed better (either in terms of response time, or both accuracy and response time) when the chart incorporated an integrated colour-based scoring-system. These results suggest that the benefits of integrated colour-based scoring-systems are relatively robust to alternative scoring-row placements.

Unsurprisingly, participants were also faster and more accurate overall when early-warning scores were present (rather than absent), suggesting that they do assist chart-users to recognise deterioration. However, it should be noted that all of the scores recorded on charts in the present study were accurate, which will not always be the case in real clinical contexts (Christofidis *et al.* 2015).

The superior performance of the drawn-dot observations and integrated colour-based scoring-system validates several recommendations, based on cognitive engineering principles (Gerhardt-Powals 1996, Horswill *et al.* 2010), made in a systematic evaluation of Australasian observation charts (Preece *et al.* 2013). These recommendations also guided the design of the ADDS chart (Horswill *et al.* 2010), which has consistently out-performed other Australian observation charts in user-performance experiments similar to the present study (Preece *et al.* 2012a, Christofidis *et al.* 2014).

The results of the present experiment also have implications for the interpretation of previous research comparing the efficacy of observation chart designs. For example, Mitchell et al. (2010) re-designed an observation chart to include several potentially user-friendly features (e.g. quasi-graphs and a colour-coded aggregate weighted scoring-system), and conducted a prospective before-and-after intervention trial where the revised chart out-performed its predecessor on several clinical outcome measures (e.g. fewer unexpected ICU admissions and deaths). However, the contribution of specific design elements cannot be assumed because the study compared charts that varied on multiple dimensions, and implementation of the re-designed chart was accompanied by changes in vital sign monitoring policy and substantial education (Mitchell *et al.* 2010). Indeed, in subsequent empirical studies, Mitchell et al.'s design yielded more errors and slower response times compared to the ADDS chart, among both novice chart-users (Preece *et al.* 2012a) and health professionals (Preece *et al.* 2012a, Christofidis *et al.* 2013), including those trained and experienced in its use (Christofidis *et al.* 2013). The present findings suggest that this may be partially attributable to Mitchell et al.'s use of written-number observations (rather than drawn dots) for most vital signs, while the results of another recent experimental study suggest that plotting blood pressure and heart rate together on the same axes may also have compromised usability (Christofidis *et al.* 2014).

Limitations

As with our previous behavioural experiments (Preece *et al.* 2012a, Christofidis *et al.* 2013, Christofidis *et al.* 2014), we have not directly demonstrated that the results generalize to real-world settings. However, given that participants were not subject to the external pressures and distractions experienced by doctors and nurses in practice, it is plausible that the between-charts differences in response times and error rates would be larger in genuine clinical environments, where the impact of poor design on cognitive load would be more crucial (Preece *et al.* 2012a).

To maximise experimental control, we only recruited naïve participants. Consequently, we cannot, strictly speaking, generalize our results to experienced chart-users. However, these findings will still almost certainly apply to health professionals because: (a) samples of novices, nurses and doctors have consistently produced similar patterns of results across charts in our previous experimental studies (Horswill *et al.* 2010, Preece *et al.* 2012, Christofidis *et al.* 2014); and (b) the effects of improved chart design on the detection of abnormal observations have been shown to outweigh health professionals' prior chart experience (Christofidis *et al.* 2013). Given that initial decisions about deteriorating patients are often made by inexperienced doctors and nurses (Endacott *et al.* 2010), the inclusion of novices was important from a pragmatic perspective: observation charts must provide effective support for health professionals of all levels, especially the least experienced. Furthermore, in the future, all clinicians will effectively be novices in relation to paper-based charts once they are used exclusively as the back-up for electronic systems (Christofidis *et al.* 2014).

CONCLUSION

Our findings suggest that chart design features have a substantial impact on chart-users' ability to recognise patient deterioration. More importantly, they further illustrate the need to objectively evaluate the efficacy of observation chart designs. In sum, we suggest that, rather than relying on chart designers' subjective judgements, or clinical trials with limited experimental control, new designs should also be evaluated objectively, through behavioural experimentation or alternative techniques that yield unbiased evidence (Preece *et al.* 2012a, Preece *et al.* 2012b, Christofidis *et al.* 2013, Christofidis *et al.* 2014). Subsequent clinical studies can then focus on broader issues, such as chart utility post-implementation (e.g., Chatterjee *et al.* 2005; Mitchell *et al.* 2010; Elliott *et al.* 2011; Bunkenborg *et al.* 2014; Elliott *et al.* 2014; Kyriacos *et al.* 2015) and subjective user experiences (e.g., Elliott *et al.* 2015). Like manufacturers of medical devices (TGA 2011), chart designers should also be required to provide objective data to support their claims.

References

ACT Health [Internet] (2011) *Compass. Early recognition of deteriorating patient*. Available from: http://health.act.gov.au/professionals/general-information/compass/.

Andre AD & Wickens CD (1995) When users want what's not best for them. *Ergonomics in Design* October, 10-13.

Australian Commission on Safety and Quality in Health Care [Internet] (March 2009) *Recognising and responding to clinical deterioration: Use of observation charts to identify clinical deterioration.* Available from: http://www.safetyandquality.gov.au/wp-content/uploads/2012/02/UsingObservationCharts-2009.pdf.

Bunkenborg G, Samuelson K, Poulsen I, Ladelund S & Akeson J (2014) Lower incidence of unexpected in-hospital death after interprofessional implementation of a bedside track-and-trigger system. *Resuscitation* 85, 424-430.

Clinical Excellence Commission [Internet] (2012) *Between the Flags. Keeping patients safe*. Available from: http://www.cec.health.nsw.gov.au/programs/between-the-flags.

Chatterjee MT, Moon JC, Murphy R & McCrea D (2005) The "OBS" chart: An evidence based approach to re-design of the patient observation chart in a district general hospital setting. *Postgraduate Medical Journal* 81, 663-666. doi: 10.1136/pgmj.2004.031872

Christofidis MJ, Hill A, Horswill MS & Watson MO (2013) A human factors approach to observation chart design can trump health professionals' prior chart experience. *Resuscitation* 84, 657-665. doi: 10.1016/j.resuscitation.2012.09.023

Christofidis MJ, Hill A, Horswill MS & Watson MO (2014). Observation charts with overlapping blood pressure and heart rate graphs do not yield the performance advantage that health professionals assume: an experimental study. *Journal of Advanced Nursing*, 70, 610-624. doi: 10.1111/jan.12223.

Christofidis MJ, Hill A, Horswill MS & Watson MO (2015). Less is more: The design of early-warning scoring systems affects the speed and accuracy of scoring. *Journal of Advanced Nursing*, 71, 1573-1586. doi: 10.111/jan.12618

Cohen J (1992) A power primer. *Psychological Bulletin* 112, 155-9. doi: 10.1037/0033-2909.112.1.155

Elliott D, McKinley S, Perry L, Duffield C, Iedema R, Gallagher R, Fry M, Roche M & Allen E (2011). Observation and Response Charts Usability Testing Report. University of Technology, Sydney.

Elliott D, McKinley S, Perry L, Duffield C, Iedema R, Gallagher R, Fry M, Roche & Allen E (2014). Clinical utility of an observation and response chart with human factors design characteristics and a track and trigger system: study protocol for a two-phase multisite multiple-methods design. *Journal of Medical Internet Research: Research Protocols* 3, e40. doi:10.2196/resprot.3300

Elliott D, Allen E, Perry L, Fry M, Duffield C, Gallagher R, Iedema R, McKinley S & Roche M (2015). Clinical user experiences of Observation and Response Charts: Focus group findings after using a form with human factors design characteristics and a track and trigger system. *BMJ Quality*

and Safety 24, 65-75. doi: 10.1136/bmjqs-2013-002777

Endacott R, Kidd T, Chaboyer W & Edington J (2007) Recognition and communication of patient deterioration in a regional hospital: a multi-methods study. *Australian Critical Care* 20, 100–105. doi: 10.1016/j.aucc.2007.05.002

Endacott R, Scholes J, Buykx P, Cooper P, Kinsman L & McConnell-Henry T (2010) Final-year nursing students' ability to assess, detect and act on clinical cues of deterioration in a simulated environment. *Journal of Advanced Nursing* 66(12), 2722-2731. doi: 10.1111/j.1365-2648.2010.05417

Faul F, Erdfelder E, Lang A-G, & Buchner A (2007) G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191.

Franklin C & Mathew J (1994) Developing strategies to prevent in hospital cardiac arrest: analyzing responses of physicians and nurses in the hours before the event. *Critical Care Medicine* 22, 189–191.

Fung K, Khan F & Dawson J (2014) The introduction of an integrated early warning score observation chart – a picture paints a thousand words. *Journal of Patient Safety* 10, 13-19. doi: 10.1097/PTS.0b013e3182948a39

Gerhardt-Powals J (1996) Cognitive engineering principles for enhancing human-computer performance. *International Journal of Human-Computer Interaction* 8,189-211.

Goldhill DR, White SA & Sumner A (1999) Physiological values and procedures in the 24 h before ICU admission from the ward. *Anaesthesia* 54, 529–534. doi: 10.1046/j.1365-2044.1999.00837

Hillman KM, Bristow PJ, Chey T, Daffurn K, Jacques T, Norman SL, Bishop GF & Simmons G (2001) Antecedents to hospital deaths. *Internal Medicine Journal* 31, 343–348. doi: 10.1046/j.1445-5994.2001.00077

Horswill MS, Preece MHW, Hill A, Christofidis MJ, Karamatic RM Hewett DJ & Watson MO (2010) *Human factors research regarding observation charts: Research project overview.* Sydney, New South Wales, Australia, Australian Commission on Safety and Quality in Health Care.

Howell DC (1997) *Statistical methods for psychology.* Duxbury Press, Belmont, California,.

Kyriacos U, Jelsma J, James M & Jordan S (2015) Early warning scoring systems versus standard observations charts for wards in South Africa: a cluster randomized controlled trial. *Trials* 16. doi: 10.1186/s13063-015-0624-2

Lawson L & Peate I (2009) *Essential Nursing Care*: *A Workbook for Clinical Practice*, Wiley-Blackwell, Chichester, West Sussex.

McDonnell A, Tod A, Bray K, Bainbridge D, Adsetts D & Walters S (2013) A before and after study assessing the impact of a new model for recognizing and responding to early signs of deterioration in an acute hospital. *Journal of Advanced Nursing* 69(1), 41-52. doi: 10.1111/j.1365-2648.2012.05986

Mitchell IA, McKay H, Van Leuvan C, Berry R, McCutcheon C, Avard B, Slater N, Neeman T &

Lamberth P (2010) A prospective controlled trial of the effect of a multi-faceted intervention on early recognition and intervention in deteriorating hospital patients. *Resuscitation* 81, 658-666. doi: 10.1016/j.resuscitation.2010.03.001

Odell M, Victor C & Oliver D (2009) Nurses' role in detecting deterioration in ward patients: systematic literature review. *Journal of Advanced Nursing* 65(10), 1992-2006. doi: 10.1111/j.1365-2648.2009.05109

Preece MHW, Hill A, Horswill MS, Karamatic R & Watson MO (2012b) Designing observation charts to optimise the detection of patient deterioration: Reliance on the subject preferences of healthcare professionals in not enough. *Australian Critical Care* 25, 238-252. doi:10.1016/j.aucc.2012.01.003

Preece MHW, Hill A, Horswill MS, Karamatic R, Hewett DG & Watson MO (2013) Applying heuristic evaluation to observation chart design to improve the detection of patient deterioration. *Applied Ergonomics* 44, 544-556. doi: 10.1016/j.apergo.2012.11.003

Preece MHW, Hill A, Horswill MS & Watson MO (2012a) Supporting the detection of patient deterioration: Observation chart design affects the recognition of abnormal vital signs. *Resuscitation* 83, 1111-1118. doi: 10.1016/j.resuscitation.2012.02.009

Preece MHW, Horswill MS, Hill A, Karamatic R, Hewett D & Watson MO (2009) *Heuristic analysis of 25 Australian and New Zealand adult general observation charts.* Sydney, New South Wales, Australia, Australian Commission on Safety and Quality in Health Care.

Preece MHW, Horswill MS, Hill A, Karamatic R & Watson MO (2010) *An online survey of health professionals' opinions regarding observation charts.* Sydney, New South Wales, Australia, Australian Commission on Safety and Quality in Health Care.

Prytherch DR, Smith GB, Schmidt P, Featherstone PI, Stewart K, Knight D & Higgins B (2005) Calculating early-warning scores - a classroom comparison of pen and paper and hand-held computer methods. *Resuscitation* 70, 173-178. doi: 10.1016/j.resuscitation.2005.12.002

Reisner AT, Chen L & Reifman J (2012) The association between vital signs and major hemorrhagic injury is significantly improved after controlling for sources of measurement variability. *Journal of Critical Care* 27*,* 533.e1-533.e10. doi: 10.1016/j.jcrc.2012.01.006

Therapeutic Goods Administration [Internet] (2011) *Australian regulatory guidelines for medical devices (ARGMD)*. Available from: http://www.tga.gov.au/pdf/devices-argmd.pdf

Chapter 5

Christofidis, M.J., Hill, A., Horswill, M.S., & Watson, M.O (2015). Less is more: the design of early-warning scoring systems affects the speed and accuracy of scoring. *Journal of Advanced Nursing, 71*(7), 1573-1586.

Table 4. Manuscript revision history for "Less is more: the design of early-warning scoring systems affects the speed and accuracy of scoring"

Date	Detail
20 May 2014	Submitted to the *Journal of Advanced Nursing*
27 August 2014	Article revised
9 December 2014	Article accepted for publication
14 June 2015	Published in print

Hypotheses

Chart-users will determine total early-warning scores more accurately when individual vital sign scoring-rows are grouped. Chart-users will make more errors in the absence of scoring-rows, despite determining early-warning scores faster.

INTRODUCTION

Many paper-based observation charts used in hospitals incorporate physiological 'track-and-trigger' systems to aid nurses and doctors in the early detection of patient deterioration (Prytherch *et al.* 2005, Subbe *et al.* 2007, Mohammed *et al.* 2009). These systems fall into three broad categories: (a) single- and multiple-parameter systems (where vital sign observations are compared with a set of criteria to determine whether one or more parameters have reached predefined thresholds); (b) aggregate weighted scoring systems (which allocate a weight or 'individual vital sign score' to each observation as a function of its level of derangement from a predetermined normal range); and (c) combination systems (which combine an aggregate weighted scoring system with a single- or multiple-parameter system) (Prytherch *et al.* 2005, Gao *et al.* 2007, Subbe *et al.* 2007, Smith *et al.* 2008, ACSQHC 2009). In the latter two system-types, individual vital sign scores are summed to provide a single score (sometimes called an 'early-warning score') that summarizes the patient's overall physiological condition (Prytherch *et al.* 2005, Mohammed *et al.* 2009). As well as assisting health professionals to recognize deterioration, these scores can be used by nurses to trigger appropriate actions, from increasing the frequency of observations through to calling for emergency assistance, depending on the magnitude of the score (Prytherch *et al.* 2005, Lawson & Peate 2009, Mohammed *et al.* 2009).

Indeed, early-warning scores have been shown to be an effective decision-making tool to help nurses assess at-risk patients (Andrews & Waterman 2005). They also empower nurses by providing objective evidence of patient deterioration and a concise and unambiguous means of communicating it to doctors (Andrews & Waterman 2005). However, these advantages are dependent on accurate scoring.

Background

Despite the relatively widespread adoption of early-warning scoring systems, the accuracy with which chart-users can determine patients' early-warning scores has received only minimal research attention (Prytherch *et al.* 2005, Smith *et al.* 2008, Mohammed *et al.* 2009). Past studies have established that errors occur frequently, both via simulations (Prytherch *et al.* 2005, Mohammed *et al.* 2009) and retrospective case-note analysis (Smith *et al.* 2008). However, further research is required to better understand their causes and potential remedies, given that every step in the process of determining a patient's early-warning score is susceptible to human error (Prytherch *et al.* 2005, Smith *et al.* 2008, Mohammed *et al.* 2009). These steps typically include: (a) collecting and recording raw vital sign data (where measurement and transcription errors may occur); (b)

scoring each observation (which may lead to 'scoring errors'); and (c) for each set of observations, summing the individual vital sign scores (where 'adding errors' may occur). Any of these errors can influence the overall score and, consequently, the appropriateness of the clinical response (Prytherch *et al.* 2005). For instance, under-scoring may delay the detection of deterioration, increasing the risk of an adverse outcome for the patient; and over-scoring may cause medical staff to be called unnecessarily, placing additional strain on finite hospital resources (Prytherch *et al.* 2005).

It has been suggested that these errors may be reduced by using a computer-based system that automates parts of the process (Prytherch *et al.* 2005, Mohammed *et al.* 2009). Nevertheless, there remains a compelling need for research on paper-based systems. Not only are they still globally ubiquitous (Preece *et al.* 2012a) but their use is likely to continue for many years to come, especially in developing countries; and they will have an even longer life as the back-up for electronic systems (Christofidis *et al.* 2014).

Several recent empirical studies have shown that improvements to observation chart design can assist both experienced and novice chart-users to detect abnormal observations more quickly and accurately (Christofidis *et al.* 2012, Preece *et al.* 2012b, Christofidis *et al.* 2013, Christofidis *et al.* 2014). However, no published study has assessed the impact of chart design on the determination of early-warning scores.

THE STUDY

Aims

The present study aimed to examine the effect of scoring-system design on the determination of early-warning scores, by systematically evaluating three alternative layouts for a colour-based early-warning scoring system. The layouts mirrored those of three general observation charts widely used in Australia, where there is unresolved debate as to which design solution is best (Horswill *et al.* 2010, Mitchell *et al.* 2010, Queensland Health 2012). The charts used in the experiment varied only in relation to the arrangement of the rows for recording individual vital sign scores. These scoring-rows were either: (a) grouped together beneath all of the vital sign data ('grouped rows'); (b) separated, with each row presented immediately below the corresponding vital sign data ('separate rows'); or (c) excluded altogether ('no rows'). All three chart designs included a row for recording overall early-warning scores at the bottom of the page.

We predicted that grouped rows (Figure 1A) would facilitate the most accurate determination of overall early-warning scores. This was the solution that we chose for the original

Adult Deterioration Detection System (ADDS) Chart (Horswill *et al.* 2010, ACSQHC 2013), which was designed by an interdisciplinary team of human factors specialists and clinicians. The chart was developed as part of a national project for the Australian Commission on Safety and Quality in Health Care and was designed with the specific aim of improving the recognition of patient deterioration. Using human factors principles (Horswill *et al.* 2010, ACSQHC 2013), we reasoned that the close proximity of the *grouped rows* to one another would allow users to sum scores without having to switch their attention (Rashid *et al.* 2012) to another part of the chart, reducing the likelihood of adding errors.

Prior to the ADDS chart, a team of experienced health professionals developed a territory-wide observation chart featuring separate rows (see Figure 1B for an illustration of this strategy) (Mitchell et al. 2010). Despite clinical improvements post-implementation (e.g., fewer unplanned ICU admissions) (Mitchell et al. 2010), we predicted that separate rows would yield more adding errors than grouped rows. To determine the overall score, separate scoring-rows require users to visually align the column of individual scores down the entire page. The interference from data recorded between the scores may cause users to accidentally skip a score or read from the wrong column.

Despite these two (albeit competing) design recommendations, an Australian state health department recently released an alternative ADDS chart design (Queensland Health 2012) that excludes individual vital sign scoring-rows altogether. Although this *no rows* strategy (see Figure 1C for an illustration) may lead to efficiency gains – by eliminating the need to record an additional 144 scores per chart (Horswill *et al.* 2010, ACSQHC 2013) – we predicted that the concurrent tasks of determining the individual vital sign scores and holding a running total in mind would induce greater cognitive load and, as a result, yield additional errors.

Design

The study used a within-subjects experimental design, with *scoring-system design* (grouped rows vs. separate rows vs. no rows) as the independent variable and participants' response times and error rates as the main outcome measures.

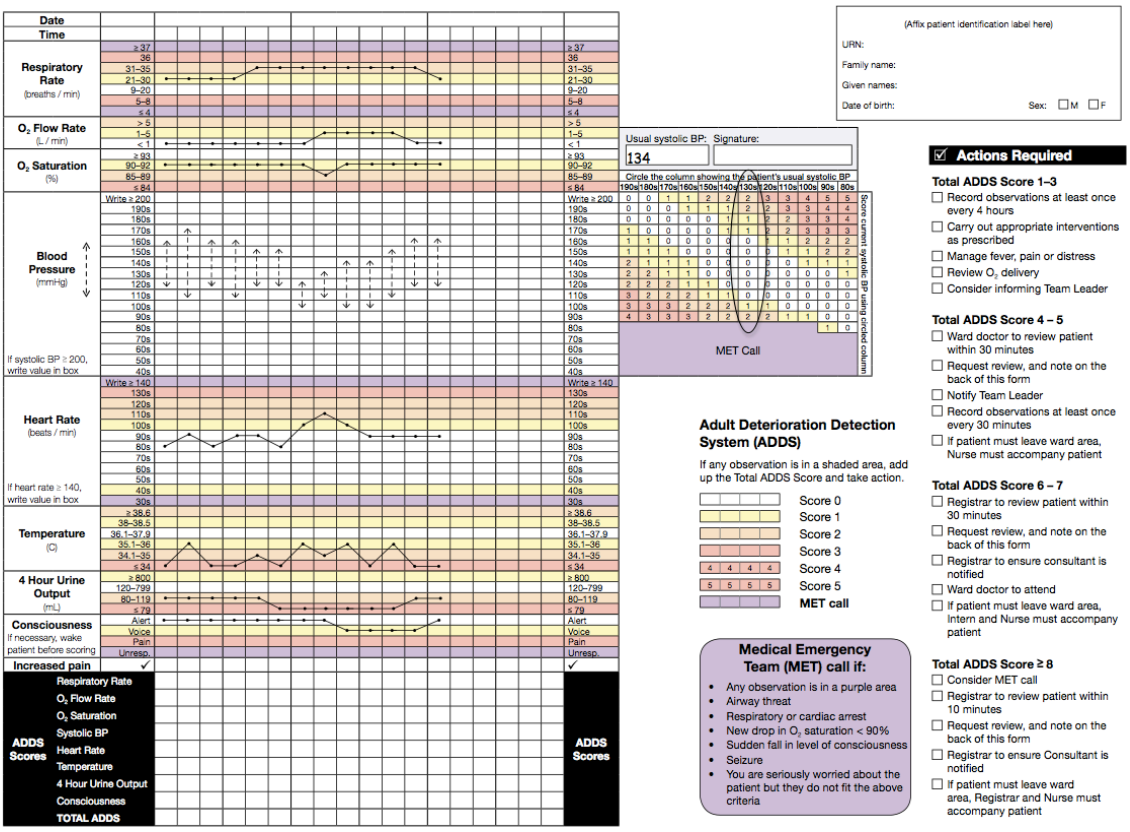
Observation chart designs

The three observation chart designs used in this study were based on the ADDS chart (Horswill *et al.* 2010, ACSQHC 2013). The ADDS was regarded as the most appropriate starting point for this study because of its superior outcomes in previous carefully-controlled human-

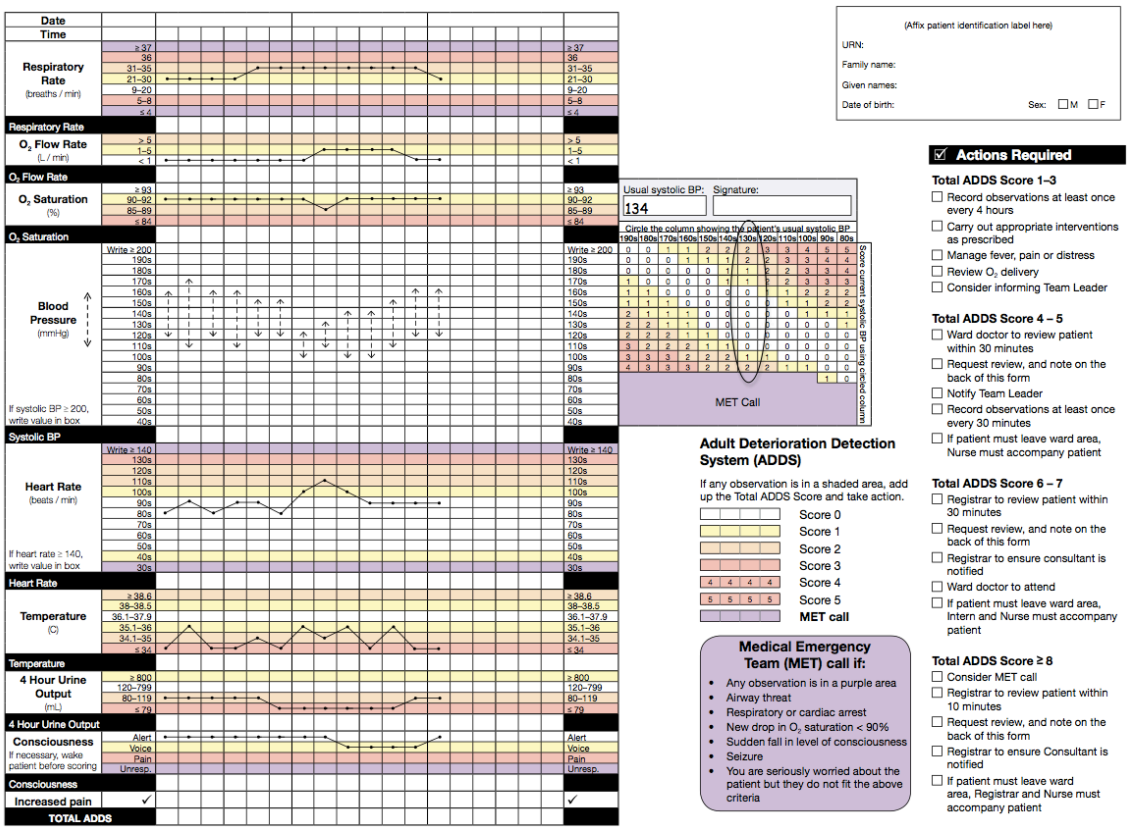
performance experiments. In these studies, participants were faster and more accurate at detecting deranged vital signs on ADDS charts, compared with other widely-used chart designs (Preece *et al.* 2012b, Christofidis *et al.* 2013). However, for the present experiment, the placement of individual vital sign scoring-rows was modified in two versions of the chart to mirror alternative designs used in Australian hopitals (as discussed above). Hence, the three charts used in the study had either: (a) grouped rows (as per the original ADDS design) (Horswill *et al.* 2010, ACSQHC 2013); (b) separate rows (as per Mitchell *et al.* 2010); or (c) *no* rows (as per Queensland Health, 2012) (Figure 1). Adobe InDesign CS5.5 (Adobe Systems Incorporated, 2011) was used to create the three designs and to plot each set of patient data (see below) on to each design. The finished charts were then colour-printed.

Fig. 1. Examples of the three chart designs used in the study, which varied according to their placement of individual vital sign scoring-rows: *grouped rows* (A); *separate rows* (B); and *no rows* (C).

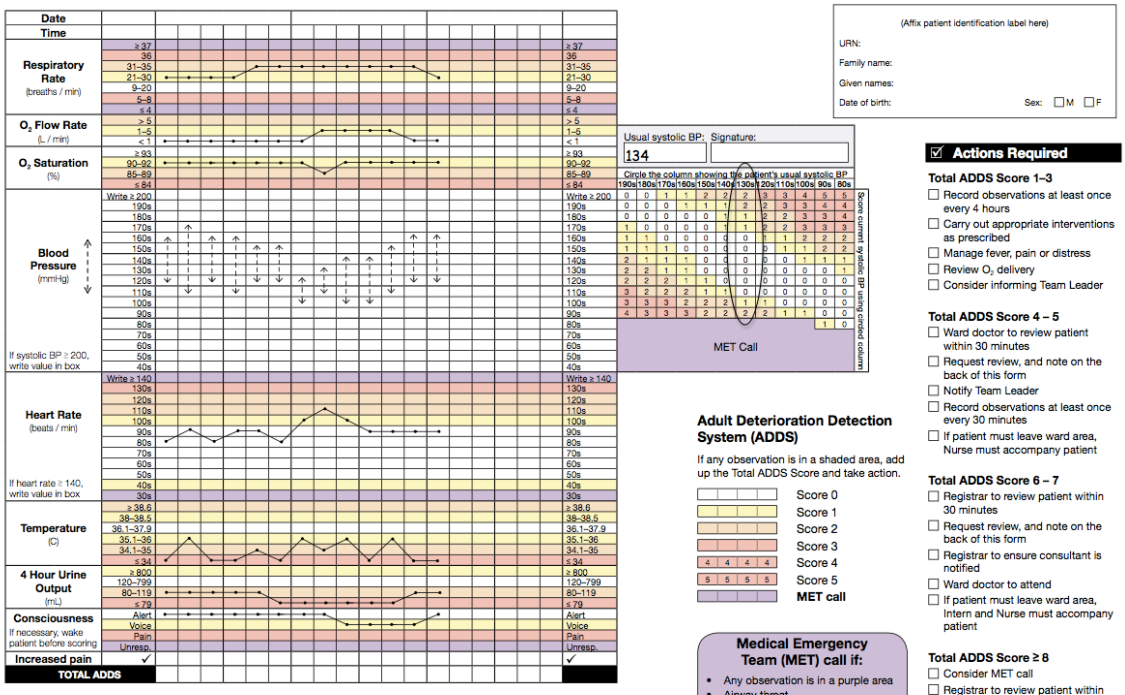
A



B


Date
Time
Respiratory Rate (breaths / min)
Respiratory Rate
O2 Flow Rate (L / min)
O2 Flow Rate
O2 Saturation (%)
O2 Saturation
Blood Pressure (mmHg)
If systolic BP ≥ 200, write value in box
Systolic BP
Heart Rate (beats / min)
If heart rate ≥ 140, write value in box
Heart Rate
Temperature (C)
Temperature
4 Hour Urine Output (mL)
4 Hour Urine Output
Consciousness
If necessary, wake patient before scoring
Alert
Voice
Pain
Unresp.
Consciousness
Increased pain
TOTAL ADDS
(Affix patient identification label here)
URN:
Family name:
Given names:
Date of birth:
Sex: M F
Usual systolic BP:
134
Signature:
Circle the column showing the patient's usual systolic BP
Score current systolic BP using circled column
MET Call
Adult Deterioration Detection System (ADDS)
If any observation is in a shaded area, add up the Total ADDS Score and take action.
Score 0
Score 1
Score 2
Score 3
Score 4
Score 5
MET call
Medical Emergency Team (MET) call if:
Any observation is in a purple area
Airway threat
Respiratory or cardiac arrest
New drop in O2 saturation < 90%
Sudden fall in level of consciousness
Seizure
You are seriously worried about the patient but they do not fit the above criteria
Actions Required
Total ADDS Score 1–3
Record observations at least once every 4 hours
Carry out appropriate interventions as prescribed
Manage fever, pain or distress
Review O2 delivery
Consider informing Team Leader
Total ADDS Score 4 – 5
Ward doctor to review patient within 30 minutes
Request review, and note on the back of this form
Notify Team Leader
Record observations at least once every 30 minutes
If patient must leave ward area, Nurse must accompany patient
Total ADDS Score 6 – 7
Registrar to review patient within 30 minutes
Request review, and note on the back of this form
Registrar to ensure consultant is notified
Ward doctor to attend
If patient must leave ward area, Intern and Nurse must accompany patient
Total ADDS Score ≥ 8
Consider MET call
Registrar to review patient within 10 minutes
Request review, and note on the back of this form
Registrar to ensure Consultant is notified
If patient must leave ward area, Registrar and Nurse must accompany patient

C


Date
Time
Respiratory Rate (breaths / min)
O2 Flow Rate (L / min)
O2 Saturation (%)
Blood Pressure (mmHg)
If systolic BP ≥ 200, write value in box
Heart Rate (beats / min)
If heart rate ≥ 140, write value in box
Temperature (C)
4 Hour Urine Output (mL)
Consciousness
If necessary, wake patient before scoring
Alert
Voice
Pain
Unresp.
Increased pain
TOTAL ADDS
(Affix patient identification label here)
URN:
Family name:
Given names:
Date of birth:
Sex: M F
Usual systolic BP:
134
Signature:
Circle the column showing the patient's usual systolic BP
Score current systolic BP using circled column
MET Call
Adult Deterioration Detection System (ADDS)
If any observation is in a shaded area, add up the Total ADDS Score and take action.
Score 0
Score 1
Score 2
Score 3
Score 4
Score 5
MET call
Medical Emergency Team (MET) call if:
Any observation is in a purple area
Airway threat
Respiratory or cardiac arrest
New drop in O2 saturation < 90%
Sudden fall in level of consciousness
Seizure
You are seriously worried about the patient but they do not fit the above criteria
Actions Required
Total ADDS Score 1–3
Record observations at least once every 4 hours
Carry out appropriate interventions as prescribed
Manage fever, pain or distress
Review O2 delivery
Consider informing Team Leader
Total ADDS Score 4 – 5
Ward doctor to review patient within 30 minutes
Request review, and note on the back of this form
Notify Team Leader
Record observations at least once every 30 minutes
If patient must leave ward area, Nurse must accompany patient
Total ADDS Score 6 – 7
Registrar to review patient within 30 minutes
Request review, and note on the back of this form
Registrar to ensure consultant is notified
Ward doctor to attend
If patient must leave ward area, Intern and Nurse must accompany patient
Total ADDS Score ≥ 8
Consider MET call
Registrar to review patient within 10 minutes
Request review, and note on the back of this form
Registrar to ensure Consultant is notified
If patient must leave ward area, Registrar and Nurse must accompany patient

Patient data

The study used nine different cases of patient data, each spanning 18 consecutive time-points, which included observations for ten vital signs: respiratory rate, oxygen delivery, oxygen saturation, systolic and diastolic blood pressure, heart rate, temperature, four hour urine output, consciousness and pain. Each case contained two sets of observations that would yield each overall early-warning score from *0* to *8* if scored and added correctly (i.e., across the nine cases, each of these 'target' scores occurred 18 times). This range was chosen to maximize content validity by reflecting the clinically-relevant values prescribed by the ADDS chart (Horswill *et al.* 2010, ACSQHC 2013). Across cases, every possible combination of individual vital sign scores that would yield each 'target' overall score was included at least once (Table 1).

To meet these criteria while maximising representativeness, each case was carefully selected from a large pool of genuine de-identified patient data collected from several Australian hospitals. The cases were only modified if a data-point was missing (where a plausible value was extrapolated or interpolated), or if the sets of observations did not meet the strict constraints of the experimental design (where some systolic blood pressure and/or oxygen delivery observations were adjusted slightly to alter their scoring range-rows). In addition, cases where one or more observations fell within a purple range-row were excluded because such observations trigger an immediate Medical Emergency Team (MET) call on ADDS charts (Figure 1), eliminating the need to determine the overall score (Horswill *et al.* 2010, ACSQHC 2013).

Table 1
Combinations of non-zero individual vital sign scores that can sum to each 'target' overall early-warning score from *0* to *8*. Across the 162 overall early-warning scores that participants were required to determine (9 cases × 18 time-points), each of these 'target' scores occurred 18 times and each possible combination of individual vital sign scores listed below was used at least once. These scores were based on the Adult Deterioration Detection System, where 8 individual vital signs are scored (Horswill et al. 2010, ACSQHC 2013).

'Target' overall early-warning score	***0***	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***
1 non-zero digit	-	*1*	*2*	*3*	*4*	*5*	†	†	†
2 non-zero digits	-	-	*1 1*	*2 1*	*3 1* *2 2*	*4 1* *3 2*	*5 1* *4 2*	*5 2* *4 3*	*5 3**
3 non-zero digits	-	-	-	*1 1 1*	*2 1 1*	*3 1 1* *2 2 1*	*3 2 1* *2 2 2*	*5 1 1* *4 2 1* *3 2 2*	*5 2 1* *4 3 1* *4 2 2* *3 3 2*
4 non-zero digits	-	-	-	-	*1 1 1 1*	*2 1 1 1*	*3 1 1 1* *2 2 1 1*	*3 2 1 1* *2 2 2 1*	*5 1 1 1* *4 2 1 1* *3 3 1 1* *3 2 2 1* *2 2 2 2*
5 non-zero digits	-	-	-	-	-	*1 1 1 1 1*	*2 1 1 1 1*	*2 2 1 1 1*	*4 1 1 1 1* *3 2 1 1 1* *2 2 2 1 1*
6 non-zero digits	-	-	-	-	-	-	*1 1 1 1 1 1*	*2 1 1 1 1 1*	*3 1 1 1 1 1* *2 2 1 1 1 1*
7 non-zero digits	-	-	-	-	-	-	-	*1 1 1 1 1 1 1*	*2 1 1 1 1 1 1*
8 non-zero digits	-	-	-	-	-	-	-	-	*1 1 1 1 1 1 1 1*

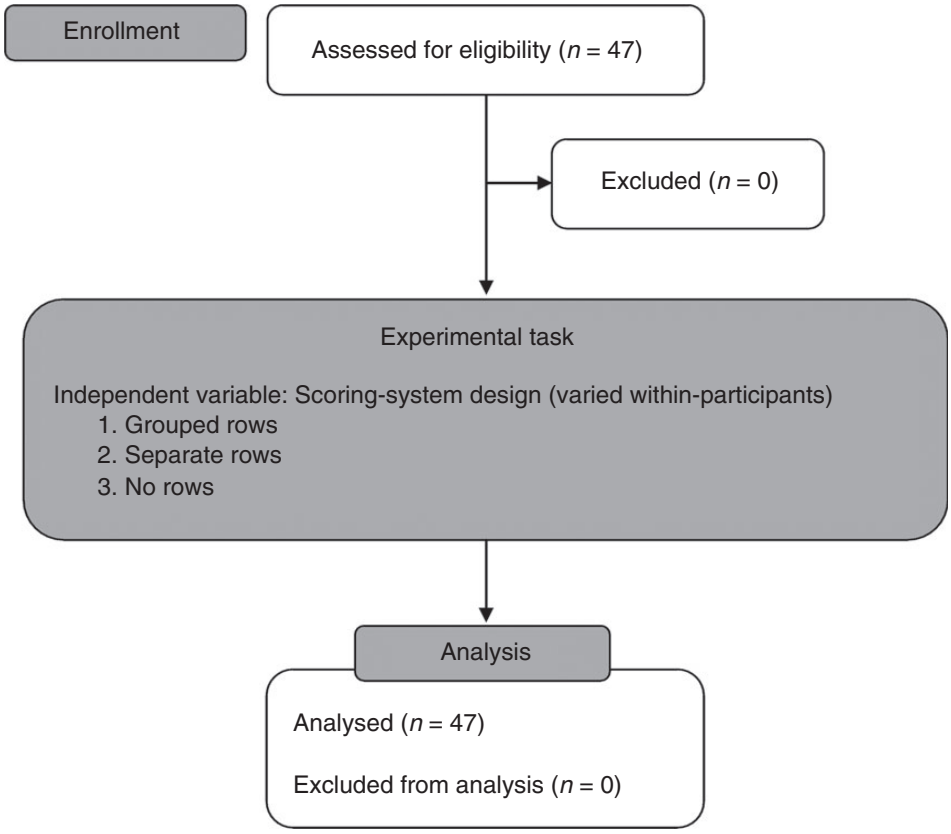
† Individual vital sign scores cannot be greater than five (Horswill et al. 2010, ACSQHC 2013).
* Only systolic blood pressure can yield individual vital sign scores of *4* or *5* (Horswill et al. 2010, ACSQHC 2013), thus the combination *4 4* cannot occur within a single time-point.

Participants

We recruited 47 novice chart-users (32 females and 15 males; mean age 21.49 years, *SD* 6.01) from a pool of undergraduate psychology students at The University of Queensland (St Lucia, Queensland, Australia). A naïve sample was deliberately selected to preclude the possibility that participants' prior chart-related preferences or experiences could advantage particular design features. It is also worth noting that, in our previous experimental studies comparing chart designs, samples of health professionals and chart novices (recruited via the psychology research participation scheme) consistently yielded very similar patterns of results across designs (Horswill *et al.* 2010, Preece *et al.* 2012b; Christofidis *et al.* 2014). Thus, we reasoned that including a group of non-naïve novices (e.g., nursing students) would have been unlikely to add additional value.

A minimum sample size of approximately 40 participants was sufficient to yield statistically significant pairwise performance differences between alternative chart designs in our previous work using similar methods (Preece *et al.* 2012b, Christofidis *et al.* 2013) where the differences were also deemed substantial enough to be of practical importance. Thus, we continued to recruit and test participants in the present study until the final sample exceeded this number. No participants were excluded from the analyses (Figure 2).

Fig. 2. Flow diagram illustrating the enrollment of participants, the within-participants experimental manipulation and the exclusions made prior to analysis.



Data collection

Participants were recruited and tested between December 2012 - January 2013 and received course credit. All participants gave informed consent; however, we did not inform them of the experimental hypotheses prior to participating.

Participants were tested individually in a quiet room and began by completing a demographic questionnaire. Next, they watched training videos that explained important background information, including: (a) the ten vital signs and their normal ranges (Horswill *et al.* 2010, ACSQHC 2013); (b) track-and-trigger systems; and (c) how to use each chart design (explained in a different random order for each participant). Participants' knowledge of key points from these videos was then tested with a 10-item multiple-choice examination. Participants who did not score 100% were required to study this information from a summary sheet and retake the examination until they did. A final training video explained the experimental protocol.

In the experiment, each participant completed nine blocks of experimental trials (one block per patient case), while standing next to a simulated patient (i.e., a mannequin in a hospital bed) to increase representativeness. In each block of trials, the participant was handed a chart attached to an open clipboard and then scored each set of observations (18 sets per block), working consecutively from the first time-point to the last. Each set of observations constituted one experimental trial and each participant completed 162 trials in total. Every time the participant recorded an overall early-warning score, they were also required to speak it aloud. This allowed the experimenter to record the response time for each set of observations using a software stopwatch. Responses were also audio recorded for verification purposes.

Each chart design was used on three blocks of trials (i.e., 54 trials per design) and the nine cases were randomly assigned to the three chart designs for each participant. To prevent order effects, the blocks were presented in a different random order for each participant.

Ethical considerations

This study was granted ethical approval in accordance with the review processes of the university ethics committees.

Data Analysis

For each set of observations, the overall early warning-score recorded by the participant was coded as correct or incorrect. For each (a) design and (b) combination of design and 'target' early-

warning score (i.e., 0-8), we calculated each participant's average response time (the mean number of seconds to record an early-warning score) and error rate (the number of incorrect early-warning scores as a percentage of all relevant early-warning scores).

For each participant, we also calculated the frequency of overall early-warning scores that were under- or over-scored on each chart design (expressed as percentages). In addition, we determined the magnitude of this under- and over-scoring for each design (i.e., each participant's mean deviation in each direction from the correct score).

For designs with individual vital sign scoring-rows, two specific error-types were coded, summed and expressed as percentages. A 'scoring error' occurred when a participant recorded an incorrect score for an individual vital sign. An 'adding error' occurred when a participant recorded an overall early-warning score that was not the sum of the individual scores recorded.

Statistical analyses were performed using IBM SPSS 21.0 (IBM Corp., Armonk, NY: USA) with statistical significance set at $\alpha = 0.05$. To compare chart designs, repeated-measures analyses of variance were conducted on response times and error rates, with η^2 calculated as the measure of effect size (Howell 1997). In addition, *t*-tests were used to compare the frequency of under-scoring vs. over-scoring, the size of under-scoring vs. over-scoring discrepancies and (for chart designs with individual vital sign scoring-rows) the percentage of time-points affected by scoring vs. adding errors, with Cohen's *d* as the effect size measure (Cohen 1992). We also examined correlations between the size of the 'target' early warning scores and the response time and error rate data for all three charts.

RESULTS

Response time

Analysis of the response time data revealed a significant main effect of scoring-row placement, $F(2, 92) = 306.99$, $p < 0.001$, $\eta^2 = 0.870$ (Figure 3A). When there were no rows for scoring individual vital signs, participants responded 6.35 seconds faster (CI 5.83-6.87) than when there were separate rows ($p < 0.001$) and 7.69 seconds faster (CI 7.17-8.20) than when there were grouped rows ($p < 0.001$). Participants were 1.34 seconds faster (CI 0.82-1.86) with separate versus grouped rows ($p < 0.001$). In addition, for each chart, response times were positively correlated with 'target' early-warning scores (grouped rows, $r = 0.98$, $p < 0.001$; separate rows, $r = 0.95$, $p < 0.001$; no rows, $r = 0.94$, $p < 0.001$), indicating that the more at risk the patient, the slower responses were likely to be.

Error rate

Analysis of the error rate data for the overall early-warning scores also yielded a significant main effect of individual vital sign score placement, $F(2, 92) = 5.57$, $p = 0.005$, $\eta^2 = 0.108$ (Figure 3B). Participants made 2.48% fewer errors (CI 0.86-4.11) when there were no rows for scoring individual vital signs, rather than separate rows ($p = 0.008$) and 2.76 % fewer errors (CI 1.01-4.50) when there were no rows than when there were grouped rows ($p = 0.007$). However, there was no significant difference between the separate and grouped rows conditions ($p = 1.00$).

Compared with over-scoring, under-scoring of overall early-warning scores occurred more frequently for the no rows design ($t(46) = -3.11$, $p = 0.003$, $d = 0.65$), affecting 1.70% more scores (CI 0.60-2.79) and for the separate rows design ($t(46) = -4.69$, $p < 0.001$, $d = 0.85$), affecting 3.20% more scores (CI 1.82-4.56) (Table 2). However, for the grouped rows design, there was no significant difference between the frequencies of under- and over-scoring ($p = 0.874$).

For the design with grouped rows, errors were 0.47 units (CI 0.06-0.88) bigger when participants over-scored compared with when they under-scored ($t(46) = -2.28$, $p < 0.05$, $d = 0.48$; Table 2). However, for the no rows design, errors were 0.38 units (CI 0.06-0.71) smaller when participants over-scored (no rows, $t(46) = 2.35$, $p < 0.05$, $d = 0.51$). For the design with separate rows, the size of the errors did not vary between under-scored and over-scored observations ($p = 0.537$).

Fig. 3. Response times (A) and error rates (expressed as percentages) (B) for recording overall early-warning scores, arranged by chart design (where only the placement of scoring-rows varied between the designs). Error bars indicate 95% confidence intervals and asterisks indicate significant differences between charts ($p < .01$).

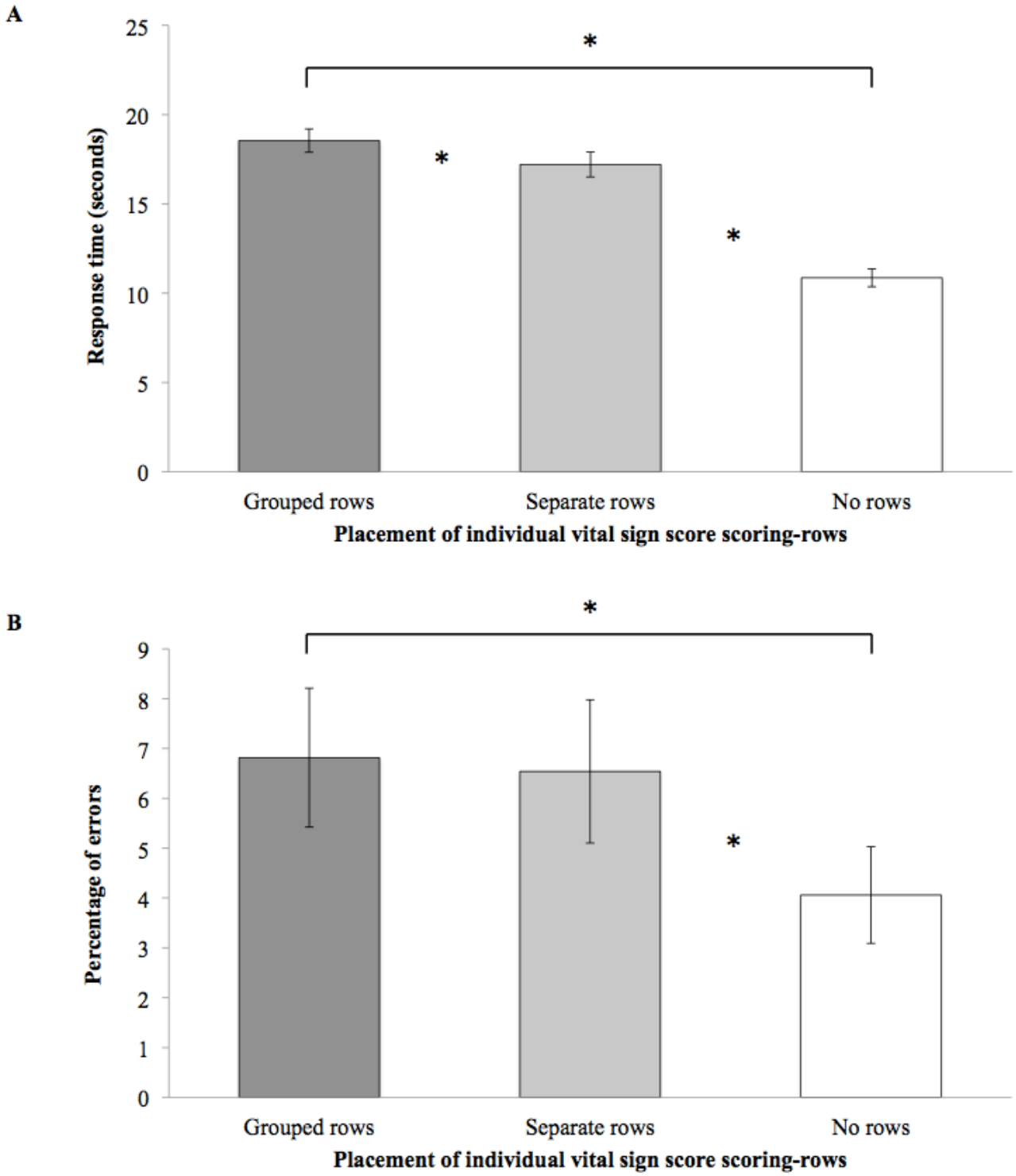


Table 2
Frequency of under-scoring and over-scoring in overall early-warning scores and mean discrepancy sizes, averaged across participants for each chart design.

Direction of discrepancy	Measure	Scoring-system design		
		Grouped rows	**Separate rows**	**No rows**
Under-scored	Percentage of overall early-warning scores under-scored (*SD*)	3.47% (*3.13*)	4.85% (*4.82*)	2.88% (*3.27*)
	Mean size of under-scoring discrepancy (*SD*)	1.00 (*0.69*)	0.77 (*0.51*)	0.87 (*0.84*)
Over-scored	Percentage of overall early-warning scores over-scored (*SD*)	3.35% (*4.19*)	1.65% (*2.29*)	1.18% (*1.70*)
	Mean size of over-scoring discrepancy (*SD*)	1.47 (*1.20*)	0.69 (*0.97*)	0.49 (*0.66*)

For both designs with individual vital sign scoring-rows (where scoring errors could be distinguished from adding errors), scoring errors affected significantly more time-points than adding errors. Specifically, scoring errors affected 3.35% more time-points (CI 1.87-4.83) on designs with grouped rows ($t(46) = -4.57$, $p < 0.001$, $d = 0.88$) and 2.56% more time-points (CI 1.03-4.09) on designs with separate rows ($t(46) = -3.37$, $p = 0.002$, $d = 0.63$) (Table 3). However, there was no significant difference between the two designs in the number of time-points affected by scoring errors ($p = 0.581$), or by adding errors ($p = 0.516$). Finally, for each chart, error rates were positively correlated with 'target' early-warning scores (grouped rows, $r = 0.87$, $p < 0.01$; separate rows, $r = 0.84$, $p < 0.01$; no rows, $r = 0.94$, $p < 0.001$), indicating that the worse state the patient was in, the greater the chance of error.

Table 3
Frequency of scoring and adding errors, averaged across participants for each chart design.

Error	Measure	Scoring-system design		
		Grouped rows	**Separate rows**	**No rows**
Scoring	Percentage of individual vital sign scores affected by scoring errors (*SD*)	0.65% (*0.62*)	0.58% (*0.63*)	-
	Percentage of time-points affected by (one or more) scoring errors (*SD*)	5.16% (*4.99*)	4.61% (*5.05*)	-
Adding	Percentage of time-points affected by adding errors (*SD*)	1.81% (*1.99*)	2.05% (*2.74*)	-
Both scoring and adding	Percentage of time-points affected by both scoring and adding errors (*SD*)	0.20% (*0.69*)	0.16% (*0.52*)	-

DISCUSSION

The results of the present study suggest that, in the case of integrated colour-based early-warning scoring systems, less is more. Contrary to hypotheses, preventing chart-users from recording individual vital sign scores yielded more efficient and accurate determination of overall scores, cutting both response times and error rates by around 40%. A potential explanation is that removing the individual vital sign scoring-rows eliminated the need for the additional visual switches (Rashid *et al.* 2012) demanded by the other two designs: between the observations and the scoring-rows at the bottom of the page (*grouped rows*), or from one scoring-row to the next (separate rows) (Figure 4). The data suggest that these switches may have impeded performance to an unexpected degree, whereas the concurrent tasks of determining each individual vital sign score and holding a running total in mind did not appear to compromise the low-level mental arithmetic required to derive an overall score on the *no rows* chart. Further, the absence of rows made this design comparatively less visually cluttered, which may have also facilitated more efficient and accurate engagement with the chart (Christofidis *et al.* 2012).

Although the two designs with individual vital sign scoring-rows did not differ in the frequency of either scoring or adding errors, participants determined the overall scores faster using separate, rather than grouped, rows. This could be due to the larger visual switches demanded by the grouped rows design (Figure 4). Because the scoring-rows are not adjacent to the corresponding vital sign data on the grouped rows chart, chart-users need to reorient themselves within a new visual space after each transition, exerting additional mental effort (Horswill *et al* 2010).

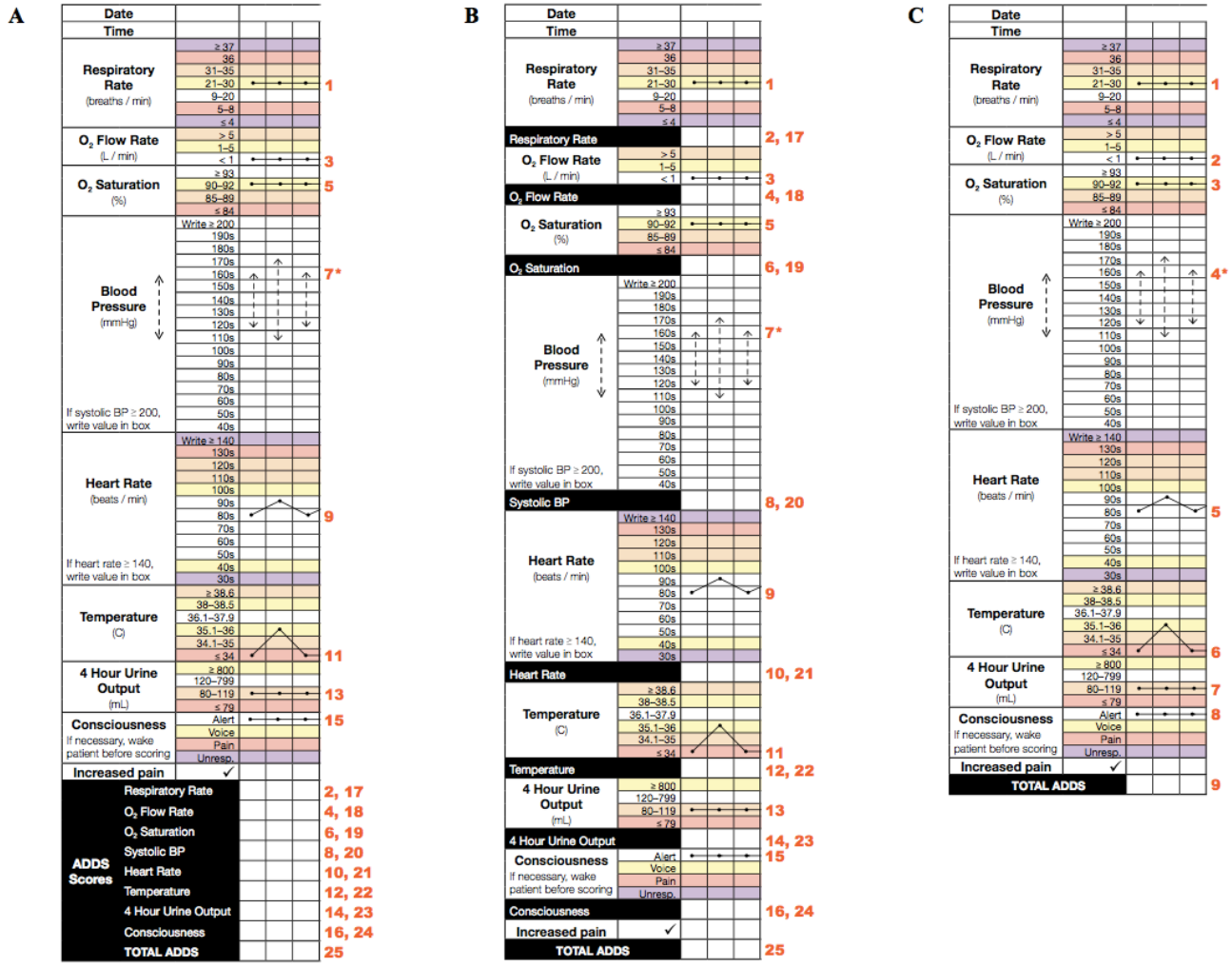
Interestingly, under-scoring of overall early-warning scores was more frequent than over-scoring for the separate rows and no rows charts (whereas they occurred at equal rates for the grouped rows chart). To calculate an overall score on the separate rows chart, users must visually align the column of individual scores down the entire page, switching from one vital sign to the next (Horswill *et al.* 2010; Figure 4). Hence, it is possible that interference from data recorded in-between the scores sometimes caused users to skip a score entirely. When completing a no scores chart, users need to remember not only the running total, but also which vital signs they have already scored. Even when working down the page from top-to-bottom, it is possible to accidentally skip over a vital sign when making visual switches between the observations and other parts of the chart (such as the scoring key).

Clinically, the implications of this study are critical. On the worst-performing chart design, incorrect overall early-warning scores were under-scored or over-scored by an average of 1.0 or 1.5 units, respectively. For some patients, this will be enough to trigger an inappropriately low (if under-scored) or high (if over-scored) response. For example, a one-unit under-score can be the difference

between a nurse being prompted to consider a MET call or merely to request a registrar review within 30 minutes (Horswill *et al.* 2010, ACSQHC 2013). This finding is even more alarming when we consider that, for all three chart designs, there were very strong positive correlations between 'target' early warning scores and both of the main outcome measures – response time and error rate. This suggests that, the more at risk the patient, the slower and more innaccurate responses are likely to be.

Some Australian hospitals have recently removed individual vital sign scoring-rows from their integrated colour-based early-warning scoring systems (Queensland Health 2012). Although we initially questioned this design decision and predicted that it would increase errors, the results of the current study support it.

Fig. 4. An illustration of the order in which a chart-user might typically attend to vital sign observation rows and scoring-rows when determining scores on each of the three chart designs: *grouped rows* (A); *separate rows* (B); and *no rows* (C). Red numerals indicate the potential order for the third column of vital sign data. Asterisks indicate that, at this step, all three charts also require the user to consult a blood pressure look-up table to the right (not pictured; Figure 1).



Limitations

The main limitation of this study is that we have not demonstrated directly that the results will generalize to real clinical settings (e.g., via a multi-site clinical trial of the three scoring system designs). Arguably, response times and error rates for all of the chart designs are likely to be greater under real-world conditions, where chart-users are faced with various external pressures and distractions. However, there are substantial costs associated with conducting clinical trials. Hence we argue that, in the chart development and validation process, it is typically more prudent to first conduct a series of lower-cost, more highly-controlled usability studies as a means of gathering preliminary evidence to inform or evaluate the major design decisions (e.g., Preece *et al.* 2012b, Christofidis *et al.* 2014). In this context, the present study serves as a template for usability studies focused on scoring system design and, to our knowledge, is the first of its kind.

In addition, we acknowledge that chart audits are required to determine whether the absence of scoring-rows impacts compliance with monitoring. That is, it is possible that the presence of scoring-rows encourages more accurate and comprehensive recording of observations. On charts with scoring rows, it is immediately evident whether all vital signs have been attended to. Hence, scoring rows may increase nurses' accountability and help them to detect their own accidental omissions.

A system without scoring-rows also relies more on trust. For example, nurses and doctors must trust that the last health professional who documented a patient's vital signs scored each observation correctly and summed the individual vital sign scores accurately. Trust is an important element in improving patient care in dynamic health care environments (Johns, 1996). If an observation chart design's lack of transparency leads nurses and doctors to distrust it, then they may resist its introduction, refuse to use it, or fail to comply properly with chart-related protocols (Preece et al., 2012a).

As with our previous behavioural experiments (Preece *et al.* 2012a, Christofidis *et al.* 2013, Christofidis *et al.* 2014), this study is also limited in that the findings may only apply to static paper-based domains, whereas hospitals will inevitably shift towards using electronic systems to record and display patient data. Indeed, compared with pen-and-paper methods, hand-held computers have already been found to help improve the accuracy and efficiency of early-warning score calculations in acute hospital care (Prytherch *et al.* 2006, Mohammed *et al.* 2009). However, we argue that paper-based observation charts are still globally ubiquitous and are likely to have a substantial shelf life, particularly in developing countries.

The recruitment of novice chart-users as participants also means that our findings, strictly speaking, cannot be generalized to experienced chart-users. Although controlling for past chart

experience was important in terms of maximising experimental control, we argue that the findings will still almost certainly apply to nurses and doctors for several reasons. First, the mechanical task of scoring invidividual vital signs and determining the total early-warning score does not rely on clinical knowledge or expertise (as opposed to the overall task of detecting deteriorating patients, where clinical judgement can be critical). Rather, it involves basic human capacities, such as visual perception, working memory and low-level addition. Second, in our previous experimental studies comparing observation chart designs, samples of chart novices and health professionals have consistently produced similar patterns of results across charts (Horswill *et al.* 2010, Preece *et al.* 2012b; Christofidis *et al.* 2014). Third, recent evidence has demonstrated that the effects of improved chart design on response times and error rates for detecting abnormal observations can outweigh health professionals' prior chart experience (Christofidis *et al.* 2013). In addition, the use of naïve participants was important because it is critical that observation charts provide effective support for clinical staff of all levels (including the least experienced), especially given that initial decisions about deteriorating patients are often made by newly-qualified nurses and doctors (Endacott *et al.* 2010). Again, the inevitable shift towards using electronic systems also means that, in the future, when it is likely that paper-based charts will be used exclusively as the back-up for electronic systems, all chart-users will effectively be novices (Christofidis *et al.* 2014). Nevertheless, it must also be emphasized that, although a well-designed observation chart can assist even the least experienced chart-user to recognize and respond to deteriorating patients, it is merely a decision-support tool and not a substitute for nurses' and doctors' good clinical judgment and training (McDonnell *et al.* 2013).

CONCLUSION

The results of this study suggest that integrated colour-based track-and-trigger systems may benefit from the exclusion of individual vital sign scoring-rows, potentially improving the effectiveness of the system and, ultimately, clinical responses. More broadly, the results demonstrate that even multi-disciplinary teams of clinicians and human factors specialists can make sub-optimal design choices and therefore that iterative empirical evaluations of clinical chart designs are essential. Because the processes involved in vital sign charting (whether computerized or paper-based) can be complex (Subbe *et al.* 2007), there remains enormous scope for further empirical usability research.

References

Andrews T & Waterman H (2005) Packaging: a grounded theory of how to report physiological deterioration effectively. *Journal of Advanced Nursing* 52, 473–481. doi: 10.1111/j.1365-2648.2005.03615.x

Australian Commission on Safety and Quality in Health Care [Internet] (2009) *Recognising and responding to clinical deterioration: Use of observation charts to identify clinical deterioration.* Available from: http://www.safetyandquality.gov.au/wp-content/uploads/2012/02/UsingObservationCharts-2009.pdf.

Australian Commission on Safety and Quality in Health Care [Internet] (2013) *Recognition and Response to Clinical Deterioration: Observation and Response Charts.* Available from: http://www.safetyandquality.gov.au/our-work/recognition-and-response-to-clinical-deterioration/observation-and-response-charts/.

Christofidis MJ, Hill A, Horswill MS & Watson MO (2012) *Human factors design and observation charts. 7th Annual International Conference on Rapid Response Systems and Medical Emergency Teams.* Sydney, New South Wales, Australia.

Christofidis MJ, Hill A, Horswill MS & Watson MO (2013) A human factors approach to observation chart design can trump health professionals' prior chart experience. *Resuscitation* 84, 657-665. doi: 10.1016/j.resuscitation.2012.09.023

Christofidis MJ, Hill A, Horswill MS & Watson MO (2014) Observation charts with overlapping blood pressure and heart rate graphs do not yield the performance advantage that health professionals assume: an experimental study. *Journal of Advanced Nursing* 70, 610-624. doi: 10.1111/jan.12223

Cohen J (1992) A power primer. *Psychological Bulletin* 112, 155-159. doi: 10.1037/0033-2909.112.1.155

Gao H, McDonnell A, Harrison DA, Moore T, Adam S, Daly K, Esmonde L, Goldhill DR, Parry GJ, Rashidian A, Subbe CP & Harvey S (2007). Systematic review and evaluation of physiological track and trigger warning systems for identifying at-risk patients on the ward. *Intensive Care Medi*cine 33, 667-679. doi: 10.1007/s00134-007-0532-3

Horswill MS, Preece MHW, Hill A, Christofidis MJ & Watson MO (2010) *Recording patient data on six observation charts: An experimental comparison.* Sydney, New South Wales, Australia, Australian Commission on Safety and Quality in Health Care.

Horswill MS, Preece MHW, Hill A, Christofidis MJ, Karamatic RM, Hewett DJ & Watson MO (2010) *Human factors research regarding observation charts: Research project overview.* Sydney, New South Wales, Australia, Australian Commission on Safety and Quality in Health Care.

Howell DC (1997) *Statistical methods for psychology.* Belmont, California, Duxbury Press.

Johns JL (1996) A concept analysis of trust. *Journal of Advanced Nursing* 24, 76-83. doi: 10.1046/j.1365-2648.1996.16310.x

Lawson L & Peate I (2009) *Essential Nursing Care. A Workbook for Clinical Practice*. Chichester, Wiley-Blackwell.

McDonnell A, Tod A, Bray K, Bainbridge D, Adsetts D & Walters S (2013) A before and after study assessing the impact of a new model for recognizing and responding to early signs of deterioration in an acute hospital. *Journal of Advanced Nursing* 69(1), 41-52. doi: 10.1111/j.1365-2648.2012.05986

Mitchell IA, McKay H, Van Leuvan C, Berry R, McCutcheon C, Avard B, Slater N, Neeman T & Lamberth P (2010) A prospective controlled trial of the effect of a multi-faceted intervention on early recognition and intervention in deteriorating hospital patients. *Resuscitation* 81, 658-666. doi: 10.1016/j.resuscitation.2010.03.001

Mohammed MA, Hayton R, Clements G, Smith G & Prytherch D (2009) Improving accuracy and efficiency of early-warning scores in acute care. *British Journal of Nursing* 18, 18-23.

Preece MHW, Hill A, Horswill MS, Karamatic R & Watson MO (2012a) Designing observation charts to optimise the detection of patient deterioration: Reliance on the subject preferences of healthcare professionals in not enough. *Australian Critical Care* 25, 238-252. doi:10.1016/j.aucc.2012.01.003

Preece MHW, Hill A, Horswill MS & Watson MO (2012b) Supporting the detection of patient deterioration: Observation chart design affects the recognition of abnormal vital signs. *Resuscitation* 83, 1111-1118. doi: 10.1016/j.resuscitation.2012.02.009

Prytherch DR, Smith GB, Schmidt P, Featherstone PI, Stewart K, Knight D & Higgins B (2005) Calculating early-warning scores - A classroom comparison of pen and paper and hand-held computer methods. *Resuscitation* 70, 173-178. doi: 10.1016/j.resuscitation.2005.12.002

Queensland Health [Internet] (2012) *NSQHS Standard 9 Clinical Deterioration Definitions sheet. National Safety and Quality Health Service*. Available from: http://www.health.qld.gov.au/psu/safetyandquality/docs/cd-audit-def.pdf.

Rashid U, Nacenta MA & Quigley A (2012) The cost of display switching: a comparison of mobile, large display and hybrid UI configurations. *AVI '12 Proceedings of the International Working Conference on Advanced Visual Interfaces*, 99-106. doi: 10.1145/2254556.2254577

Smith AF & Oakey RJ (2006) Incidence and significance of errors in a patient 'track and trigger' system during an epidemic of Legionnaires' disease: retrospective casenote analysis. *Anaesthesia* 61, 222-228. doi: 10.1111/j.1365-2044.2005.04513

Smith GB, Prytherch DR, Schmidt PE & Featherstone PI (2008) Review and performance evaluation of aggregate weighted 'track and trigger' systems. *Resuscitation* 77, 170-179. doi: 10.1016/j.resuscitation.2007.12.004.

Subbe CP, Gao H & Harrison DA (2007) Reproducibility of physiological track-and-trigger warning systems for identifying at-risk patients on the ward. *Intensive Care Medicine* 33, 619-624. doi: 10.1007/s00134-006-0516-8

Chapter 6

The primary purpose of this thesis was to resolve contentious hospital observation chart design decisions using a behavioural experimental approach. Observation charts are essential cognitive artifacts that represent the past actions and shared intentions of the hospital ward team, in addition to the current state of the patient (Norman, 1992; Rogers, Patterson, & Render, 2012; Sela & Auerbach-Shpak, 2014). However, these charts have traditionally been developed without direct reference to how users process information (Gerhardt-Powals, 1996). Without an understanding of human-system interactions, these designers risk: (a) developing ineffective and inefficient chart interfaces; (b) missing opportunities for novel designs; and (c) alienating their users (Gerhardt-Powals, 1996; Gillan & Schvaneveldt, 1999). This laissez-faire approach to design may be due, in part, to clinicians' perceptions of chart-related tasks. Rather than viewing the tasks as potentially crucial for patient survival, staff tend to perceive the measuring and recording of observations as a simple low-priority activity (Boulanger & Toghill, 2009; Mitchell et al., 2010; NICE, 2007a). As such, chart tasks tend to be left to the most junior clinical staff (Mitchell, 2012).

Effective document design is important to many different domains, and hence has been the focus of research for many years (Carliner, Verckens, & de Waele, 2006; Ganier, 2004; Hoeken & Korzilius, 2003). This research has tended to be multi-disciplinary, involving fields such as psychology, linguistics, graphic design, education, and technical communication, and hence research strategies have been correspondingly diverse (Carliner et al., 2006; Spyridakis & Wenger, 1992). Nonetheless, the general strategy of adopting a systematic research-based approach, in which effectiveness is formally evaluated against some performance criterion (Schriver, 1993), has been found to be able to yield effective outcomes. To give one example, researchers from the flight industry employed task observations, interviews and survey data to redevelop work cards that control aircraft inspection and maintenance tasks, finding significant improvements among wing inspectors following pre- and post- usability tests (Drury, Sarac, & Driscoll, 1997; Patel, Drury, & Lofgren, 1994).

The developers of the ADDS chart also followed a systematic research-based approach to design their form, where a criterion-based outcome was used to test design effectiveness. As with the aviation example above, they started with observations of the tasks involving patient charts, as well as informal interviews with users. They then conducted a heuristic evaluation of existing charts to guide design, adapting usability principles from computer-based design to fit the chart context. The criterion-based performance of the resulting chart was evaluated by, for example, determining whether it yielded lower users' error rates and judgement times than alternative chart designs.

Discussion of findings with respect to human factors design principles

Although the ADDS chart was designed to account for users' psychological processes (Preece et al., 2010c), some clinicians have questioned whether several of its design features represent best practice and have argued in favour of pre-existing chart designs. The designers of the ADDS chart lacked objective evidence to appropriately respond to these claims, as their original supporting study compared the chart to alternative designs that differed on multiple dimensions (Preece et al., 2012b). They were also unable to make definitive human factors arguments in favour of their design decisions. This is because the abstract nature of human factors design principles means that while one principle might support one particular design option, another principle (or in some cases, even the same principle) can support a different design option, including the alternatives that some clinicians preferred. The aim of this thesis was to attempt to resolve these issues by conducting a series of behavioural experiments that directly addressed each of the points of contention regarding best practice observation chart design.

Chapter 2 reported a study designed to address clinicians' concerns that the ADDS chart might be problematic for health professionals who were accustomed to alternative chart designs (Christofidis et al., 2013). When we considered this argument from a human factors perspective, we explored the possibility of 'expertise reversal', a well-known effect within the instructional learning literature. We speculated that the psychological processes involved in learning through instruction might be comparable to those involved in interpreting an observation chart. As such, we considered the possibility that the ADDS chart design might be counterproductive for experienced chart-users with acquired expectations (Kalyuga et al., 1998; Kirschner et al., 2006). It was critical to address this possibility because systems designed without attention to the end-user can put stress on individuals' capabilities (Wickens et al., 2004).

We found that participants were more accurate and faster at detecting abnormal observations when using the ADDS chart compared to several existing chart designs; even those that they were highly experienced with in their occupational role. Despite potentially bringing acquired knowledge schemas to the task, the instructional schemas of the ADDS chart did not appear to cognitively overload users' working memory resources (Kalyuga et al., 2003). That is, compared to instruction that relied more on pre-existing schemas for direction (i.e., the chart that users were familiar with), a system high in instructional guidance (i.e., the ADDS chart) did not compromise performance. Arguably, this is attributable to the user-friendly design of the ADDS chart, which was specifically developed to minimise users' cognitive load (Gerhardt-Powals, 1996; Nielsen, 1993). As described in Chapter 1, the ADDS chart designers sought to minimise cognitive load by adopting coloured range rows to signal observations that have crossed particular thresholds of abnormality.

Presumably, this design feature reduced users' need to mentally compare the observations to remembered normal ranges or to a look-up table. Like the ADDS chart, the existing chart that one of the highly experienced participant groups were trained in also used an integrated colour-based scoring system (see Chapter 2, Figure 1c). However, it differed from the ADDS chart in its use of written-number (vs. drawn-dot) observations. Once again, we attribute the performance differences between the designs (in part) to a comparatively reduced cognitive load: the drawn-dots may have prevented users from automatically reading the numerical observations and/or comparing them with clinical criteria stored in memory. Our findings may benefit future chart designs. We demonstrated that it is possible to design a chart that can advantage both chart novices (Preece et al., 2012b) and health professionals (Christofidis et al., 2013), in spite of the incongruent expectations and understandings that experienced users may bring. In this instance, trade-off design decisions based on the end-user were unnecessary. That is, novice users' needs were not at odds with the needs of experienced users (and vice versa). The results also suggest that old technologies need not constrain new ones (Thomas & Schneider, 1984). This supports the idea that, rather than preserve the status quo in fear of poor performance with change, work systems (like hospital charts) might benefit from a process of continuous improvement (Salvendy, 1997).

Chapter 3 described a study designed to address clinicians' arguments that the ADDS chart's use of separate blood pressure and heart rate plots would make the detection of deterioration more difficult, compared to charts that overlap these vital signs (because separate plots preclude the use of the 'Seagull Sign') (Christofidis et al., 2014). When we addressed this view from a human factors perspective, we explained the conflict that can arise when one principle (in this case, 'proximity compatibility') can be applied in more than one way. We considered that the best application of this principle might depend on which type of cognitive processing best advantages users' detection of deterioration. We speculated that if integrative processing benefits users during this task, then overlapping graphs might be beneficial.

We found that participants were more accurate and faster when using separate vital sign graphs, compared to when graphs overlapped. This result demonstrates the advantage of distant display proximity, suggesting that the task of detecting deterioration using vital sign data may benefit from independent processing (Wickens & Carswell, 1995), where users can separately assess each vital sign for deranged observations. Accordingly, integrative processing appears to have disadvantaged users. Although we speculated that close display proximity could exploit the physiological interrelationship between blood pressure and heart rate (e.g., where a borderline observation of one vital sign could cue users to notice an abnormal observation of the other), in line with Chapter 3's hypotheses, overlapping plots yielded no performance advantage. The visual clutter created by the overlapping plots may have made observations more difficult for users to

perceptually separate from one another. Our results also suggest that an integrated colour-based track-and-trigger system should adopt separate, rather than overlapping, vital sign graphs. It is probable that participants have confused heart rate's colour-coding for that of blood pressure on overlapping plots (where one colour-code system is assigned to heart rate and another to systolic blood pressure in the look-up table). We also found that the availability of the Seagull Sign failed to improve users' performance. This could be due to the extreme spatial proximity that the visual cue demands. Although the Seagull Sign may attract users' attention to a particular time-point, having to then discriminate between the two observations may have made it more difficult for users to actually identify the abnormality. Another possibility is that the Seagull Sign is not visually salient enough to direct users' attention to a given time-point in the first place.

Chapter 4 involved a study that focused on addressing clinicians' concerns that the ADDS chart's use of drawn-dot observations, an integrated colour track-and-trigger system and grouped scoring-rows might not support users' detection of abnormal vital signs (Christofidis et al., 2016). Indeed, from a human factors perspective, we were able to make several arguments in favour of alternative design options. For example, compared to drawn-dots, written-numbers might add redundancy (i.e., an abnormal observation recorded as a high or low value may be more noticeable). Compared to an integrated colour-based system, a non-integrated tabular track-and-trigger system might represent a more simplistic display that only includes information that users need. Finally, compared to grouped scoring-rows, separate rows might provide more immediate redundancy because of the increased proximity between the data for each vital sign and its corresponding score (Gerhardt-Powals, 1996; Nielsen, 1993; Wickens et al., 2004).

We found that participants were more accurate and faster using drawn-dot observations (vs. written numbers) and an integrated colour-based scoring system (vs. a non-integrated tabular system). Our results are consistent with the proposal that these two ADDS chart features minimised cognitive load and data-driven searches, two design principles that Preece et al. applied from the software and web design domains (see Chapter 1), by avoiding the need for users to have to read numerical observations and remember (or refer to) normal vital sign references ranges (Gerhardt-Powals, 1996; Nielsen, 1993). We also found that when participants had access to scores, they were faster using separate (rather than grouped) rows. On these charts, users may have been more likely to consult each scoring-row immediately after they assessed the observations of a particular vital sign. Thus, when participants noticed an abnormal score, the distance in which they moved their attention from the score to the corresponding observation would have been much shorter than on charts with grouped rows. In other words, charts with separate rows may provide users with more immediate redundant cues, leading to significantly faster recognition of deterioration. This finding demonstrates the challenge of applying human factors design principles. In Chapter 1, we described

the various ways in which the ADDS chart groups information that will be used together to decrease users' search time. Although we used this principle to argue in favour of the ADDS chart's use of grouped rows, in retrospect, it may better explain the human factors rationale for utilising separate rows. That is, it may be more advantageous to closely position an observation with its corresponding individual vital sign score than it is to position scores close to one another (Gerhardt-Powals, 1996; Nielsen, 1993; Wickens et al., 2004).

The study reported in Chapter 5 addressed clinicians' views that the ADDS chart's use of grouped scoring-rows might impair users' recording of individual vital sign scores (Christofidis et al., 2015). In addressing this concern, we theorised that when users determine an individual score and then move their attention to another section of the chart to record the score, the mental effort required to reorient their attention to the new visual space after a large visual switch might lead to recording errors. Thus, we raised the possibility that charts with grouped rows may actually impair users' recording of individual scores. Unexpectedly, we found that participants were more accurate and faster at calculating overall scores when they were prevented from recording individual vital sign scores altogether. Although we hypothesised that simultaneously determining each score while holding a running total in mind would hinder users' mental arithmetical calculations, our results suggest that the load on working memory was substantially less than we anticipated. For each progressively determined early-warning score, participants only had to remember to one digit (i.e., the intermediate score after each addition). Further, the retroactive interference in determining each individual vital sign score may not have been enough to prolong the storage period in which intermediate scores were held (Wickens & Hollands, 2000). At most, this is a two-step operation where users: (a) assess the colour of the range row that the observation lies in, and (b) cross-reference to the scoring key to determine the appropriate individual vital sign score. (Alternatively, users may remember which scores correspond to which colours, in which case, the task only involves the first step.) The cognitive loads associated with these tasks are substantially less than more typical illustrations of retroactive interference (e.g., forgetting a phone number before dialing because someone asked a question during the retention interval) (Wickens & Hollands, 2000).

The performance benefits associated with excluding scoring-rows may be partly explained by another human factors design principle: 'to minimise information access cost' (Wickens et al., 2004). When users choose information from a display, a certain amount of selective attention is required (Czaja & Sharit, 2012). If this selective attention has to 'move' from one display location to another to access information, there is typically a cost in time or effort (Wickens et al., 2004). To minimise this access cost, Wickens et al. (2004) proposed that frequently retrieved display elements should be positioned in a way where the cost of travelling between them is small. Arguably, the chart design without individual vital sign scoring rows adheres most to this principle. Of the charts

examined, this design requires the least amount of travelling between display elements. Rather than switching their attention back-and-forth (i.e., from the observations to the scoring-rows), users only need to move their attention progressively down the chart from one observation to the next. Per time-point, this involves 16 fewer visual switches than on either of the charts with scoring-rows (Rashid, Nacenta, & Quigley, 2012). From a practical perspective, removing individual vital sign scoring-rows may also resolve the potential design conflict highlighted in Chapter 4, where participants detected abnormal observations faster using separate (vs. grouped) rows when scores were recorded, but grouped rows when scores were absent (however, future studies should examine whether the performance advantage associated with excluding scoring-rows also holds for other chart-based tasks, e.g., detecting abnormal vital sign observations). The principle of minimising information access cost may also account for why separate rows yielded faster response times than grouped rows. As mentioned in Chapter 5, although both charts required the same overall number of visual switches (i.e., 25), arguably, the time cost of travelling from an observation to the corresponding individual vital sign scoring-row (and then to the next observation) would be greater when rows are grouped because the distances between these display elements are greater.

We also found, contrary to our predictions, that scoring errors affected more time-points than adding errors (for charts with scoring-rows): on the worst performing chart, only 1.81% of time-points were affected by adding errors, compared to the 5.16% that were affected by (one or more) scoring errors. The discrepancy between our hypotheses and our findings could be because, once again, we overestimated the cognitive demands of the task. Compared to other arithmetical tasks in nursing (e.g., dosage calculations that can involve the use of fractions, percentages, decimals and ratios) (Aschenbrenner & Venable, 2009), the addition of single digit individual scores is much less taxing. This finding suggests that if individual vital sign scores are to be included on an observation chart, then preference should be given to designs that facilitate the most accurate scoring.

This thesis demonstrates the utility of a human factors approach to chart design that is moderated by empirical testing. In Chapters 2, 3 and 4 we found that the ADDS chart design (or particular design features that it incorporates), which was developed by human factors researchers, performed significantly better than existing charts that were designed and supported by clinicians. Interestingly, in almost all of our findings, participants were faster using chart designs that they also made fewer errors with (suggesting the absence of speed-accuracy trade-offs).

Human factors design, which accounts for users' information-processing capabilities and limitations (Gerhardt-Powals, 1996; Proctor & Van Zandt, 2008; Rebelo & Soares, 2014), has been emphasised in the patient safety arena in recent years. This is fortunate for paper-based hospital charts which have traditionally been developed by clinicians who (a) have not received design

training, and (b) may be subject to cognitive biases that stem from overconfidence and lack of feedback. This approach has led to the widespread implementation of potentially dangerous chart designs. Chapter 2, for example, illustrated the substantial error rates associated with two existing Australian charts without track-and-trigger systems. For one group who were experienced in using observation charts, mean error rates reached 37.2% for the numerical chart, and 38.4% for the graphical chart. The traditional approach of relying on clinicians' opinions has also lead to strong cultural support for certain chart designs. In Chapter 3, we described clinicians' justification for the use of overlapping blood pressure and heart rate graphs based on the potential availability of the Seagull Sign. Despite any empirical evidence that the visual cue assisted chart-users to detect deterioration, it had been widely endorsed by health care staff in Australia and the United Kingdom. Our finding that the overlapping blood pressure and heart rate plots required to use the Seagull Sign can actually impair performance demonstrates the danger of cultural beliefs in health care. Pervasive false beliefs are somewhat unsurprising given that peoples' notions about how well they perform often fail to correlate with objective performance (Dunning et al., 2003).

The value of empirically-based evaluation approaches to chart design

The development of the ADDS chart also illustrates the importance of adopting a human factors approach at the beginning of the process of designing a system, so that higher-level decisions flow on to affect more detailed decisions (Proctor & Van Zandt, 2008). Too often, human factors experts are consulted only after a system has already been designed. Given the time and money that has already been invested, designers are likely to resist responding to the criticisms and suggested changes made by human factors experts. This can lead to an unsafe system that fails to support both user performance and satisfaction (Wickens et al., 2004). This has been the case for several paper-based medical charts designed in Australia. Early human factors input can not only make the design more effective and user-friendly (Rebelo & Soares, 2014), but can also reduce the costs involved in development (Sela & Auerbach-Shpak, 2014). In some circumstances, financial constraints may compromise the redesign of a system following a human factors analysis late in the design cycle (Wickens & Hollands, 2000).

This thesis also highlights the need to empirically evaluate design. Although evaluation is critical for all systems (Salvendy, 1997), the ADDS chart is a particularly good candidate given the complexity that surrounds the application of human factors design principles. These principles are intended to act as guides so that they can be applied to a variety of systems across many different industries. However, sometimes designers can be forced to choose between multiple conflicting design principles when attempting to solve particular design problems. To complicate the matter,

sometimes a single principle can be applied in more than one way. As discussed in Chapter 1, in instances where designers cannot rely on guidelines for unambiguous direction, there is often no simple resolution to the design issue (Proctor & Van Zandt, 2008; Wickens et al., 2004). The successful application of these design principles can be so unintuitive that even human factors specialists can make decisions that are less than optimal. For instance, in Chapter 5, we found the ADDS chart's use of grouped scoring-rows led users to make more errors and take longer to record overall vital sign scores, compared to alternative design options that Preece et al. (2010c) rejected. Although experienced designers may be tempted to trust their own intuition when faced with competing design considerations, their decisions are still grounded in opinion. Designers experienced in human factors can also be affected by their conceptual knowledge of a system's design. This can result in interfaces that are comprehensible to designers but unintelligible to users (Nielsen, 1993).

Our approach to evaluation demonstrates the value of laboratory-based behavioural trials. Laboratory experiments allow designers to manipulate the variables that they anticipate will affect user performance, while holding other variables constant (Wickens & Hollands, 2000). In our highly controlled usability studies (described in Chapters 3, 4 and 5), we were able to select individual chart elements of interest and control all other aspects of design. These trials are also relatively inexpensive to run, in the context of the cost of patient harm. If this experimental approach is adopted by future chart designers, evaluation should: (a) begin early in the development process so that preliminary evidence can inform major design decisions; (b) occur at a number of points during the development process to facilitate continuous iterative improvement; and (c) involve human factors specialists (Proctor & Van Zandt, 2008; Salvendy, 1997; Wickens & Hollands, 2000). That is, evaluation should not be regarded as separate from the design process: it should be regarded as an integral part of it.

Finally, this thesis highlights the importance of clinical context when design features from one chart type are applied to another. With limited empirical evidence surrounding chart design, there is a risk that future designers will apply features of the ADDS chart (including those supported by this thesis) to alternative charts. Because designers may not understand the human factors rationale behind the feature, they might apply it in an inappropriate way. For instance, after the development of the ADDS chart, we sought to improve the design of a state-wide blood glucose and insulin chart (Christofidis et al., 2012). Analogous to vital signs, clinicians aim to maintain hospitalised patients' blood glucose levels within a physiological reference range to prevent clinically significant hyper- and hypoglycemic events. In light of this thesis's findings, we suggested the use of an integrated colour-based track-and trigger system to facilitate the faster detection of abnormal blood glucose levels. However, because insulin charts are fundamentally

different from general observation charts, if blood glucose levels were recorded with drawn-dots (as per the ADDS chart design), usability could be compromised. First, time increments between blood glucose level readings can vary greatly. If the time-axis is not carefully examined, a busy clinician could interpret and act on a trend line of five blood glucose level readings taken at two hour intervals (i.e., over a 10 hour period) in the same way as a trend line of five readings taken at 15-minute intervals (i.e., over a 75 minute period). Second, straight trend lines that would invariably be drawn to connect one recording to the next may encourage users to incorrectly infer direct linear increases and decreases in blood glucose levels between adjacent recordings. Instead, we recommended the use of written-numbers which we hypothesised would encourage closer examination of the relationship between blood glucose level readings and the time-axis and discourage the assumption of linearity (Christofidis et al., 2012). Although this potential solution appears intuitive, defining the real world factors that are likely to affect the use of a system can be difficult (Proctor & Van Zandt, 2008).

Experimental limitations and their implications for future research

The four experiments presented in this thesis are adaptations of a similar paradigm to that used by Preece et al. (2012b). In laboratory-based settings, we measured the accuracy and efficiency with which participants performed realistic clinical tasks using different observation chart designs. Consequently, several experimental limitations apply to all the studies in the thesis.

Representative design

The first limitation involves the representativeness of the study. Representativeness refers to the extent to which the conditions encountered in the experiment map onto conditions beyond the experiment (which, in this case, might include nurses working with patients on hospital wards). That is, the representativeness of the experiments may have implications for the extent to which the findings of the experiments can be generalized (Araújo, Davids, & Passos, 2007; Hammond, 1998). (See Hammond (1998) and Araújo et al. (2007) for discussions of representative design in psychological research.) Although the results of Joshi et al. (2014) demonstrate the clinical efficacy of the ADDS design (e.g., reduced illness severity at intensive care unit admission), it is possible that some of our laboratory results may not directly transfer to other clinical environments because of factors not accounted for by our experiments. Given the need to extend research findings to real-world systems, generalisability is a critical goal in human factors (Salvendy, 1997; Wickens & Hollands, 2000). In this section, we will consider several aspects of representativeness including the

experimental environment, the generation and presentation of stimuli, and real-world performance factors.

Experimental environment

Rather than evaluate chart designs in actual clinical settings (e.g., hospital wards), we elected to test participants in a quiet room (clinicians in hospital training rooms, and novice chart-users in a university laboratory). Despite the advantages of laboratory-based behavioural experiments previously discussed, it is possible that the findings obtained in these highly-regulated environments may not completely generalise to real-world conditions (Wickens & Hollands, 2000). That is, there could be an incongruence between the test (laboratory) and target (real world) situation that may affect the interpretation of outcomes (Salvendy, 1997). For example, compared to the quiet test settings, hospital ward environments involve chart-users being exposed to noise (e.g., alarms), distractions (e.g., background conversation) and interruptions (e.g., questions from patients). Chart-users are also likely to be more stressed because of health professionals' high workload under time-pressure (Carayon, 2012). Arguably, information processing and clinical decision-making will be less optimal in these circumstances, such that users are likely to make more errors and take longer than our results suggest (for all of the chart designs examined).

The disparity between the test and real-world environments could also substantially limit the generalisability of specific findings. For example, in Chapter 5, contrary to our predictions, participants determined overall early-warning scores more accurately and efficiently using chart designs without individual vital sign scoring-rows (vs. with rows). However, it is possible that in real clinical settings, chart-users might be comparatively more susceptible to the aforementioned external influences when using this design. As described in Chapter 1, if chart-users are interrupted while they calculate an early warning-score, they may try to recall where they were up to in the calculation process, increasing the risk of a mistake. Alternatively, users may simply start the calculation from scratch, increasing the time it takes to perform the task. That is, the benefits of writing down individual vital sign scores may only become apparent in more challenging settings.

Stimulus generation

It is also important to consider the extent to which experimental stimuli map onto real-world stimuli (Salkind, 2010). The plotted vital sign observations used in this thesis reflect real physiological data, collected from large patient cohorts across several Australian hospitals. The observations were entered into an Excel spreadsheet (Microsoft Corporation 2011) from which patient cases were extracted according to pre-defined time-points. For example, in Chapter 3, a single case was taken from the dataset after every 13 consecutive time-points. Although authentic,

our use of patient data was highly controlled. For example, to standardise the amount of deterioration, each 'abnormal' patient case presented to participants in Chapters 2, 3 and 4 only included one abnormal observation from one vital sign. Although these cases were taken from real patient data, an isolated physiological derangement only reflects a small percentage of the hospital patient population. For example, in one cross-sectional survey, there was an average of 4.4 abnormal vital sign observations for each general ward admission (where on average, 1.2 vital signs were abnormal) (Harrison, Jacques, Kilborn, & McLaws, 2005). Also, the patient data presented to participants was recorded completely and accurately. Although this was essential from an experimental perspective, in practice, vital signs are not always recorded appropriately (e.g., in one reported case, 25% of observational data were missing) (Endacott et al., 2007; Leuvan & Mitchell, 2008).

Stimulus presentation

The presentation of test stimuli may also impact representativeness. In Chapter 2 and 5, we presented participants with real paper observation charts. For each trial, participants opened a closed chart, at which time the experimenter started an electronic stopwatch (using a computer program that was specifically designed for the studies). The experimenter then stopped the watch when participants made a verbal response ('normal' or 'abnormal' in Chapter 2; the overall early-warning score in Chapter 5). To reduce the risk of inter-individual differences, the same experimenter measured all participants' responses and response times. Participants' verbal responses were also audio recorded in case a trial needed verifying (e.g., if the response time for a given trial went unrecorded, the experimenter could later listen to the audio recording and re-time that trial). Although paper charts were used to simulate a realistic interaction between the user and the tool, this approach relied on the experimenter's own accuracy and reaction time to record participants' responses. Chapter 2 is also somewhat limited in that the experimenter stopped the watch when participants responded 'normal' or 'abnormal'. It was only after the watch was stopped that participants were required to specify which vital sign was abnormal. Although we do not anticipate that this would significantly impact the overall results, a more robust experimental strategy could have been to ask participants to instead say aloud the abnormal vital sign from the outset.

In Chapters 3 and 4 we presented chart designs on computer monitors, using a software package called SuperLab (Cedrus Corporation, 2007), rather than paper. Participants responded by clicking on the appropriate area of the screen (these areas included onscreen buttons in Chapter 3 and relevant vital sign graphing areas in Chapter 4). The software recorded participants' responses and response times for each trial, avoiding the human limitations of the experimenter noted

previously. This method was used to improve measurement accuracy and also to allow for more trials to be included in each study, as it facilitated the presentation of a large number of charts in rapid succession. This was especially advantageous in Chapter 4's factorial design experiment where multiple design elements were compared (indeed, over a thousand stimulus items were used in this experiment alone). Although this was not completely ideal from a fidelity perspective, we argue that this approach was still likely to map onto the same psychological processes (Salvendy, 1997) involved in the detection of abnormal vital signs. This methodology is also novel. To our knowledge, it the marks the first experimental use of computer software to present observation charts to participants (and then record their task responses).

Finally, it could be argued that some of the percentages were calculated from a relatively small number of repetitions (for example, each data point in Chapter 2 was calculated from 8 trials). The number of trials was limited by the session length, where adding more trials would risk introducing participant fatigue effects and testing participants over multiple sessions would risk participant attrition. However, a counterargument is that if the number of trials was inappropriately small then we would predict that these would introduce noise into the data as a result of under-sampling participants' behaviour. This in turn would be predicted to increase the chances of non-significant results (i.e. a Type II error, due to insufficient psychometric reliability in our measurements). However, given that all of our studies did yield statistically reliable results, then this could be argued to indicate that our measurements did have an appropriate level of reliability.

Real-world performance factors

The laboratory experiments presented in this thesis are also limited in that they do not examine the effect of chart design on many real-world performance factors. This is, in part, because the general observation chart serves multiple purposes in clinical practice. Addressing the effect of design on performance can become complex when a document seeks to achieve several purposes and communicative effects (Lentz & Pander Maat, 2004). Although the ADDS chart was designed to improve health professionals' detection of deterioration, it also serves many other roles. Indeed, differences in how health professionals engage with observation charts elucidate several roles that are not captured by this thesis. Informal observations across various hospital wards have revealed that nurses tend to engage in a prescribed sequence of actions. For instance, every few hours (the exact interval depends on the clinical state of the patient and/or local hospital protocols), a ward nurse will typically measure a patient's vital signs; document the corresponding values on the observation chart; calculate the early warning score (as per the chart design); review the recorded observations for abnormalities; and when necessary, escalate the clinical response (i.e., contact the treating doctor or make a Medical Emergency Team call). Our experiments, however, have only

captured two of these roles: that is, detecting abnormal vital sign observations (Chapter 2, 3 and 4) and determining early warning scores (Chapter 5).

The limited scope of this thesis restricts our conclusions about the usability of the ADDS chart in-practice, especially the effect of design on clinical decision-making. The incorporation of escalation protocols (i.e., recommended response actions based on the degree of physiological deterioration) illustrates how one design feature can affect the ways in which a chart is actually used. In a focus group study with clinical ward staff, Elliot et al. (2015) found that escalation protocols empowered nurses with less clinical experience. This subset of nurses reported that the inclusion of such protocols 'permitted' them to call for assistance, mitigating the riskier wait-and-see approach that can exist in hospitals. For these health professionals, decision-making was made more straightforward. However, more experienced nurses reported that the very same design feature compromised their professional autonomy. In clinical instances where professional judgment failed to align with protocol, nursing staff resented that they were not allowed to amend the calling criteria or escalate the response to a level they deemed more appropriate (e.g., to have a patient reviewed by a doctor instead of a senior nurse) (Elliott et al., 2015). This attitude can impact on clinical practice. Nurses admitted to falsifying abnormal vital signs (when they judged that a given observation was acceptable for the patient, despite falling out of the reference range) rather than accurately recording the observation, precluding the need to follow the appropriate recommendations or justify their omission (Elliott et al., 2015).

This thesis also fails to capture many doctor roles. Doctors typically refer to a patient's observation chart to answer clinical questions, monitor physiological trends, and guide treatment decisions (Elliott et al., 2015). For example, if a patient was admitted to the Emergency Department with right-sided abdominal pain and vomiting, the treating doctor may inspect the documented temperature observations to assess for fever (clinical features suggestive of appendicitis). The presence or absence of febrile observations would influence the doctor's differential diagnoses and subsequent management plan (e.g., calling for an urgent surgical consultation vs. ordering an abdominal CT scan). Similarly, if a cardiac patient was being treated with anti-hypertensive medications, the doctor may refer to the chart to assess the patient's blood pressure observations over a number of days. The physiological trend in the data would then help the doctor decide whether the dose should be adjusted. These aspects of chart-use and clinical decision-making are critical to examine given the potential influence of design. For example, despite our experimental evidence in support of drawn-dot observations (vs. written-number), Elliot et al.'s (2015) focus groups reported that the use of vital sign ranges, rather than exact numbers, hindered inter-professional communication. In clinical reviews, for instance, doctors reportedly insisted that nurses provide them with exact values. Consequently, nurses had to re-measure patient vital signs, creating

redundancies in workload. Once again, nurses admitted to violating proper chart protocols. Some staff reported purposely documenting observations in written-number format, despite being aware that observations should be recorded as drawn-dots.

Chapter 4 also flagged the possibility that particular design features may affect real-world compliance with a chart. We speculated that including, rather than excluding, individual vital sign scoring-rows might encourage health professionals to document observations more accurately and comprehensively. We consider a scenario where a nurse has failed to record a particular vital sign observation. Respiratory rate observations, for example, are often neglected (in one reported case, 75% of the time) because they are one of the few vital signs that are measured without a manual instrument or electronic machine (Chatterjee et al., 2005; Leuvan & Mitchell, 2008). When users sum individual vital sign scores to calculate an overall score, they may be less likely to purposely exclude the score of a missing vital sign if scoring-rows are included, because a blank row would provide clear evidence of the user's omission (e.g., to other staff members). On charts without rows, however, neglecting a particular vital sign might be less noticeable. If the missing vital sign observation goes unnoticed by other staff, there are no other cueing features. The presence of a blank row might also help users to notice their accidental omissions in measuring and/or recording vital sign observations. It is critical to assess the social and organisational variables that might influence chart-use (Proctor & Van Zandt, 2008) as research has demonstrated that some nurses are more diligent in their recording of vital signs than others (Endacott et al., 2007). However, we do acknowledge that some compliance issues may be beyond the scope of design and may require other strategies such as education. For instance, when respiratory rate is documented, it is frequently recorded as 20 breaths per minute, a number that might suggest overgenerous rounding or even data fabrication (Chatterjee et al., 2005).

Future research

To address the above limitations, future research could evaluate the effect of chart design in real clinical environments with a random cross-section of genuine chart-users. For example, clinical trials could compare the effect of varying chart designs on health professionals' abilities (e.g., to detect deterioration or calculate overall warning scores) across multiple hospital sites. To date, clinical evaluations of (modified) ADDS charts have only evaluated the effect of newly implemented early-warning scoring systems in terms of cardiac arrest rates (Drower et al., 2013) and clinical user experiences (Elliott et al., 2014). However, future studies need to carefully consider the financial costs involved in clinical trials, as well as the potential ethical issues surrounding the comparison of chart designs using real patients' clinical deterioration as the dependent variable (e.g., where the effects of an inferior design, allocated to one hospital ward, are

compared to that of a superior design, allocated to another hospital ward). Future studies, laboratory or hospital-based, could also explore the effect of chart design on those aspects of chart-use that are yet to be examined. This could include the above-mentioned tasks, either in isolation or combination for a more complete picture. Additional studies could also go on to explore some of the real-world performance factors that may be affected by chart design (e.g., the actual incidence of nurses falsifying patient observations to avoid escalating a response when subjectively deemed unnecessary).

Chart audits could also be valuable. Retrospective chart audits are a common way for hospitals to identify user errors and adverse events (Fitzpatrick & Kazer, 2011). However, audits can also evaluate improvements. For example, a recent retrospective audit of the Q-ADDS chart in an Intensive Care Unit revealed a reduction in patients' illness severity at admission and their length of stay (Joshi et al., 2014). Future chart audits could also inform the results of this thesis. For example, a post-implementation audit of the chart designs in Chapter 5 (i.e., grouped vs. separate vs. no scoring-rows) could be used to examine the effect of scoring-rows on health professionals' compliance with vital sign measurements

Representativeness of participants

Another generalisability issue relates to the representativeness of the participants. In Chapters 2 and 3, we recruited purposive samples of doctors and nurses who volunteered in response to flyers advertising a study that aimed to improve the identification of patient deterioration. Although the samples were drawn from the population of interest (Salvendy, 1997), this recruitment method ran the risk of only including health professionals who were particularly interested in improving patient safety. In light of the abovementioned differences in nurses' levels of diligence (Endacott et al., 2007), it is possible that our samples of health professionals would perform better than a random sample of doctors and nurses. In human factors research, it is necessary to evaluate a 'generalised' user's interaction with a system, rather than just one particular type of user. In this way, designers can be more confident that the design will be appropriate for a broad class of system users (Wickens & Hollands, 2000). In Chapters 4 and 5, we only employed convenience samples of novices. By relying on undergraduate psychology students as the novice chart-user group, we need to be cautious when generalising the results of these two chapters to health professionals (especially those who are not novices). Although it was important to assess novices, user testing with real clinical staff is still fundamental in human factors design because it provides direct information about how these end-users interact with the interface (Nielsen, 1993). However, as detailed in the preceding chapters, we argue that the overall pattern of results are still very likely to apply to health

professionals (see Chapters 4 and 5 for details), given, for example, the strong similarities between the performance outcomes of health professionals and psychology students found in previous experiments.

Sample size

It is also important to consider the recruitment of doctors and nurses in Chapters 2 and 3. In both experiments, there are marked differences between the numbers of novice chart-users vs. health professionals. In Chapter 2, the difference between the chart experience groups is especially pronounced. This is largely attributable to the types of hospitals from which participants were recruited. Health professionals experienced with the multiple parameter track-and-trigger chart (n = 64) were recruited from a large major metropolitan hospital, while those experienced with a no track-and-trigger graphical chart (n = 37) were sourced from significantly smaller, more regional hospitals with less staff members. Similarly, in Chapter 3, there is a substantial difference between the number of Seagull Sign trained nurses (n = 41) and novices (n = 65). This reflects the relative difficulty of recruiting health professionals compared to first-year undergraduate students. Because we employed mixed-design ANOVAs in both chapters, where chart experience group (between-groups) and chart type (within-groups) comprise the independent and repeated factors respectively, the differences in sample sizes between the independent groups do not compromise the overall results. However, it is important to consider the statistical power of the individual groups. For each chapter, a power analysis was performed using G*Power 3.1.9.2 (Faul, Erdfelder, Lang, & Buchner, 2007) to calculate the sample size necessary to detect a medium effect size (partial $\eta^2 = 0.06$) for the least sensitive effect (Cohen, 1988), with alpha set at 0.05 and 80% power (see Table 5). We conservatively estimated a correlation among repeated measures of .85, based on the results of Preece et al. (2012b).

Table 5. For each chapter, the minimum sample size according to the power analysis and the actual sample size included in the final statistical analysis

	Minimum sample size according to power analysis	Sample size included in final analysis
Chapter 2	110	101
Chapter 3	135	186
Chapter 4	108	205
Chapter 5	10	47

Given the aforementioned difficulties in recruiting health professionals, the sample size included in the final analyses of Chapter 2 is just short of the minimum prescribed by the power analysis. We recruited and tested until we reached approximately 40 participants per group, as this

sample size yielded significant pairwise performance differences (deemed substantial enough to be of practice importance) between alternative chart designs in Preece et al. (2012b). Nonetheless, it should be noted that, in spite of this, we obtained statistically significant results, indicating that inadequate power was unlikely to be a problem in this study.

Sensitivity and response bias

In the preceding chapters, we did not address a critical element of chart-users' responses. When participants gave an incorrect response, were they more likely to mistake a normal patient case for an abnormal case? Alternatively, were they more likely to mistake an abnormal case for a normal case? The latter possibility represents a failure to detect physiological deterioration that, as highlighted in Chapter 1, can result in delayed or missed intervention. It would be especially concerning if chart-users were more likely to miss abnormal observations when using particular chart designs. For example, although the ADDS chart designs yielded low overall error rates, it is possible that when chart-users *do* make errors using the novel designs, they are frequently missing deterioration (that is, more often than they are mistaking a normal case as abnormal).

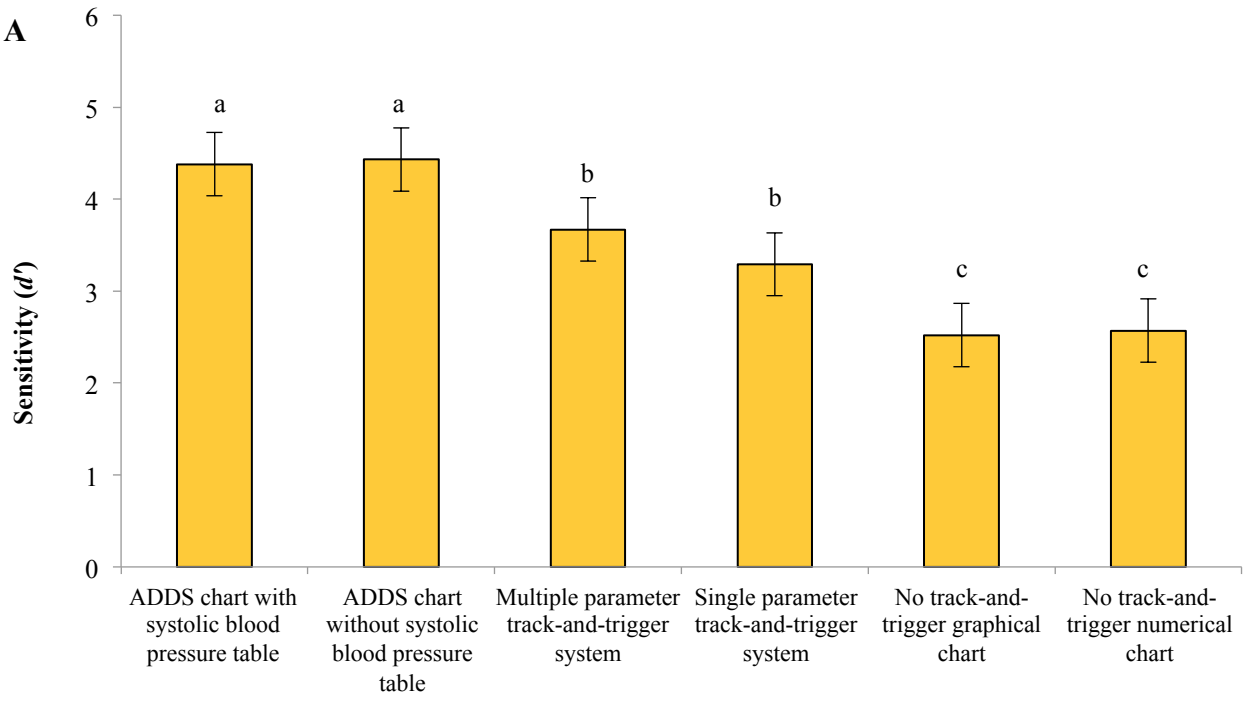
To evaluate this, we redefined the dependent variables in Chapter 2, 3 and 4 in Signal Detection Theory terms: a 'hit' occurred when a participant (correctly) detected an abnormal observation when one was present, and a 'false alarm' occurred when a participant (incorrectly) detected an abnormal observation when one was *not* present (i.e., when all observations were normal). For each participant and chart design, we calculated the hit rate (the number of hits divided by the number of abnormal trials) and false alarm rate (the number of false alarms divided by the number of normal trials). These rates were converted into measures of sensitivity (*d'*), where a high sensitivity index indicates a more accurate distinction between signal and noise. We also calculated response bias (β) to account for participants' potential response strategies. That is, in cases of uncertainty, participants may be more likely to give a positive response (i.e., employing a liberal strategy, indicated by β indices less than 1) or negative response (i.e., using a more conservative strategy, denoted by β indices greater than 1) (Stanislaw & Todorov, 1999).

To compare chart designs, repeated-measures analyses of variance were conducted on *d'* and β using IBM SPSS 21.0 (IBM Corp., Armonk, NY: USA) with statistical significance set at $\alpha = 0.05$. The measure of effect size was calculated using η^2 (Howell 1997). Figures 18, 19 and 20 illustrate: (A) the mean sensitivity values for each chart design, where 0 indicates an inability to distinguish abnormal cases from normal cases (where the greater the value above 0, the better participants were at distinguishing abnormal from normal cases); and (B) the mean response bias values for each chart design, where a response bias of 1 indicates that participants favour neither the

'abnormal' nor 'normal' response, values less than 1 signify a bias towards responding 'abnormal', and values greater than 1 denote a bias towards the 'normal' response.

Chapter 2

Analysis of the sensitivity index revealed a significant main effect of chart design, $F(4.33, 359.05) = 38.66$, $p < 0.001$, $\eta^2 = 0.94$ (see Figure 18A for pairwise comparisons between chart designs). Participants were significantly more accurate at differentiating between normal and abnormal patient cases using the ADDS chart designs, compared to the four alternative charts. Response bias indices less than 1 (see Figure 18B) demonstrate that, across all charts, participants responded liberally (i.e., in cases of uncertainty, participants deemed a patient case abnormal rather than normal). Analysis of the response bias index also revealed a significant main effect of chart design, $F(4.31, 358.01) = 13.49$, $p < 0.001$, $\eta^2 = 0.56$ (see Figure 18B). Participants favoured a liberal response significantly more when detecting deterioration on the ADDS chart designs (as well as the multiple parameter track-and-trigger system chart), compared to the single parameter track-and-trigger system chart, no track-and-trigger graphical chart and no track-and-trigger numerical chart. This finding suggests that on these chart designs, participants who were uncertain were more likely to judge a case as abnormal, erring on the side of caution.



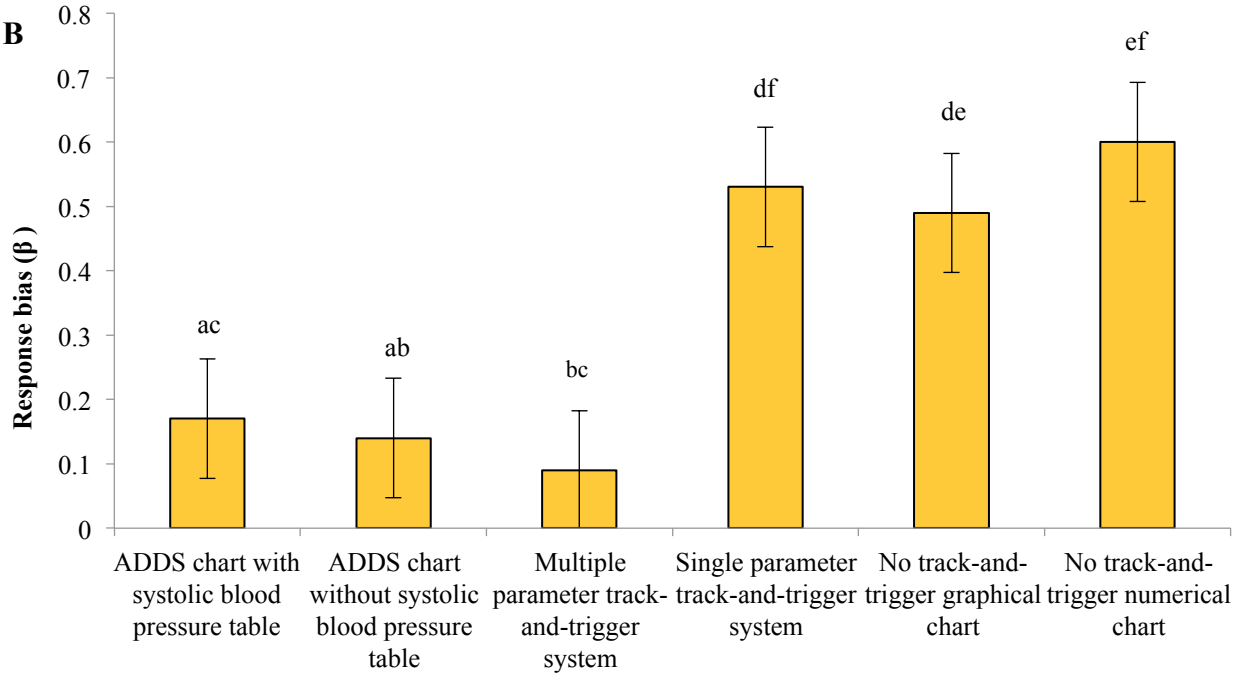


Figure 18. Measures of sensitivity (A) and response bias (B) for detecting abnormal observations on the six chart designs in Chapter 2. Error bars indicate 95% confidence intervals. Within each group, different letters indicate significant differences at the 5% level.

Table 6. Mean (SD) hit and false alarm rates on the six chart designs in Chapter 2

	Hit rate	Miss rate	False alarm rate	Correct rejection rate
ADDS chart with systolic blood pressure table	0.97 (0.09)	0.03 (0.09)	0.09 (0.14)	0.91 (0.14)
ADDS chart without systolic blood pressure table	0.97 (0.11)	0.03 (0.11)	0.08 (0.16)	0.92 (0.16)
Multiple parameter track-and-trigger system	0.89 (0.19)	0.11 (0.19)	0.14 (0.18)	0.86 (0.18)
Single parameter track-and-trigger system	0.93 (0.15)	0.07 (0.15)	0.23 (0.19)	0.77 (0.19)
No track-and-trigger graphical chart	0.86 (0.20)	0.14 (0.20)	0.30 (0.16)	0.70 (0.16)
No track-and-trigger numerical chart	0.87 (0.21)	0.13 (0.21)	0.31 (0.15)	0.69 (0.15)

Chapter 3

Analysis of the sensitivity index revealed a significant main effect of chart design, $F(2.67, 408.64) = 23.94$, $p < 0.001$, $\eta^2 = 0.87$ (see Figure 19A for pairwise comparisons). Participants were significantly more accurate at differentiating between normal and abnormal patient cases using the ADDS chart style design (i.e., separate blood pressure and heart rate graphs with a track-and-trigger system), compared to the three alternative chart extracts. Once again, response bias indices less than 1 (see Figure 19B) demonstrate that across all designs, participants responded liberally (i.e., when uncertain, participants were more likely to judge a case as abnormal over normal). Analysis of the response bias index also revealed a significant main effect of chart design, $F(2.64, 403.24) = 2.96$, $p < 0.05$, $\eta^2 = 0.44$ (see Figure 19B). Compared to the extract with overlapping graphs and no track-

and-trigger system, participants favoured the 'abnormal' response significantly less when detecting deterioration on the ADDS chart style design. (There was no significant difference between the ADDS chart style design and the other two extracts.) The comparatively high miss rates for the extract with overlapping graphs and no track-and-trigger system (see Table 7) suggests that participants may have found it difficult to differentiate between abnormal heart rate and blood pressure observations using this design.

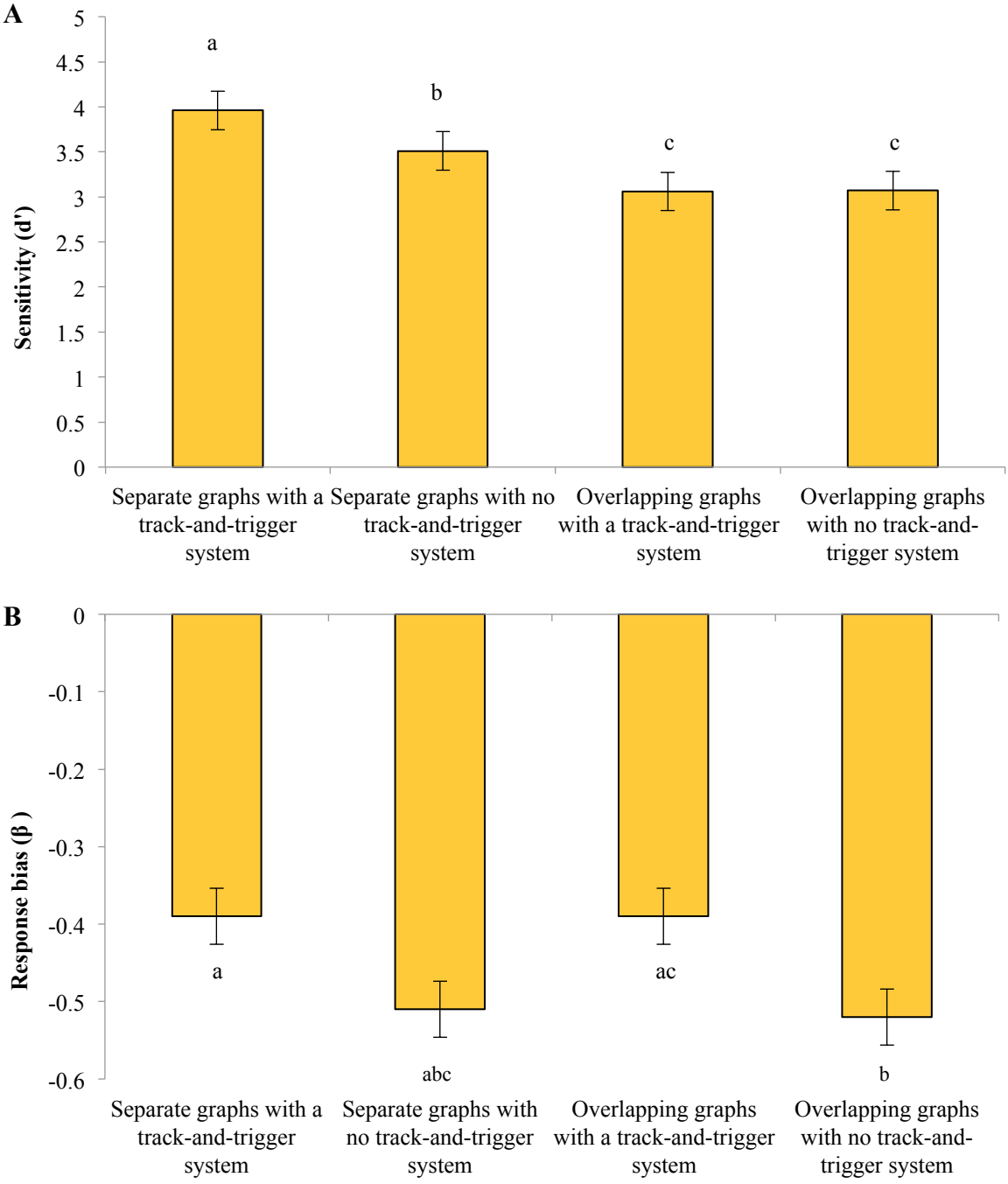


Figure 19. Measures of sensitivity (A) and response bias (B) for detecting abnormal observations on the four chart extracts. Error bars indicate 95% confidence intervals. Within each group, different letters indicate significant differences at the 5% level.

Table 7. Mean (SD) hit and false alarm rates on the four chart extracts in Chapter 3

	Hit rate	Miss rate	False alarm rate	Correct rejection rate
Separate graphs with a track-and-trigger system	0.85 (0.20)	0.15 (0.20)	0.04 (0.09)	0.96 (0.09)
Separate graphs with no track-and-trigger system	0.78 (0.23)	0.22 (0.23)	0.06 (0.14)	0.94 (0.14)
Overlapping graphs with a track-and-trigger system	0.78 (0.22)	0.22 (0.22)	0.12 (0.23)	0.88 (0.23)
Overlapping graphs with no track-and-trigger system	0.74 (0.23)	0.26 (0.23)	0.08 (0.15)	0.92 (0.15)

Chapter 4

Analysis of the sensitivity index revealed a significant main effect of chart design, $F(6.55, 1225.07) = 3.529$, $p < 0.05$, $\eta^2 = 0.96$ (see Figure 20A for pairwise comparisons). Once again, participants were significantly more accurate at differentiating between normal and abnormal patient cases using the ADDS chart style design (i.e., an integrated colour track-and-trigger system with grouped scoring-rows and drawn-dot observations) compared to the alternative chart designs. However, analysis of the response bias index revealed no significant main effect of chart design, $F(6.56, 1227.09) = 1.55$, $p = 0.152$, $\eta^2 = 0.47$.

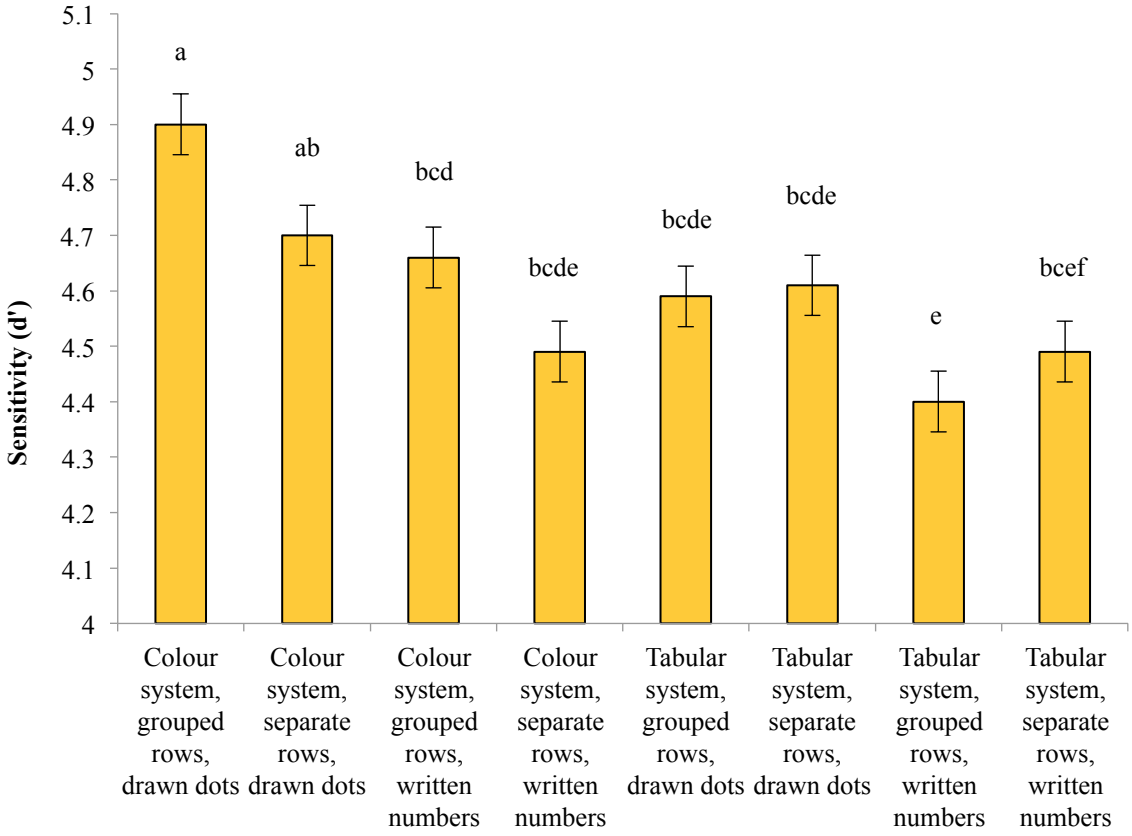


Figure 20. Measures of sensitivity for detecting abnormal observations on the eight chart designs. Error bars indicate 95% confidence intervals. Within each group, different letters indicate significant differences at the 5% level.

Table 8. Mean (SD) hit and false alarm rates on the eight chart designs in Chapter 4

	Hit rate	Miss rate	False alarm rate	Correct rejection rate
Integrated colour-based scoring-system, grouped scoring-rows, drawn-dot observations	0.92 (0.15)	0.08 (0.15)	0.01 (0.05)	0.99 (0.05)
Integrated colour-based scoring-system, separate scoring-rows, drawn-dot observations	0.90 (0.17)	0.10 (0.17)	0.02 (0.07)	0.98 (0.07)
Integrated colour-based scoring-system, grouped scoring-rows, written-number observations	0.89 (0.20)	0.11 (0.20)	0.02 (0.07)	0.98 (0.07)
Integrated colour-based scoring-system, separate scoring-rows, written-number observations	0.87 (0.18)	0.13 (0.18)	0.02 (0.10)	0.98 (0.10)
Non-integrated tabular scoring-system, grouped scoring-rows, drawn-dot observations	0.89 (0.19)	0.11 (0.19)	0.03 (0.09)	0.97 (0.09)
Non-integrated tabular scoring-system, separate scoring-rows, drawn-dot observations	0.88 (0.18)	0.12 (0.18)	0.02 (0.07)	0.98 (0.07)
Non-integrated tabular scoring-system,	0.88 (0.20)	0.12 (0.20)	0.04 (0.10)	0.96 (0.10)

grouped scoring-rows, written-number observations				
Non-integrated tabular scoring-system, separate scoring-rows, written-number observations	0.88 (0.20)	0.12 (0.20)	0.03 (0.10)	0.97 (0.10)

These findings suggest that the superior performance of the ADDS chart designs were not compromised by high rates of missed deterioration. Across Chapter 2, 3 and 4, we found that participants were more accurate at differentiating between normal and abnormal patient cases using the ADDS chart style designs and were likely to favour a conservative, safer response. This provides further evidence that the ADDS chart designs support chart-users' detection of deterioration.

Application of the results

This thesis is also limited in that the results may only apply to paper-based domains. In a field where the use of information and computer technologies is continually expanding and updating (Dekker, 2011), paper-based systems are becoming increasingly obsolete (Proctor & Van Zandt, 2008). Hospital observation charts are no exception: it is inevitable that computerised forms will eventually replace paper charts. Already, a number of electronic systems have been developed in Australia and overseas to assist health professionals to collect vital sign observations, detect deterioration and escalate levels of care. Empirical evidence has supported the transition to computerised vital sign monitoring systems. Positive clinical outcomes have been reported in the literature, including improvements in clinical attendance to deteriorating patients, patient mortality and length of stay in hospital, as well as the time staff spend recording vital signs (Bellomo et al., 2012; Jones et al., 2011). Computerised systems have also been shown to be very well accepted by clinical staff (Wood & Finkelstein, 2013). In one study, nurses perceived the computer-based system as more accurate, fast, simple and convenient than pen-and-paper methods (Prytherch et al., 2006). Although these findings are promising, the human-system interaction is almost always complex regardless of the medium (Prytherch et al., 2006). Consequently, computerised systems also require careful consideration from a human factors perspective. For example, it is currently unknown which empirically-supported features of the ADDS chart could be used successfully in an equivalent computerised system. One electronic automated advisory vital signs monitoring system, which has been clinically evaluated in recent years (Bellomo et al., 2012), displays observations (for a given time-point) on-screen as a set of numerical values. Although our findings support the use of drawn-dot observations, an important empirical question is whether, in the context of an electronic system, there are clinical and human factors advantages to presenting observations in

numerical form. At present, it is also unclear which of the elements that are necessary on paper should be retained by computerised systems. For example, the above-mentioned vital signs monitor presents early-warning scores in small text adjacent to each corresponding vital sign observation. However, the system also provides staff with automated messages to signal what action to take when scores reach a particular threshold (Bellomo et al., 2012). Arguably, the score becomes unnecessary in this case and may actually increase the visual clutter on the monitor. Future investigations are necessary to assess which paper-based chart design features do and do not successfully translate to computerised systems.

Also pertinent to this question is the role and effectiveness of automation; that is, when mechanical or computer components assume the tasks that were otherwise performed by a user (Dekker, 2011; Wickens & Hollands, 2000; Wickens et al., 2004). For example, the previously mentioned electronic vital signs monitoring system: (a) transfers and displays patient vital sign data electronically via a direct physical link with monitoring devices; (b) uses this data to calculate early-warning scores (displayed using colour densities that aim to correlate with the level of severity); and (c) alerts users to necessary actions based on the early-warning score (e.g., to increase the frequency of observations). The system also reminds users of when to measure vital signs, stores data for review, and displays vital sign trends on request (Bellomo et al., 2012). Although automated systems have improved the accuracy and efficiency of tasks across various industries, many unanticipated issues have arisen (Salvendy, 1997). Because these systems still involve human users, many of their shortcomings are grounded in the limitations of attention, perception and cognition (Wickens et al., 2004).

First, automated systems can be more complicated than their manual counterparts. Because of their complex algorithms, automated systems may complete tasks in very different ways to human users. If a system's logic is poorly understood, users can sometimes perceive the system to be acting incorrectly (Salvendy, 1997; Wickens & Hollands, 2000; Wickens et al., 2004), especially if they are busy and distracted with other tasks (Dekker, 2011). These 'automation induced surprises' can become problematic if the user assumes that the system has failed and inappropriately intervenes (Wickens & Hollands, 2000). For example, health professionals (particularly those without exposure to paper-based early-warning systems) may not fully understand the way in which an automated vital sign monitor translates a patient's physiological data into an early-warning score. If users then fail to attend to the patient appropriately (e.g., by deciding not to phone the on-call clinician) because they perceive the automated message as incorrect based on the monitor's early-warning score, then the patient is may deteriorate further.

Conversely, automated systems can also be overly trusted. If users perceive a system as being highly reliable, they can become complacent and neglect to monitor its operation (Wickens &

Hollands, 2000). If the automated system then fails, complacent users will be slower to detect the failure and subsequently less likely to respond appropriately (Wickens & Hollands, 2000; Wickens et al., 2004). This is a critical possibility for automated vital sign monitors. Unlike paper charts, hardware can break, software can crash, and electricity can disconnect. If health professionals become too complacent with the automated system, they may fail to notice that their patients have been unmonitored for hours. If information communication technology fails in hospitals, then staff may have to return to using paper-based charts without notice. This is not a hypothetical possibility. In 2002, the network at Boston's Beth Israel Deaconess Medical Centre repeatedly crashed. Over a four-day period, hospital staff had to revert to paper-based systems (e.g., medical records, prescription forms, lab request forms) that had not been used for years. Critically, many newer members of staff (e.g., interns) had no prior experience with the paper forms (Berinato, 2003). Similarly, in 2015, staff at London's Hillingdon and Mount Vernon hospitals had to transfer to paper-based manual processes for several days following a problem with the network infrastructure (Flinders, 2015).

Automated systems may also lead to the de-skilling of staff. When an automated system assumes responsibility for a task, users' skills can gradually degrade if those skills are not used. Over time, de-skilling can increase users' reliance on the automation as well as the likelihood that users will inappropriately respond to a failing system (Wickens & Hollands, 2000; Wickens et al., 2004). De-skilling as a result of the introduction of automated vital sign monitoring is a real possibility. In the above scenarios, where vital sign monitors fail, health professionals may be ill equipped to monitor patients using traditional manual techniques. For example, manual blood pressure measurements are highly dependent on correct user handling (Tholl, Forstner, & Anlauf, 2004). Staff may also end up lacking practice at documenting vital sign observations by hand and calculating early-warning scores. There are reports that nurses have already expressed fears about de-skilling after the introduction of paper-based early-warning systems (Elliott et al., 2014). Following the implementation of a suite of track-and-trigger charts (modified versions of the ADDS chart), experienced nursing staff were reportedly concerned that the new systems would de-skill staff and replace clinical judgment (Elliott et al., 2014). While this does not necessarily mean that these fears are justified in this context, it does suggest that the possibility needs to be considered. Future research could explore ways to address these potential issues. Subsequent studies examining computerised vital sign monitoring systems could assess which functions would be better allocated to the user and which functions would be better allocated to the automation, based on the relative capabilities of each (Wickens et al., 2004).

The transition to computer-based vital sign monitoring marks a significant juncture in the effort to improve the detection of deteriorating hospitalised patients. Instead of design being guided

by clinicians' opinions (as paper-based chart design has been for decades), computerised systems could be developed using a structured human factors approach from the very beginning of the design process. As discussed earlier, the design of the ADDS chart is already associated with an 11% reduction in mortality amongst intensive care unit admissions (Joshi et al., 2014) as well as a 45% reduction in the incidence of cardiac arrests (Drower et al., 2013). These findings suggest that a human factors approach to design and iterative empirical evaluations can significantly improve health professionals' clinical monitoring of hospitalised patients and their recognition of physiological deterioration. Despite these promising possibilities, we need to acknowledge that chart design, whether paper or electronic, only represents one piece of the puzzle, because a comprehensive human factors approach encompasses a user's interaction with all components of a system (Salvendy, 1997). This suggests that future research should also consider the design of the physical equipment with which health professionals work, the nature of each task that they do, the environment that surrounds them, and the training that they receive (Wickens et al., 2004). Interactions between these components are likely and need to be understood.

Investigated human factors principles

Although this thesis represents the first systematic examination of observation chart design features, our experiments are primarily centered on a single design, which again limits the generalisability of our findings. Although a number of different approaches could have been used to resolve contentious hospital observation chart design decisions, we chose to center our investigations on the tool that represented best practice. At the inception of this project, the ADDS chart was the most empirically supported chart design reported in the literature (Preece et al., 2012b). In addition, the general observation chart only represents one type of medical chart. Attention also needs to be paid to mission-critical charts that have been identified as contributing to adverse events. For instance, in 2011, we sought to improve the design of hospital insulin charts given the potential for patient harm (Christofidis et al., 2012): in a country where the prevalence of diabetes among hospitalised patients is estimated at 24.7% (Bach et al., 2014), and poor glycemic control has been associated with acute cardiovascular events, disability and death (Montori, Bistrian, & McMahon, 2002). Future research is needed to identify which critical medical charts need immediate review using human factors principles and empirical assessment with behavioural and clinical studies.

We also acknowledge that several design principles used in the development of the ADDS chart were not investigated. The previously discussed principles of limiting data-driven tasks, minimising users' cognitive load, and displaying information that will be used together close

together (Gerhardt-Powals, 1996; Nielsen, 1993) only represent a sub-set of those outlined in Chapter 1. Still unknown are the specific effects of: (a) displaying relationships between interface elements (e.g., thicker horizontal lines between adjoining vital signs, ruled off-date rows, thicker vertical lines after every three time columns); (b) constraining the use of colour (to minimise visual clutter); (c) maintaining consistency (e.g., using the same formatting for related labels); (d) speaking users' language (e.g., using common abbreviations and terminology); and (e) displaying information to match users' tasks (Nielsen, 1993). For example, Preece and colleagues applied the principle of displaying information to match users' tasks by positioning the ADDS chart instructions towards the top of the outside front page so that they are immediately available when a user first looks at the chart. However, in our experiments, participants watched a training video that explained how to use the chart, precluding the need to read the instructions. Further, although this principle led the ADDS chart designers to order the vital signs according to their importance, the potential effects of this ordering by priority were not captured by the experimental task.

It is also important to consider that, in terms of human factors design principles, the design of the ADDS chart is influenced largely by the work of Gerhardt-Powals (1996), Nielsen (1993) and Zhu et al. (2005). Although the design principles expounded by these authors are well established in the wider literature, it is possible that alternative principles could be better applied to the ADDS chart design. Indeed, in Chapter 5, we employed a principle that had not been considered in the design of the ADDS chart (i.e., to minimise information access cost) to rationalise the performance benefits associated with excluding scoring-rows (Wickens et al., 2004). Given that the application of the above human factors design principles may not be optimising the ADDS chart's usability, it is critical that future studies explore their effect. We also recommend that future researchers apply and test validated design principles from other domains (e.g., aviation, military and other areas of health care) to develop a stronger evidence-base for the recent human factors approach to chart design, especially those design principles that can be implemented in multiple ways or those that conflict with other principles.

Conclusion

The novel ADDS chart, designed using human factors design principles, supports chart-users' detection of patient deterioration. Despite clinicians' arguments that specific aspects of the design cannot be regarded as best practice, we found that: (1) even health professionals experienced with alternative chart designs can perform better with the ADDS chart; (2) blood pressure and heart rate are better presented as plots that are separated (even for health professionals who prefer overlapping graphs); and (3) chart-users' performance with drawn-dot observations, an integrated colour-based

scoring-system, and grouped scoring-rows corresponds to apriori predictions based on human factors design principles. Although the ADDS chart's use of individual vital sign scoring-rows was not supported, this finding does demonstrate that behavioural experiments should inform best design practice, rather expert opinion. Despite continuous innovations in the health care industry, too often there is a gap between evidence and practice (Grol & Grimshaw, 2003). Therefore, we would argue that it is critical to patient safety that individual hospitals and health services only implement observation chart designs that are supported by empirical evidence.

References

ACSQHC. (2008). Recognising and responding to clinical deterioration: Background paper. Sydney, NSW: Australian Commission on Safety and Quality in Health Care.

ACSQHC. (2009). Recognising and responding to clinical deterioration: Use of observation charts to identify clinical deterioration. Sydney, NSW: Australian Commission on Safety and Quality in Health Care.

ACSQHC. (2010). National consensus statement: essential elements for recognising and responding to clinical deterioration. Sydney, NSW: Australian Commission on Safety and Quality in Health Care.

Araújo, D., Davids, K., & Passos, P. (2007). Ecological Validity, Representative Design, and Correspondence Between Experimental Task Constraints and Behavioral Setting: Comment on Rogers, Kadar, and Costall (2005). *Ecological Psychology, 19*(1), 69-78.

Aschenbrenner, D. S., & Venable, S. J. (2009). *Drug Therapy in Nursing*: Wolters Kluwer Health/Lippincott Williams & Wilkins.

Bach, L. A., Ekinci, E. I., Engler, D., Gilfillan, C., Hamblin, P. S., MacIsaac, R. J., . . . Wyatt, S. (2014). The high burden of inpatient diabetes mellitus: the Melbourne Public Hospitals Diabetes Inpatient Audit. *Med J Aust, 201*(6), 334-338.

Bellomo, R., Ackerman, M., Bailey, M., Beale, R., Clancy, G., Danesh, V., . . . Tangkau, P. (2012). A controlled trial of electronic automated advisory vital signs monitoring in general hospital wards. *Crit Care Med, 40*(8), 2349-2361. doi: 10.1097/CCM.0b013e318255d9a0

Berinato, S. (2003). All systems down. Retrieved September 11, 2015, from http://www.computerworld.com/article/2581420/disaster-recovery/all-systems-down.html

Boulanger, C., & Toghill, M. (2009). How to measure and record vital signs to ensure detection of deteriorating patients. *Nurs Times, 105*(47), 10-12.

Bright, D., Walker, W., & Bion, J. (2004). Clinical review: Outreach - a strategy for improving the care of the acutely ill hospitalized patient. *Crit Care, 8*(1), 33-40. doi: 10.1186/cc2377

Brunstein, A., Betts, S., & Anderson, J. R. (2009). Practice enables successful learning under minimal guidance. *Journal of Educational Psychology, 101*(4), 790-802. doi: 10.1037/a0016656

Buist, M., Bernard, S., Nguyen, T. V., Moore, G., & Anderson, J. (2004). Association between clinically abnormal observations and subsequent in-hospital mortality: a prospective study. *Resuscitation, 62*(2), 137-141. doi: 10.1016/j.resuscitation.2004.03.005

Campbell, J. I. D. (1992). *The Nature and Origin of Mathematical Skills*: Elsevier Science.

Carayon, P. (2012). Human Factors and Ergonomics in Health Care and Patient Safety. In P. Carayon (Ed.), *Handbook of Human Factors and Ergonomics in Health Care and Patient Safety* (Second Edition ed., pp. 3-15). Boca Raton FL: CRC Press Taylor & Francis Group.

Carliner, S., Verckens, J. P., & de Waele, C. (2006). *Information and Document Design: Varieties on Recent Research*: John Benjamins Publishing Company.

Chatterjee, M. T., Moon, J. C., Murphy, R., & McCrea, D. (2005). The "OBS" chart: an evidence based approach to re-design of the patient observation chart in a district general hospital setting. *Postgrad Med J, 81*(960), 663-666. doi: 10.1136/pgmj.2004.031872

Choudhry, N. K., Fletcher, R. H., & Soumerai, S. B. (2005). Systematic review: the relationship between clinical experience and quality of health care. *Ann Intern Med, 142*(4), 260-273.

Christofidis, M. J., Hill, A., Horswill, M. S., & Watson, M. O. (2013). A human factors approach to observation chart design can trump health professionals' prior chart experience. *Resuscitation, 84*(5), 657-665. doi: 10.1016/j.resuscitation.2012.09.023

Christofidis, M. J., Hill, A., Horswill, M. S., & Watson, M. O. (2014). Observation charts with overlapping blood pressure and heart rate graphs do not yield the performance advantage that health professionals assume: an experimental study. *Journal of Advanced Nursing, 70*(3), 610-624. doi: 10.1111/jan.12223

Christofidis, M. J., Hill, A., Horswill, M. S., & Watson, M. O. (2015). Less is more: the design of early-warning scoring systems affects the speed and accuracy of scoring. *Journal of Advanced Nursing, 71*(7), 1573-1586. doi: 10.1111/jan.12618

Christofidis, M. J., Hill, A., Horswill, M. S., & Watson, M. O. (in press). Observation chart design features affect the detection of patient deterioration: A systematic experimental evaluation. *Journal of Advanced Nursing*.

Christofidis, M. J., Horswill, M. S., Hill, A., McKimmie, B. M., Visser, T., & Watson, M. O. (2012). Task Analysis and Heuristic Analysis of Insulin Charts: Final Report prepared for the Australian Commission on Safety and Quality in Health Care: School of Psychology, The University of Queensland.

Cioffi, J., Salter, C., Wilkes, L., Vonu-Boriceanu, O., & Scott, J. (2006). Clinicians' responses to abnormal vital signs in an emergency department. *Aust Crit Care, 19*(2), 66-72.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*: L. Erlbaum Associates.

Czaja, S. J., & Sharit, J. (2012). *Designing Training and Instructional Programs for Older Adults*: Taylor & Francis.

Darby, B., Mitchell, I., Van Leuvan, C., Kingbury, A., & McKay, H. (2012). *Seagulls could save lives.* Paper presented at the Official 2012 Program of the 7th International Conference on Rapid Response Systems and Medical Emergency Teams, Sydney, NSW.

Dekker, S. (2011). *Patient Safety: A Human Factors Approach*: Taylor & Francis.

Drews, F. A., & Kramer, H. S. (2012). Human–Computer Interaction Design in Health Care. In P. Carayon (Ed.), *Handbook of Human Factors and Ergonomics in Health Care and Patient Safety* (Second Edition ed., pp. 265-279). Boca Raton FL: CRC Press Taylor & Francis Group.

Drower, D., Mckeany, R., Jogia, P., & Jull, A. (2013). Evaluating the impact of implementing an early warning score system on incidence of in-hospital cardiac arrest. *NZ Med J, 126*, 26-34.

Drury, C. G., Sarac, A., & Driscoll, D. M. (1997). Documentation design aid development *Human Factors in Aviation Maintenance--Phase VII: Progress Report* (pp. 75-107). Washington, DC: Federal Aviation Administration/Office of Aviation Medicine.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12*(3), 83-87. doi: 10.1111/1467-8721.01235

Eastwood, K. J., Boyle, M. J., Williams, B., & Fairhall, R. (2011). Numeracy skills of nursing students. *Nurse Educ Today, 31*(8), 815-818. doi: 10.1016/j.nedt.2010.12.014

Elliott, D., Allen, E., Perry, L., Fry, M., Duffield, C., Gallagher, R., . . . Roche, M. (2014). Clinical user experiences of observation and response charts: focus group findings of using a new format chart incorporating a track and trigger system. *BMJ Qual Saf*. doi: 10.1136/bmjqs-2013-002777

Elliott, D., Allen, E., Perry, L., Fry, M., Duffield, C., Gallagher, R., . . . Roche, M. (2015). Clinical user experiences of observation and response charts: focus group findings of using a new format chart incorporating a track and trigger system. *BMJ Qual Saf, 24*(1), 65-75. doi: 10.1136/bmjqs-2013-002777

Endacott, R., Kidd, T., Chaboyer, W., & Edington, J. (2007). Recognition and communication of patient deterioration in a regional hospital: a multi-methods study. *Aust Crit Care, 20*(3), 100-105. doi: 10.1016/j.aucc.2007.05.002

Ericsson, K. A., & Ward, P. (2007). Capturing the naturally occurring superior performance of experts in the laboratory: Toward a science of expert and exceptional performance. *Current Directions in Psychological Science, 16*(6), 346-350. doi: 10.1111/j.1467-8721.2007.00533.x

FAA. (2008). *Instrument Flying Handbook*: Skyhorse Pub.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods, 39*(2), 175-191.

Fitzpatrick, J. J., & Kazer, M. (2011). *Encyclopedia of Nursing Research, Third Edition*: Springer Publishing Company.

Flinders, K. (2015). Network failure crashed frontline services at London hospital. Retrieved September 11, 2015, from http://www.computerweekly.com/news/4500247512/Network-failure-crashed-frontline-services-at-London-hospital

Franklin, C., & Mathew, J. (1994). Developing strategies to prevent inhospital cardiac arrest: analyzing responses of physicians and nurses in the hours before the event. *Crit Care Med, 22*(2), 244-247.

Ganier, F. (2004). Factors affecting the processing of procedural instructions: implications for document design. *Professional Communication, IEEE Transactions on, 47*(1), 15-26.

Gao, H., McDonnell, A., Harrison, D. A., Moore, T., Adam, S., Daly, K., . . . Harvey, S. (2007). Systematic review and evaluation of physiological track and trigger warning systems for identifying at-risk patients on the ward. *Intensive Care Med, 33*(4), 667-679. doi: 10.1007/s00134-007-0532-3

Gerhardt-Powals, J. (1996). Cognitive engineering principles for enhancing human-computer performance. *International Journal of Human-Computer Interaction, 8*, 189-211.

Gillan, D. J., & Schvaneveldt, R. W. (1999). Applying cognitive psychology: Bridging the gulf between basic research and cognitive artifacts. In F. T. Durso, R. Nickerson, R. Schvaneveldt, S. Dumais, M. Chi & S. Lindsay (Eds.), *The handbook of applied cognition* (pp. 3-31). Chichester, England: Wiley.

Goldhill, D. R., White, S. A., & Sumner, A. (1999a). Physiological values and procedures in the 24 h before ICU admission from the ward. *Anaesthesia, 54*(6), 529-534.

Goldhill, D. R., Worthington, L., Mulcahy, A., Tarling, M., & Sumner, A. (1999b). The patient-at-risk team: identifying and managing seriously ill ward patients. *Anaesthesia, 54*(9), 853-860.

Grol, R., & Grimshaw, J. (2003). From best evidence to best practice: effective implementation of change in patients' care. *Lancet, 362*(9391), 1225-1230. doi: 10.1016/s0140-6736(03)14546-1

Hammond, K. R. (1998). Ecological validity: Then and now. Retrieved June 5, 2016, from http://www.albany.edu/cpr/brunswik/notes/essay2.html

Hammond, N. E., Spooner, A. J., Barnett, A. G., Corley, A., Brown, P., & Fraser, J. F. (2013). The effect of implementing a modified early warning scoring (MEWS) system on the adequacy of vital sign documentation. *Aust Crit Care, 26*(1), 18-22. doi: 10.1016/j.aucc.2012.05.001

Harrison, G. A., Jacques, T. C., Kilborn, G., & McLaws, M. L. (2005). The prevalence of recordings of the signs of critical conditions and emergency responses in hospital wards - the SOCCER study. *Resuscitation, 65*(2), 149-157. doi: 10.1016/j.resuscitation.2004.11.017

Hillman, K. M., Bristow, P. J., Chey, T., Daffurn, K., Jacques, T., Norman, S. L., . . . Simmons, G. (2001). Antecedents to hospital deaths. *Intern Med J, 31*(6), 343-348.

Hoeken, H., & Korzilius, H. (2003). Conducting experiments on cultural aspects of document design: Why and how? *Communications, 28*, 285-304.

Horswill, M. S., Preece, M. H. W., Hill, A., Christofidis, M. J., & Watson, M. O. (2010). Human factors research regarding observation charts: Research project overview: Report prepared for the Australian Commission on Safety and Quality in Health Care's Program for Recognizing and Responding to Clinical Deterioration.

Jacques, T., Harrison, G. A., McLaws, M. L., & Kilborn, G. (2006). Signs of critical conditions and emergency responses (SOCCER): A model for predicting adverse events in the inpatient setting. *Resuscitation, 69*(2), 175-183. doi: 10.1016/j.resuscitation.2005.08.015

Johnstone, C. C., Rattray, J., & Myers, L. (2007). Physiological risk factors, early warning scoring systems and organizational changes. *Nurs Crit Care, 12*(5), 219-224. doi: 10.1111/j.1478-5153.2007.00238.x

Jones, S., Mullally, M., Ingleby, S., Buist, M., Bailey, M., & Eddleston, J. M. (2011). Bedside electronic capture of clinical observations and automated clinical alerts to improve compliance with an Early Warning Score protocol. *Crit Care Resusc, 13*(2), 83-88.

Joshi, K., Landy, M., Anstey, C., Gooch, R., & Campbell, V. (2014). *Effect on admission severity of illness and ICU outcomes using the Adult Deterioration Detection System for MET call activation*. Queensland Government.

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review, 19*(4), 509-539. doi: 10.1007/s10648-007-9054-3

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38*(1), 23-31. doi: 10.1207/S15326985EP3801_4

Kalyuga, S., Chandler, P., & Sweller, J. (1998). Levels of expertise and instructional design. *Human Factors, 40*(1), 1.

Kalyuga, S., Chandler, P., Tuovinen, J., & Sweller, J. (2001). When problem solving is superior to studying worked examples. *Journal of Educational Psychology, 93*(3), 579-588. doi: 10.1037/0022-0663.93.3.579

Kalyuga, S., & Renkl, A. (2010). Expertise reversal effect and its instructional implications: Introduction to the special issue. *Instructional Science, 38*(3), 209-215. doi: 10.1007/s11251-009-9102-0

Kalyuga, S., & Sweller, J. (2004). Measuring knowledge to optimize cognitive load factors during instruction. *Journal of Educational Psychology, 96*(3), 558-568. doi: 10.1037/0022-0663.96.3.558

Kansal, A., & Havill, K. (2012). The effects of introduction of new observation charts and calling criteria on call characteristics and outcome of hospitalised patients. *Crit Care Resusc, 14*(1), 38-43.

Karwowski, W. (2006). *International Encyclopedia of Ergonomics and Human Factors, Second Edition - 3 Volume Set*: Taylor & Francis.

Kause, J., Smith, G., Prytherch, D., Parr, M., Flabouris, A., & Hillman, K. (2004). A comparison of antecedents to cardiac arrests, deaths and emergency intensive care admissions in Australia and New Zealand, and the United Kingdom - the ACADEMIA study. *Resuscitation, 62*(3), 275-282. doi: 10.1016/j.resuscitation.2004.05.016

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational psychologist, 41*(2), 75-86.

Knight, S., Calvesbert, K., Clarke, J., & Williamson, J. (2002). Developing a nursing observation chart. *Emerg Nurse, 10*(3), 16-17. doi: 10.7748/en2002.06.10.3.16.c1062

Koutoukidis, G., Stainton, K., & Hughson, J. (2012). *Tabbner's Nursing Care: Theory and Practice*: Elsevier Health Sciences.

Kyun, S., Kalyuga, S., & Sweller, J. (2013). The effect of worked examples when learning to write essays in English literature. *Journal of Experimental Education, 81*(3), 385-408. doi: 10.1080/00220973.2012.727884

Lee, J. D., Kirlik, A., & Dainoff, J. (2013). *The Oxford Handbook of Cognitive Engineering*: OUP USA.

Lentz, L., & Pander Maat, H. (2004). Functional Analysis for Document Design. *Technical communication, 51*(3), 387-398.

Leuvan, C. H., & Mitchell, I. (2008). Missed opportunities? An observational study of vital sign measurements. *Crit Care Resusc, 10*(2), 111-115.

Lockwood, C., Conroy-Hiller, T., & Page, T. (2004). Vital signs. *JBI Reports, 2*(6), 207-230.

McBride, M., & Burgman, M. (2012). What Is Expert Knowledge, How Is Such Knowledge Gathered, and How Do We Use It to Address Questions in Landscape Ecology? In A. H.

Perera, C. A. Drew & C. J. Johnson (Eds.), *Expert Knowledge and Its Application in Landscape Ecology* (pp. 11-38). New York: Springer

Mitchell, I. A. (2012). *Putting the chart into perspective*. Paper presented at the 7th International Conference on Rapid Response Systems and Medical Emergency Teams, Sydney, New South Wales, Australia.

Mitchell, I. A., McKay, H., Van Leuvan, C., Berry, R., McCutcheon, C., Avard, B., . . . Lamberth, P. (2010). A prospective controlled trial of the effect of a multi-faceted intervention on early recognition and intervention in deteriorating hospital patients. *Resuscitation, 81*(6), 658-666. doi: 10.1016/j.resuscitation.2010.03.001

Mohammed, M., Hayton, R., Clements, G., Smith, G., & Prytherch, D. (2009). Improving accuracy and efficiency of early warning scores in acute care. *Br J Nurs, 18*(1), 18-24.

Montori, V. M., Bistrian, B. R., & McMahon, M. (2002). Hyperglycemia in acutely ill patients. *JAMA, 288*(17), 2167-2169. doi: 10.1001/jama.288.17.2167

Mumpower, J. L., & Stewart, T. R. (1996). Expert judgement and expert disagreement. *Thinking & Reasoning, 2*(2-3), 191-212. doi: 10.1080/135467896394500

NICE. (2007a). Acutely ill patients in hospital: Recognition of and response to acute illness in adults in hospital. *National Institute for Health and Clinical Excellence: Guidance*. London.

NICE. (2007b). Recognising and responding appropriately to early signs of deterioration in hospitalised patients. *National Institute for Health and Clinical Excellence: Guidance*. London.

Nielsen, J. (1993). *Usability Engineering*. Cambridge: AP Professional.

Nielsen, J., & Mack, R. L. (1994). *Usability Inspection Methods*. Michigan: Wiley.

Norman, D. (1992). Design principles for cognitive artifacts. *Research in Engineering Design, 4*(1), 43-50. doi: 10.1007/BF02032391

Nückles, M., Hübner, S., Dümer, S., & Renkl, A. (2010). Expertise reversal effects in writing-to-learn. *Instructional Science, 38*(3), 237-258. doi: 10.1007/s11251-009-9106-9

Nwulu, U., Westwood, D., Edwards, D., Kelliher, F., & Coleman, J. J. (2012). Adoption of an electronic observation chart with an integrated early warning scoring system on pilot wards: a descriptive report. *Comput Inform Nurs, 30*(7), 371-379. doi: 10.1097/NXN.0b013e318251074a

Odell, M., Victor, C., & Oliver, D. (2009). Nurses' role in detecting deterioration in ward patients: systematic literature review. *J Adv Nurs, 65*(10), 1992-2006.

Oksa, A., Kalyuga, S., & Chandler, P. (2010). Expertise reversal effect in using explanatory notes for readers of Shakespearean text. *Instructional Science, 38*(3), 217-236. doi: 10.1007/s11251-009-9109-6

Patel, S., Drury, C. G., & Lofgren, J. (1994). Design of workcards for aircraft inspection. *Applied Ergonomics, 25*(5), 283-293. doi: http://dx.doi.org/10.1016/0003-6870(94)90042-6

Phua, D. H., & Tan, N. C. (2013). Cognitive aspect of diagnostic errors. *Ann Acad Med Singapore, 42*(1), 33-41.

Preece, M. H., Hill, A., Horswill, M. S., Dunbar, N., Adams, J. L., Christofidis, M. J., & Watson, M. O. (2010a). Developer's Guide for Observation and Response Charts: Report prepared for the Australian Commission on Safety and Quality in Health Care's Program for Recognising and Responding to Clinical Deterioration.

Preece, M. H., Hill, A., Horswill, M. S., Karamatic, R., Hewett, D. G., & Watson, M. O. (2013). Applying heuristic evaluation to observation chart design to improve the detection of patient deterioration. *Appl Ergon, 44*(4), 544-556. doi: 10.1016/j.apergo.2012.11.003

Preece, M. H., Hill, A., Horswill, M. S., Karamatic, R., & Watson, M. O. (2012a). Designing observation charts to optimize the detection of patient deterioriation: Reliance on the subjective preferences of healthcare professionals is not enough. *Aust Crit Care, 25*(4), 238-252. doi: 10.1016/j.aucc.2012.01.003

Preece, M. H., Hill, A., Horswill, M. S., & Watson, M. O. (2012b). Supporting the detection of patient deterioration: Observation chart design affects the recognition of abnormal vital signs. *Resuscitation, 83*(9), 1111-1118. doi: 10.1016/j.resuscitation.2012.02.009

Preece, M. H., Horswill, M. S., Hill, A., Karamatic, R., & Watson, M. O. (2010b). An online survey of health professionals' opinions regarding observation chart: Report prepared for the Australian Commission on Safety and Quality in Health Care's Program for Recognising and Responding to Clinical Deterioration.

Preece, M. H., Horswill, M. S., Hill, A., & Watson, M. O. (2010c). The Development of the Adult Deterioration Detection System (ADDS) Chart: Report prepared for the Australian Commission on Safety and Quality in Health Care's program for Recognising and Responding to Clinical Deterioration.

Proctor, R. W., & Van Zandt, T. (2008). *Human Factors in Simple and Complex Systems, Second Edition*: Taylor & Francis.

Proctor, R. W., & Vu, K. P. L. (2006). *Stimulus-Response Compatibility Principles: Data, Theory, and Application*: Taylor & Francis.

Prytherch, D. R., Smith, G. B., Schmidt, P., Featherstone, P. I., Stewart, K., Knight, D., & Higgins, B. (2006). Calculating early warning scores - a classroom comparison of pen and paper and hand-held computer methods. *Resuscitation, 70*(2), 173-178. doi: 10.1016/j.resuscitation.2005.12.002

Rashid, U., Nacenta, M. A., & Quigley, A. (2012). *The cost of display switching: a comparison of mobile, large display and hybrid UI configurations*. Paper presented at the Proceedings of the International Working Conference on Advanced Visual Interfaces, Capri Island, Italy.

Rebelo, F., & Soares, M. (2014, 19-23 July 2014). *Advances in Ergonomics In Design, Usability & Special Populations.* Paper presented at the 5th AHFE Conference, Krakow, Poland.

Reisner, A. T., Chen, L., & Reifman, J. (2012). The association between vital signs and major hemorrhagic injury is significantly improved after controlling for sources of measurement variability. *J Crit Care, 27*(5), 533.e531-510. doi: 10.1016/j.jcrc.2012.01.006

Reisslein, J., Atkinson, R. K., Seeling, P., & Reisslein, M. (2006). Encountering the expertise reversal effect with a computer-based environment on electrical circuit analysis. *Learning and Instruction, 16*(2), 92-103. doi: 10.1016/j.learninstruc.2006.02.008

Rey, G. D., & Buchwald, F. (2011). The expertise reversal effect: cognitive load and motivational explanations. *J Exp Psychol Appl, 17*(1), 33-48. doi: 10.1037/a0022243

Robb, G., & Seddon, M. (2010). A multi-faceted approach to the physiologically unstable patient. *Qual Saf Health Care, 19*(5), e47. doi: 10.1136/qshc.2008.031807

Rogers, M. L., Patterson, E. S., & Render, M. L. (2012). Cognitive Work Analysis in Health Care. In P. Carayon (Ed.), *Handbook of Human Factors and Ergonomics in Health Care and Patient Safety* (Second Edition ed., pp. 465-474). Boca Raton FL: CRC Press Taylor & Francis Group.

Roller, J. E., Prasad, N. H., Garrison, H. G., & Whitley, T. (1992). Unexpected emergency department death: Incidence, causes, and relationship to presentation and time in the department. *Ann Emerg Med, 21*(6), 743-745.

Salden, R. J. C. M., Aleven, V., Schwonke, R., & Renkl, A. (2010). The expertise reversal effect and worked examples in tutored problem solving. *Instructional Science, 38*(3), 289-307. doi: 10.1007/s11251-009-9107-8

Salkind, N. J. (2010). Encyclopedia of Research Design. Thousand Oaks, CA: SAGE Publications.

Salvendy, G. (1997). *Handbook of Human Factors and Ergonomics*. New Jersey: Wiley.

Schein, R. M., Hazday, N., Pena, M., Ruben, B. H., & Sprung, C. L. (1990). Clinical antecedents to in-hospital cardiopulmonary arrest. *Chest, 98*(6), 1388-1392.

Schnotz, W. (2010). Reanalyzing the expertise reversal effect. *Instructional Science, 38*(3), 315-323. doi: 10.1007/s11251-009-9104-y

Schriver, K. A. (1993). Quality in document design: Issues and controversies. *Technical communication*, 239-257.

Sela, R., & Auerbach-Shpak, Y. (2014). The user-centered design of a radiotherapy chart. In Y. Donchin & D. Gopher (Eds.), *Around the patient bed: human factors and safety in health care*. Boca Raton: CRC Press, Taylor & Francis Group.

Smith, D. L., & Fernhall, B. (2011). *Advanced Cardiovascular Exercise Physiology*. USA: Human Kinetics.

Spyridakis, J. H., & Wenger, M. J. (1992). Writing for Human Performance: Relating Reading Research to Document Design. *Technical communication, 39*(2), 202-215.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*(1), 137-149. doi: 10.3758/bf03207704

Tholl, U., Forstner, K., & Anlauf, M. (2004). Measuring blood pressure: pitfalls and recommendations. *Nephrology Dialysis Transplantation, 19*(4), 766-770. doi: 10.1093/ndt/gfg602

Thomas, J. C., & Schneider, M. L. (1984). *Human Factors in Computer Systems*. USA: Ablex Publishing Corporation.

Tuovinen, J. E., & Sweller, J. (1999). A comparison of cognitive load associated with discovery learning and worked examples. *Journal of Educational Psychology, 91*(2), 334-341. doi: 10.1037/0022-0663.91.2.334

van Gog, T., Ericsson, K. A., Rikers, R. M. J. P., & Paas, F. (2005). Instructional design for advanced learners: Establishing connections between the theoretical frameworks of cognitive load and deliberate practice. *Educational Technology Research and Development, 53*(3), 73-81. doi: 10.1007/BF02504799

Wickens, C. D., & Carswell, C. M. (1995). The proximity compatibility principle: Its psychological foundation and relevance to display design. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 37*(3), 473-494. doi: 10.1518/001872095779049408

Wickens, C. D., & Hollands, J. G. (2000). *Engineering Psychology and Human Performance*. New Jersey: Prentice Hall.

Wickens, C. D., Lee, J. D., Liu, Y., & Gordon Becker, S. E. (2004). *An Introduction to Human Factors Engineering*. New Jersey: Pearson Prentice Hall.

Wickens, C. D., & McCarley, J. S. (2007). *Applied Attention Theory*. Florida: Taylor & Francis.

Wood, J., & Finkelstein, J. (2013). Comparison of automated and manual vital sign collection at hospital wards. *Stud Health Technol Inform, 190*, 48-50.

Zeitz, K., & McCutcheon, H. (2006). Observations and vital signs: ritual or vital for the monitoring of postoperative patients? *Appl Nurs Res, 19*(4), 204-211. doi: 10.1016/j.apnr.2005.09.005

Zheng, N., & Xue, J. (2009). *Statistical Learning and Pattern Analysis for Image and Video Processing*. London: Springer.

Zhu, W., Vu, K. P. L., & Proctor, R. W. (2005). Evaluating web usability. In R. W. Proctor & K. P. L. Vu (Eds.), *Handbook of Human Factors in Web Design* (pp. 321-337). New Jersey: Lawrence Erlbaum Associates.