Less is more: the design of early-warning scoring systems affects the speed and accuracy of scoring

Melany J. Christofidis, Andrew Hill, Mark S. Horswill & Marcus O. Watson

Abstract
Aim. To evaluate the effect of early-warning scoring system design on the speed and accuracy of scoring.

Background. Despite the widespread implementation of early-warning scoring systems in hospitals, the speed and accuracy with which chart-users determine patients’ early-warning scores has received minimal research attention.

Design. Within-subjects, with scoring-system design as the independent variable.

Methods. Forty-seven novice chart-users were presented with realistic vital sign observations recorded on charts with three different scoring-system designs. The rows for recording individual vital sign scores were either: (1) grouped together beneath all of the vital sign rows; (2) separated, with each row presented immediately below the corresponding vital sign row; or (3) excluded altogether. Participants’ response times and error rates for determining the overall scores were measured for 54 time-points per design. Data were collected in December 2012–January 2013.

Results. Contrary to predictions, participants responded fastest and made the fewest errors when using the chart design without individual vital sign scoring-rows. For the other two designs, participants were faster when the rows for scoring individual vital signs were separated (vs. grouped), but accuracy did not differ. For both of these designs, significantly more time-points were affected by scoring errors compared with adding errors. Finally, data for patients with more serious derangements yielded greater response times and error rates on all three charts.

Conclusion. Early-warning scoring systems may be more effective without individual vital sign scoring-rows. Even when charts are designed by multi-disciplinary teams of human factors specialists and clinicians, empirical evaluations are essential.

Keywords: design, deterioration, early-warning scoring systems, human factors, nursing, observation chart
Why is this research needed?

- Paper-based observation charts used in hospitals commonly incorporate early-warning scoring systems to aid nurses and doctors in the early detection of deteriorating patients.
- However, errors in determining patients’ early-warning scores, which can affect the appropriateness of the clinical response, occur frequently.
- Although recent studies have demonstrated that improvements to chart design can assist chart-users to detect abnormal observations more quickly and accurately, no published study has assessed the impact of design on the determination of early-warning scores.

What are the key findings?

- Preventing chart-users from recording individual vital sign scores yielded more efficient and accurate determination of overall scores: participants responded fastest and made the fewest errors when using the chart design without individual vital sign scoring-rows.
- Although the two designs with individual vital sign scoring-rows did not differ in the frequency of either scoring errors (i.e. determining individual vital sign scores) or adding errors (i.e. summing individual scores to determine the overall score), participants determined the overall scores faster using separate, rather than grouped, rows.
- Across all chart designs, response times and error rates positively correlated with ‘target’ early-warning scores (i.e. the magnitude of the correct total score), indicating that the more at risk the patient, the slower and more inaccurate responses were likely to be.

How should the findings be used to influence policy/practice/research/education?

- Contrary to our predictions, the findings 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 the clinical responses of both nurses and doctors.
- Our results also demonstrate that iterative empirical evaluations are essential as even multi-disciplinary teams of clinicians and human factors specialists can make sub-optimal design choices.

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. 2006, 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. 2006, Gao et al. 2007, Subbe et al. 2007, Smith et al. 2008, Australian Commission on Safety & Quality in Health Care (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 summarises the patient’s overall physiological condition (Prytherch et al. 2006, Mohammed et al. 2009). As well as assisting health professionals to recognise 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. 2006, 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. 2006, Smith et al. 2008, Mohammed et al. 2009). Past studies have established that errors occur frequently, both via simulations (Prytherch et al. 2006, 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. 2006, 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. 2006). 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. 2006).

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. 2006, 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 backup 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, 2013, 2014, Preece et al. 2012b). However, no published study has assessed the impact of chart design on the determination of early-warning scores.

The study

Aims

This 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. 2010b, 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. 2010b, 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. 2010b, 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. 2010a, 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. 2010b,
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 hospitals (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. 2010b, 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.

**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, 4 hour urine output, consciousness and pain. Each case contained two sets of observations that would yield each overall early-warning score from 0–8 if scored and added correctly (i.e. across the nine cases, each of these ‘target’ scores occurred 18 times). This range was chosen to maximise content validity by reflecting the clinically relevant values prescribed by the ADSS chart (Horswill et al. 2010b, 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 ecological validity, 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 in 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. 2010b, ACSQHC 2013).

**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. 2010b, 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 this study until the final sample exceeded this number. No participants were excluded from the analyses (Figure 2).

**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. 2010b, 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 ecological validity. 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 one experimental trial and each participant completed 18 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 Research Ethics Committee 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 $\eta^2$ calculated as the measure of

---

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 eight individual vital signs are scored (Horswill et al. 2010b, Australian Commission on Safety & Quality in Health Care 2013).

<table>
<thead>
<tr>
<th>‘Target’ overall early-warning score</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 non-zero digit</td>
<td>–</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>2 non-zero digits</td>
<td>–</td>
<td>–</td>
<td>11</td>
<td>21</td>
<td>31</td>
<td>41</td>
<td>51</td>
<td>52</td>
<td>53†</td>
</tr>
<tr>
<td>3 non-zero digits</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>111</td>
<td>211</td>
<td>211</td>
<td>311</td>
<td>321</td>
<td>511</td>
</tr>
<tr>
<td>4 non-zero digits</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>1111</td>
<td>2111</td>
<td>2111</td>
<td>3111</td>
<td>3211</td>
</tr>
<tr>
<td>5 non-zero digits</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>11111</td>
<td>21111</td>
<td>21111</td>
<td>22111</td>
</tr>
<tr>
<td>6 non-zero digits</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>111111</td>
<td>211111</td>
<td>211111</td>
</tr>
<tr>
<td>7 non-zero digits</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>1111111</td>
<td>2111111</td>
</tr>
<tr>
<td>8 non-zero digits</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>11111111</td>
</tr>
</tbody>
</table>

* Individual vital sign scores cannot be greater than five (Horswill et al. 2010b, ACSQHC 2013).
† Only systolic blood pressure can yield individual vital sign scores of 4 or 5 (Horswill et al. 2010b, ACSQHC 2013), thus the combination 44 cannot occur within a single time-point.
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 vs. 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 ($t(46) = -2.28,$...
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 ($P = 0.537$).

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.

**Discussion**

The results of this 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. Furthermore, 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. 2010b).

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. 2010b, 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. 2010b, 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 inaccurate 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 this study support it.

---

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

<table>
<thead>
<tr>
<th>Error</th>
<th>Measure</th>
<th>Scoring-system design</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Grouped rows</td>
</tr>
<tr>
<td>Scoring</td>
<td>Percentage of individual vital sign scores affected by scoring errors ($%$)</td>
<td>65% (62)</td>
</tr>
<tr>
<td></td>
<td>Percentage of time-points affected by (one or more) scoring errors ($%$)</td>
<td>16% (99)</td>
</tr>
<tr>
<td>Adding</td>
<td>Percentage of time-points affected by adding errors ($%$)</td>
<td>81% (199)</td>
</tr>
<tr>
<td>Both scoring and adding</td>
<td>Percentage of time-points affected by both scoring and adding errors ($%$)</td>
<td>20% (69)</td>
</tr>
</tbody>
</table>

---

© 2015 John Wiley & Sons Ltd
Figure 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 generalise 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 low-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, this 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 has an impact on 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 healthcare 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, 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 generalised 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 individual 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. 2010b, 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 emphasised that, although a well-designed observation chart can assist even the least experienced chart-user to recognise 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 computerised or paper-based) can be complex (Subbe et al. 2007), there remains enormous scope for further empirical usability research.

Acknowledgements

We thank Lynnette Adams and David Collard for providing an Adobe InDesign CS5.5 copy of the ADDS chart. We also acknowledge staff from The University of Queensland for accommodating and supporting this research project, and the students who volunteered to participate.
Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The study was conducted as part of the first author’s PhD project, which is supported by an Australian Postgraduate Award and a Queensland Government Smart Futures Scholarship.

Conflicts of interest

No conflict of interest has been declared by the authors.

Author contributions

All authors have agreed on the final version and meet at least one of the following criteria [recommended by the ICMJE (http://www.icmje.org/ethical_1author.html)]:

- substantial contributions to conception and design, acquisition of data, or analysis and interpretation of data;
- drafting the article or revising it critically for important intellectual content.

References


The Journal of Advanced Nursing (JAN) is an international, peer-reviewed, scientific journal. JAN contributes to the advancement of evidence-based nursing, midwifery and health care by disseminating high quality research and scholarship of contemporary relevance and with potential to advance knowledge for practice, education, management or policy. JAN publishes research reviews, original research reports and methodological and theoretical papers.

For further information, please visit JAN on the Wiley Online Library website: www.wileyonlinelibrary.com/journal/jan

**Reasons to publish your work in JAN:**

- **High-impact forum:** the world’s most cited nursing journal, with an Impact Factor of 1.527 – ranked 14/101 in the 2012 ISI Journal Citation Reports © (Nursing (Social Science)).
- **Most read nursing journal in the world:** over 3 million articles downloaded online per year and accessible in over 10,000 libraries worldwide (including over 3,500 in developing countries with free or low cost access).
- **Fast and easy online submission:** online submission at http://mc.manuscriptcentral.com/jan.
- **Positive publishing experience:** rapid double-blind peer review with constructive feedback.
- **Rapid online publication in five weeks:** average time from final manuscript arriving in production to online publication.
- **Online Open:** the option to pay to make your article freely and openly accessible to non-subscribers upon publication on Wiley Online Library, as well as the option to deposit the article in your own or your funding agency’s preferred archive (e.g. PubMed).