

Received: 2018.06.29
Accepted: 2018.08.30
Published: 2018.12.15

Comprehensive Evaluation of User Interface for Ventilators Based on Respiratory Therapists' Performance, Workload, and User Experience

Authors' Contribution:
Study Design A
Data Collection B
Statistical Analysis C
Data Interpretation D
Manuscript Preparation E
Literature Search F
Funds Collection G

ABCDEF 1,2 **Mingyin Jiang**
AEFG 1,2 **Shenglin Liu**
BCD 1,2 **Jiaqi Gao**
AF 1,2 **Qingmin Feng**
ADEFG 1,2 **Qiang Zhang**

1 Department of Medical Engineering, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, P.R. China
2 Healthcare Ergonomics Lab, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, P.R. China

Corresponding Author: Qiang Zhang, e-mail: whuhcce@163.com

Source of support: This study was supported by the National Key R&D Program of China (Number: 2016YFC0106702)

Background: Poor ergonomic design of ventilators can result in human errors. In this study, we evaluated the ergonomics of ventilators through respiratory therapists' performance, workload, and user experience.





Material/Methods: Sixteen respiratory therapists were recruited to this usability study of 3 ventilators. Participants had to perform 7 tasks on each ventilator. Respiratory therapists' performance was measured by task errors of all tasks for each participant. Workload was measured by objective measurement (blink rate and duration) and by subjective measurement (NASA-TLX). User experience was assessed by the USE Questionnaire.

Results: For task errors, significant differences were found among ventilators ($p < 0.05$) and the Evital 4 received higher task errors when compared to the Servo I ($p < 0.05$). For blink rate, significant differences were found in tasks of starting the ventilator, ventilator monitoring values recognition, ventilator setting parameters modification, alarm parameter recognition, and resetting among ventilators ($p < 0.05$). Furthermore, blink duration was also found to be significant differently in tasks of starting the ventilator, mode and setting parameters recognition, ventilator monitoring values recognition, ventilator mode modification, and alarm parameter recognition and resetting, as well as in the average of all tasks ($p < 0.05$). For perceived workload, the Evital 4 received higher NASA-TLX scores among ventilators. For user experience, the Servo I received the highest scores on the USE Questionnaire among the ventilators.

Conclusions: The study provides a comprehensive evaluation method of user interface based on respiratory therapists' performance, workload, and user experience. In addition, this study suggests that the ergonomic design of the Evital 4 is poor. Finally, we found that eye motion (blink rate and duration) may be useful to assess the ergonomics of a user interface.

MeSH Keywords: **Critical Care • Equipment Design • Equipment Safety • Ventilators, Mechanical**

Full-text PDF: <https://www.medscimonit.com/abstract/index/idArt/911853>

 4673  5  —  35



Background

Mechanical ventilation is an important component in critical care practice and is widely used for patient safety during pre-hospital transportation and in-hospital transportation [1]. The use of ventilators in clinical settings is always a risk for the patient; one significant use-related risk for ventilators is human error [2,3]. In intensive care units (ICUs), the high workload and rapid response to multitasking required of medical personnel is a root cause of medical errors [4], and 67% of the incidents are caused by human error [5,6]. Mechanical ventilation is an important part of respiratory care, which accounts for nearly 25% of the daily ICU workload, and it can easily produce human errors [7,8]. Poorly designed ventilator user interfaces can negatively affect user performance. Patient safety specialists have confirmed that defects in the user interfaces of medical devices are the root cause of adverse events [9–11].

The evaluation of a medical device's user interface can be achieved through usability testing [12]. Usability testing has played an increasingly important role in medical device design in recent years, with the US Food and Drug Administration (FDA) requiring manufacturers to apply human factors engineering to product design and development processes to meet minimum use safety requirements [13]. Several studies on the evaluation of user interfaces in ventilators have been conducted, and these studies have confirmed that ventilator user interfaces can cause human error and reduce the performance quality of basic tasks [14–19]. However, these studies evaluating the ergonomics of ventilators mainly relied on measurement that lack objective performance measurement data on users, such as perceived workload (via NASA-TLX), task failure, task completion time, or subjective evaluation.

Given the shortcomings of these methods, the present study evaluated the ergonomics of the ventilator user interface through respiratory therapists' performance, workload, and user experience. To address the lack of objective user performance measurement data, our study used eye motions (via blink rate and blink duration) as physical responses to evaluate the physiological workload of participants when performing ventilator operation tasks. Blink rate and duration are important indicators of mental fatigue [20,21]. Studies have found that increased workload leads to a reduction in blink rate and shorter blink duration [22,23]. This workload evaluation method has been successfully applied in various fields, such as surgeons [22], radiation therapists [24], drivers [25], and pilots [26].

In this study, our primary purpose was to quantitatively evaluate the ergonomics of the ventilators' user interfaces using respiratory therapists' performance, workload, and user experience based on objective and subjective measures, which

provide an evaluation method of usability from the perspective of the user.

Material and Methods

Ventilators

The 3 tested ventilators were the Evita 4 (Draeger, Lubeck, Germany; version of software: 04.24 07/12/11), Servo I (Maquet, Solna, Sweden; version of software: v5.00.00), and Boaray 5000D (Probe, Shenzhen, China; version of software: 0A_006_V06.10.02_151119). Each machine was equipped with a standard double-limb circuit and was connected to a test lung (Venti-Plus™, GaleMed, Taipei, Taiwan, China). More details about the tested ventilators can be found in the supplemental materials.

Participants

Sixteen respiratory therapists, who are routinely responsible for daily ventilator operation, participated in the ventilator usability test. All of the respiratory therapists had a basic knowledge of mechanical ventilation and ventilator operation experience. More details about the participants can be found in the supplemental materials.

Specific test tasks

The participants were asked to accomplish 7 specific tasks on each ventilator: (1) start the ventilator, (2) mode and setting parameters recognition, (3) ventilator monitoring values recognition, (4) ventilator setting parameters modification, (5) ventilator mode modification, (6) alarm parameter recognition and resetting, and (7) respond to alarm. More details can be found in the supplemental materials.

Performance measure

Performance was measured by the percentage of test tasks (total of 7) with failures for each participant. An arbitrary upper-limit task completion time was determined and the correct answer had to be given in less than 3 min [15,17,18]. The task was identified as a failure if participants could not provide the correct answer or exceeded the time limit. Hence, a lower task errors percentage indicates a better user performance.

Workload evaluation

Workload was evaluated through objective (physiological) and subjective (perceived) measurements.

Physiological workload

The physiological workload was evaluated by blink rate and blink duration. We sampled eye motions data at 50 Hz using the Tobii Glasses 2 Eye Tracker (Tobii Technology, Danderyd, Sweden) when participants were performing test tasks. Before performing tasks on ventilators, each participant had to conduct a pupil calibration process with the eye tracker based on the manufacturer's recommendations. Participant's blink rate and blink duration data were collected with the eye tracker when they were performing tasks on the ventilators.

Perceived workload

Perceived workload was measured using the NASA-TLX questionnaires. The NASA-TLX evaluation of workload relies on 6 different psychological dimensions: mental demand, temporal demand, physical demand, frustration, performance, and effort. The result of the NASA-TLX is a score ranging from 0 to 100, where higher scores correspond to a higher mental workload and to the user interface being considered difficult to use. The NASA-TLX has been widely used in various studies for workload measures, such as therapy plan systems [24], ventilators [18], and monitoring devices [27].

User experience

User experience was assessed by the USE Questionnaire. The USE Questionnaire measured the usability dependent on 4 dimensions: usefulness, ease of use, ease of learning, and satisfaction [28]. The USE questionnaire contains 30 questions, and each question was evaluated by a Likert-type scale from 1 to 7 (from "strongly disagree" to "strongly agree", respectively). The USE Questionnaire has been used in various studies to evaluate user's perceived usability, such as in the evaluation of the software developed for radiologists [29], mobile application [30], and M-health care application [31]. Therefore, it enabled us to measure the participants' user experience.

Study protocol

This study was conducted in an unoccupied ICU treatment room of a university hospital, in Wuhan, Hubei province, China. The Ethics Committee of Tongji Medical College, Huazhong University of Science and Technology approved the study (IORG No: IORG0003571).

The 3 tested ventilators were evaluated in a random order for each participant (Supplementary Table 1, Supplementary Materials). Before the start of each task, participants were required to wear the eye tracker and to undergo the calibration process. After calibration, a researcher would stand near the ventilator to be tested and dictate the task to the participants.

Only 1 attempt was allowed for each task [15,17,18]; participants were allowed to perform the task with the ventilator when a start signal was given by the researcher. The participants were instructed to inform the researcher immediately when they accomplished the task. A clinical medical engineer would check the performance of the participants throughout each test task process. After this, the participants were then allowed to perform the next task.

After the participants performed all 7 tasks on 1 ventilator, they moved to the next ventilator to repeat the tasks. The participants were allowed to rest when needed.

Outcome measures

While the participants performed each task, their eye motion data were recorded by the eye tracker. After the participants completed each task, the task would be identified as a failure if participants could not provide the correct answer or exceeded the time limit. After all 7 tasks had been accomplished on 1 ventilator, the participants had to complete the NASA-TLX and the USE Questionnaire before beginning the tasks on the next ventilator. The NASA-TLX was used to evaluate mental workload when the participant accomplished the 7 tasks on a ventilator. The USE Questionnaire was conducted to evaluate the user experience on each tested ventilator.

Statistical analyses

Values are expressed as the mean \pm SD. The task error, blink rate, blink duration, NASA-TLX scores, and USE Questionnaire ratings were compared. The difference among ventilators for each task was assessed by the Friedman nonparametric test where $p < 0.05$ was considered significant. Post hoc multiple comparison tests were conducted using the Dunn-Bonferroni test [32]. The correlations between physiological workload (average blink rate and blink duration of all tasks) and perceived workload (NASA-TLX scores) as well as the correlation between workload (average blink rate, average blink duration and NASA-TLX scores) and user experience (the USE Questionnaire overall average scores) for each ventilator were analyzed using the Spearman correlation coefficient. The analyses were performed using the statistics software SPSS 20 (IBM Corporation, Armonk, New York).

Results

Performance

Task errors (% of tasks) for each ventilator were analyzed. The task errors when participants performed tasks on the 3 ventilators were significantly different ($p < 0.05$). The E vital 4

received the highest task errors (18.769 ± 6.846), the task errors for Boaray 5000D were 10.725 ± 6.395 , and the Servo I received the lowest (8.938 ± 7.150). Post hoc comparisons of task errors with Bonferroni correction were performed. The participants made fewer errors with the Servo I compared to the Evital 4 ($p<0.05$). Most task errors were made in the tasks of mode and setting parameters recognition and in ventilator monitoring values recognition for the 3 ventilators.

Workload evaluation

Physiological workload

Table 1 shows the blink rate for participants when accomplishing the 7 tasks on the 3 ventilators. For tasks 1, 3, 4, and 6, the blink rate showed significant differences among the ventilators ($p<0.05$). The average blink rate of all tasks was not significantly different among the ventilators ($p=0.06$). Post hoc multiple comparisons of blink rate among ventilators were analyzed using the Dunn-Bonferroni test (Table 1). After Bonferroni correction, 5 out of 13 comparisons were statistically significant. The lowest blink rate was recorded while accomplishing tasks on the Evital 4 compared with the other ventilators (except for task 1).

Table 2 shows blink duration during performance of the 7 tasks; we found that the blink durations were significantly different for tasks 1, 2, 3, 5, and 6 among the 3 machines ($p<0.05$). Furthermore, the average blink duration of all tasks showed significant differences among the ventilators ($p<0.05$). Post hoc multiple comparisons of the blink durations among ventilators are displayed in Table 2. After Bonferroni correction, 7 out of 15 comparisons were statistically significant. Moreover, for the post hoc multiple comparisons of the average blink duration of all tasks, the Evital 4 resulted in a poorer performance than that of the Servo I and Boaray 5000D ($p<0.05$).

According to the results in Tables 1 and 2, performance of the tasks (except for task 1) on the Evital 4 were more difficult than on the Servo I or Boaray 5000D, as shown by lower blink rates and shorter blink durations. It was easy for participants to perform the tasks on the Servo I and Boaray 5000D as shown by higher blink rates and longer blink durations (except for task 1). Overall, participants performing tasks on the Evital 4 received a larger physiological workload than they received on the other ventilators.

Perceived workload

Table 3 lists the results of each dimension for the 3 ventilators. The Evital 4 achieved the highest TLX value (40.042 ± 12.304 , $p<0.05$) and the Servo I achieved the lowest (23.750 ± 7.628 , $p<0.05$). For the post hoc comparisons of global task index and

dimensions (Table 3), the Evital 4 resulted in a higher task load index scores than did the Boaray 5000D and Servo I ($p<0.05$). Moreover, higher mental demand and performance dimension scores were observed for the Evital 4 than those for the Servo I ($p<0.05$). Details can be found in Table 3.

Correlation between physiological workload and perceived workload

The correlation between physiological workload and perceived workload were analyzed. For the Boaray 5000D, significantly negative correlations were shown between NASA-TLX workload and the average blink duration of all tasks ($r=-0.51$, $p=0.04$). Furthermore, a significantly negative correlation was also shown for the Servo I between the NASA-TLX workload and the average blink rate ($r=-0.59$, $p=0.02$). The other correlations were not significantly different. According to the results of correlation between physiological workload and perceived workload, the NASA-TLX workload showed a negative r value with average blink rate and blink duration, which suggests that a lower blink rate and shorter blink durations matched the increasing workload of the participants.

User experience

Table 4 shows the results of user experience for each ventilator. The Servo I received the highest overall average scores for the USE Questionnaire (5.684 ± 0.900 , $p<0.05$) and the Servo I received the lowest (4.894 ± 0.981 , $p<0.05$). Furthermore, statistically significant differences were also found in usefulness and satisfaction ($p<0.05$). For the post hoc comparisons of the overall average scores, usefulness, and satisfaction (Table 4), the Evital 4 resulted in lower scores than did the Servo I ($p<0.05$). Moreover, lower scores were observed in all evaluation dimensions for the Evital 4 than for the Servo I and Boaray 5000D in the USE Questionnaire ($p<0.05$). Details can be found in Table 4.

Correlation between workload and user experience

The correlation between workload and user experience were analyzed. For the Servo I, significant positive correlations were shown between user experience and average blink rate for all tasks ($r=0.632$, $p<0.01$). Furthermore, a significantly negative correlation was shown for the Boaray 5000D between user experience and NASA-TLX workload ($r=-0.562$, $p=0.02$). The other correlations were not significantly different. According to the results, the NASA-TLX workload had a negative r value with user experience, which suggests that a worse user experience matched the higher workload. Similarly, the average blink rate and blink duration had positive r values with user experience, which suggests that a higher blink rate and longer blink duration matched the better user experience.

Table 1. Blink rate for each task in ventilators and mean differences ($M_D=Mr_1-Mr_2$) with the results of post hoc multiple comparisons of blink rate.

Task	Mean blink rate for each task among ventilators			p	Post hoc tests with Bonferroni correction			p
	Evital 4	Servo I	Boaray 5000D		Mr ₁	Mr ₂	M _D (Mr ₁ -Mr ₂)	
	Mean ±SD	Mean ±SD	Mean ±SD					
Task 1: start the ventilator (blinks/second)	0.201±0.557	0.115±0.593	0.122±0.067	p<.01	Servo I	Evital 4	-0.086	p<.01
					Boaray 5000D	Evital 4	-0.079	0.02
					Servo I	Boaray 5000D	-0.007	1.00
Task 2: mode and setting parameters recognition (blinks/second)	0.164±0.078	0.201±0.542	0.184±0.035	0.82				
Task 3: ventilator monitoring values recognition (blinks/second)	0.103±0.074	0.186±0.041	0.156±0.044	0.01	Servo I	Evital 4	0.083	p<.01
					Boaray 5000D	Evital 4	0.053	0.34
					Servo I	Boaray 5000D	0.030	0.34
Task 4: ventilator setting parameters modification (blinks/second)	0.120±0.052	0.198±0.038	0.168±0.030	0.01	Servo I	Evital 4	0.078	0.01
					Boaray 5000D	Evital 4	0.048	0.23
					Servo I	Boaray 5000D	0.030	0.65
Task 5: ventilator mode modification (blinks/second)	0.105±0.047	0.146±0.036	0.136±0.049	0.08				
Task 6: alarm parameter recognition and resetting (blinks/second)	0.108±0.051	0.151±0.067	0.179±0.056	0.02	Servo I	Evital 4	0.043	0.16
					Boaray 5000D	Evital 4	0.071	0.01
					Servo I	Boaray 5000D	-0.028	1.00
Task 7: respond to alarm (blinks/second)	0.128±0.079	0.148±0.085	0.131±0.038	0.36				
Average blink rate of all tasks (second)	0.132±0.029	0.163±0.016	0.153±0.015	0.06				

Positive M_D values representing Mr_1 lower workload than Mr_2 .

Discussion

The aim of this study was to evaluate the ergonomics of 3 intensive care ventilators through respiratory therapists' performance, workload, and user experience. As the test tasks were the same for all tested ventilators, the results of this study showed that the ergonomic design of the user interface has an influence on respiratory therapists' performance, workload, and user experience. Traditionally, when a medical device-related adverse event occurs, it is natural to attribute the cause to the human. User interface design defects can serve as an example

of potential causes that could lead to adverse events. In this study, several issues have been identified and show room for improvement. For instance, the power switch presents several problems. As to be expected given that 91% of Chinese people are right-handed [33], in the study, researchers noted that all the participants looking for the power switch ("on/off" button) start from the right-hand side of the machine, but only the Evita 4 put it on the right side. For the Boaray 5000D and Servo I, whose switches are on the back left of the machine, and particularly for the Servo I, where the power switch is hidden behind a sliding cover, it was difficult for participants to

Table 2. Blink duration for each task in ventilators and mean differences ($M_D=Mr_1-Mr_2$) with the results of post hoc multiple comparisons of blink duration.

Task	Mean blink rate for each task among ventilators			p	Post hoc tests with Bonferroni correction			p
	Evital 4	Servo I	Boaray 5000D		Mr ₁	Mr ₂	M _D (Mr ₁ -Mr ₂)	
	Mean ±SD	Mean ±SD	Mean ±SD					
Task1: start the ventilator (second)	0.117±0.170	0.101±0.019	0.101±0.009	p<.01	Servo I	Evital 4	-0.016	0.01
					Boaray 5000D	Evital 4	-0.016	0.02
					Servo I	Boaray 5000D	0.000	1.00
Task2: mode and setting parameters recognition (second)	0.097±0.014	0.123±0.026	0.111±0.026	0.02	Servo I	Evital 4	0.026	0.02
					Boaray 5000D	Evital 4	0.014	0.40
					Servo I	Boaray 5000D	0.012	0.65
Task3: ventilator monitoring values recognition (second)	0.091±0.010	0.123±0.010	0.111±0.020	p<.01	Servo I	Evital 4	0.032	p<.01
					Boaray 5000D	Evital 4	0.020	0.01
					Servo I	Boaray 5000D	0.012	1.00
Task4: ventilator setting parameters modification (second)	0.104±0.018	0.123±0.034	0.114±0.027	0.45				
Task5: ventilator mode modification (second)	0.093±0.014	0.118±0.021	0.112±0.018	0.01	Servo I	Evital 4	0.025	0.01
					Boaray 5000D	Evital 4	0.019	0.28
					Servo I	Boaray 5000D	0.006	0.75
Task6: alarm parameter recognition and resetting (second)	0.094±0.013	0.109±0.019	0.118±0.022	0.03	Servo I	Evital 4	0.015	0.23
					Boaray 5000D	Evital 4	0.024	0.04
					Servo I	Boaray 5000D	-0.009	1.00
Task7: respond to alarm (second)	0.097±0.023	0.112±0.025	0.104±0.023	0.29				
Average blink duration of all tasks (second)	0.099±0.009	0.116±0.013	0.111±0.009	p<.01	Servo I	Evital 4	0.017	0.01
					Boaray 5000D	Evital 4	0.005	0.01
					Servo I	Boaray 5000D	0.012	1.00

Positive M_D values representing Mr_1 lower workload than Mr_2 .

find the switch and turn it on, thereby making it hard to use in urgent environments. This conclusion is confirmed by the results of blink rate and blink duration for task 1 in Tables 1 and 2. Gonzalez-Bermejo [16] showed that the power switch

on the back of the ventilator was difficult to find, suggesting that the power switch should be placed on the front panel of the ventilator. On a related note, Laurence [17] found that a

Table 3. NASA-TLX workload scores for each ventilator and mean differences ($M_D=Mr_1-Mr_2$) with the results of post hoc multiple comparisons of NASA-TLX workload scores.

NASA-TLX workload	NASA-TLX workload scores			P	Post hoc tests with Bonferroni correction			P
	Evital 4	Servo I	Boaray 5000D		Mr ₁	Mr ₂	M _D (Mr ₁ -Mr ₂)	
	Mean ±SD	Mean ±SD	Mean ±SD					
TLX (task load index)	40.042± 12.304	23.750± 7.628	26.083± 5.791	p<.01	Servo I	Evital 4	-16.292	p<.01
					Boaray 5000D	Evital 4	-13.959	0.02
					Servo I	Boaray 5000D	-2.333	1.00
Mental demand	10.375± 4.286	4.625± 5.513	8.333± 4.286	p<.01	Servo I	Evital 4	-5.750	p<.01
					Boaray 5000D	Evital 4	-2.042	0.28
					Servo I	Boaray 5000D	-3.708	0.47
Physical demand	2.168± 3.783	0.833± 1.450	0.958± 1.088	0.57				
Temporal demand	8.083± 5.158	6.250± 4.021	6.083± 3.593	0.50				
Performance	8.958± 5.817	3.792± 3.300	4.667± 2.271	0.01	Servo I	Evital 4	-5.166	0.02
					Boaray 5000D	Evital 4	-4.291	0.05
					Servo I	Boaray 5000D	-0.875	1.00
Effort	7.625± 5.497	5.333± 3.578	3.333± 3.070	0.08				
Frustration	2.833± 3.197	2.917± 2.920	2.708± 2.412	0.77				

Positive M_D values representing Mr₁ lower workload than Mr₂.

cover over the power switch was good for safety but difficult for the user to easily and quickly switch on and off.

The most significant usability problem noted by our study was the different terminology in use between ventilators. In the mode and setting parameters recognition and ventilator monitoring values recognition tasks, participants got confused with the different terminology that each ventilator used. Most ventilation terminology was presented in English acronyms and different ventilator manufacturers used different acronyms for the same terms. However, the use of different acronyms among ventilators resulted in unnecessary user confusion, added additional operational error, and increased the workload for task completion. Several studies have observed this problem [14–16,18] and have confirmed that heterogeneous terminology of ventilation modes and set/monitoring parameters increases operational failures. Thus, it seems necessary to have a standardized terminology and ensure that

the unified terminology is easy for the user to read and understand. Furthermore, the ventilators' alarms also need additional attention. The visual and auditory stimulation and display interface for the alarms need improvement so that users can quickly recognize the occurrence of the alarm and easily read and understand the alarm content.

However, published studies mostly have focused on perceived workload (via NASA-TLX), task failure, task completion time, or subjective evaluation to evaluate the ergonomics of the user interface for ventilators, the studies lack objective user performance data, and few studies have used the physical signs of users to evaluate the ergonomics of the user interface for ventilators. In this study, we applied the use of eye motions (blink rate and blink duration) indicators as physiological signs of workload to evaluate the ergonomics of the user interface. Published studies have proved that blink rate and blink duration are strongly associated with workload [22–24]. The lower

Table 4. The USE Questionnaire scores for each ventilator and mean differences ($M_D=Mr_1-Mr_2$) with the results of post hoc multiple comparisons of the USE Questionnaire scores.

User experience	The USE Questionnaire scores			P	Post hoc tests with Bonferroni correction			P
	Evital 4	Servo I	Boaray 5000D		Mr1	Mr2	M_D (Mr_1-Mr_2)	
	Mean \pm SD	Mean \pm SD	Mean \pm SD					
The overall average scores	4.894 \pm 0.981	5.684 \pm 0.900	5.328 \pm 0.747	p<.01	Servo I	Evital 4	0.790	p<.01
					Boaray 5000D	Evital 4	0.434	0.34
					Servo I	Boaray 5000D	0.356	0.34
Usefulness	4.885 \pm 0.961	5.878 \pm 1.032	5.360 \pm 0.716	p<.01	Servo I	Evital 4	0.993	p<.01
					Boaray 5000D	Evital 4	0.475	0.56
					Servo I	Boaray 5000D	0.518	0.10
Ease of use	4.976 \pm 1.028	5.688 \pm 0.811	5.221 \pm 0.795	0.06				
Ease of learning	4.828 \pm 0.835	5.563 \pm 0.824	5.188 \pm 0.854	0.07				
Satisfaction	4.812 \pm 1.155	5.527 \pm 1.035	5.536 \pm 0.764	0.03	Servo I	Evital 4	0.715	0.04
					Boaray 5000D	Evital 4	0.724	0.65
					Servo I	Boaray 5000D	-0.009	0.65

Positive M_D values representing Mr_1 lower workload than Mr_2 .

blink rate and shorter blink duration mean an increasing mental workload, which can be used to evaluate the ergonomic design of the user interface. The blink rate results during the 7 tested tasks among the 3 ventilators are presented in Table 1, and significant differences were found in tasks 1, 3, 4, 6. For the Servo I, the data from Table 1 show that the blink rates were higher than those for with the Evital 4 in tasks 3 and 4 ($p<.01$; $p<.01$, respectively). For the Boaray 5000D, the blink rate was higher than it was for the Evital 4 in task 6 ($p=0.01$). For the Evital 4, the blink rate was higher than it was for the Servo I and Boaray 5000D in task 1 ($p<.01$; $p=0.02$, respectively). These results show that the Servo I outperformed the Evital 4 in tasks 3, 4, and the Boaray 5000D performed better in task 6 than did the Evital 4. Furthermore, the Evital 4 outperformed the Servo I and Boaray 5000D in task 1. The results of the blink duration lead to a conclusion similar to those from Table 2. The longer blink duration for the Evital 4 compared to that of the Servo I and Boaray 5000D in task 1 means a lower workload for participants when performing task 1 on the Evital 4 ($p=0.01$ and $p=0.02$, respectively). The Servo I outperformed the Evital 4 in task 2, 3 and 5 ($p=0.02$ and $p<.01$; $p=0.01$, respectively). The Boaray 5000D outperformed the Evital 4 in tasks 3 and 6 ($p=0.01$ and $p=0.04$, respectively).

Furthermore, the Evital 4 had the shorter average blink duration of all tasks than did the Servo I and Boaray 5000D ($p=0.01$ and $p=0.01$, respectively). For the eye motions data, a lower blink rate and shorter blink duration mean participants experienced poor performance on the Evital 4 (except task 1) and this also confirmed that blink rate and duration can be effective indicators to evaluate the ergonomics of the user interface.

The above conclusions were also supported by task errors and user experience, for which the Evital 4 received the higher task errors and lower USE Questionnaire scores. Furthermore, participants thought the Servo I would have more usefulness and higher user satisfaction ($p<.01$ and $p=0.04$, respectively) than the Evital 4.

The correlation between subjective and objective workload measures were analyzed. Two of 6 correlations were significant among ventilators. Furthermore, the NASA-TLX scores have a negative correlation with blink rate and blink duration, which means the lower blink rate and shorter blink duration matches the increasing mental workload. These conclusions are also supported by the correlation between workload and user

experience, in which higher blink rates, longer blink duration, and lower NASA-TLX scores matched a better user experience.

Our results show that using blink rate and blink duration to evaluate the ergonomic design of ventilators can be an effective method. These physical signs of the user can be useful indicators to show the ergonomics of ventilator from the users' experience and can be reliable objective indicators of user performance data to evaluate the medical device user interface.

Several ergonomic studies of ventilators have demonstrated that the manufacturer should pay more attention to ergonomic design of ventilator user interfaces. For instance, Hodges [34] studied the speed and ability of nursing and medical staff to successfully activate capnography before and after a specific episode of training and assessment, finding that the ergonomic design of the ventilator user interface affects capnography activation. Marjanovic [35] used psycho-cognitive scales (system usability scale and mental workload) and physiological measurements (pupil diameter, heart and respiratory rate, and thoracic volume variations) to assess 20 senior ICU physicians completing 11 specific tasks for each ventilator, finding that some ventilators show low ergonomics performance and a high risk of user errors. Compared with our study, these studies evaluating the ergonomics of ventilators mainly relied on perceived workload (via NASA-TLX), task completion time, subjective evaluation (SUS), and physiological measurements (via pupil diameter and heart and respiratory rate), which fail to explore the relationships between subjective and objective data. In our study, the correlations between subjective and objective data for each ventilator were analyzed, and these corrections show that a good ergonomic design of the user interface was associated with better user experience and lower NASA-TLX scores.

Limitations of the study

In this study, there are several limitations that should be recognized. First, the participants in our study were respiratory therapists and represent only 1 category of ventilator users. Thus, the results of this study cannot be directly applied to other user categories. Second, compared with other studies [14,16–18], testing only 3 ventilators may be insufficient. However, the 3 ventilators include almost all of those in use in our local region intensive care units and which were available for our study. The user interface of other ventilator brands may have some usability issues, but our intent was not to compare the usability among different manufacturers. Finally, the chosen ventilators can perform more functions than were tested in this study, but the tasks that are representative of those performed by respiratory therapists, such as parameters modification, are always practical requirements in ICUs.

Conclusions

This study provides a comprehensive method to evaluate the usability of ventilators from the perspective of respiratory therapists' performance, workload, and user experience. The Evita 4 resulted in poor performance in the tests when participants performed tasks on it. The results of this study show that the 3 ventilators tested had usability shortcomings in the design of the user interface that increased the mental workload of the user and can lead to failures. Therefore, optimizing the design of the user interface is needed to reduce these failures. Furthermore, this study proved that eye motion data (blink rate and duration) is useful in evaluating the ergonomics of the user interface.

Acknowledgments

We acknowledge the support of respiratory therapists from Union Hospital, Tongji Medical College, and Huazhong University of Science and Technology, who made this study possible.

Conflict of interests

None.

Supplemental Materials

Ventilators

The 3 tested ventilators were the Evita 4 (Draeger, Lubeck, Germany; version of software: 04.24 07/12/11), Servo I (Maquet, Solna, Sweden; version of software: v5.00.00), and Boaray 5000D (Probe, Shenzhen, China; version of software: 0A_006_V06.10.02_151119). Each machine was equipped with a standard double-limb circuit and was connected to a test lung (Venti Plus™, GaleMed, Taipei, Taiwan, China). These 3 ICU ventilators are now in extensive use in medical institutions in our local area and have similar user interface design features. Several new-generation ventilators have been developed by manufacturers, such as the V500 and Servo U. However, these ventilators are rarely used in our local area, making them unavailable for our tests. However, the tested ventilators reflect the actual application of ventilators in our local medical institutions.

Participants

Sixteen respiratory therapists, who are routinely responsible for daily ventilator operation, participated in the ventilator usability test. All of the respiratory therapists had a basic knowledge of mechanical ventilation and ventilator operation experience. Before the formal study, we provided operational training on the 3 tested ventilators (Evita 4, Servo I, and Boaray

5000D) for all participants. All participants were given a series of learning goals and were required to familiarize themselves with the tested ventilators. An expert was available to answer the participants' questions. When participants felt able to use the ventilators on a real patient, a test, including setting values modification, browsing the menu, and searching for monitoring values, was conducted for participants to demonstrate their ability to independently use the ventilators on the patient. A pilot study with 3 participants was performed to improve the test flow and analyze the reliability of the study data.

Tasks to accomplish

1. Start the ventilator

With the ventilator completely assembled and connected to the power and test lung, the participants had to start the ventilator. The task stop signal was given when the first insufflation was produced by the ventilator.

2. Mode and setting parameters recognition

The tester would set a specific ventilation mode and turn it on. The participant had to respond to 2 questions: First, the participant identified the ventilation mode of the tested ventilator. In this study, we selected 2 ventilation modes – VC-IMV or PC-CSV – which were alternated on each tested ventilator. Second, they had to recognize the setting parameters in the current running mode. In the VC-IMV mode, the participants needed to recognize the setting parameters as follow: inspired oxygen fraction (FIO₂), tidal volume (VT), respiratory rate (RR), and positive end-expiratory pressure (PEEP), for which the setting values are FIO₂ 0.3, VT 600 ml, RR 18/min, and PEEP 8 cm H₂O, respectively. In the PC-CSV mode, the participants needed to recognize the following setting parameters: inspired oxygen fraction (FIO₂), respiratory rate (RR), positive end-expiratory pressure (PEEP), and inspiratory pressure (P_{insp}), for which the setting values are FIO₂ 0.5, RR 14/min, PEEP 5 cm H₂O, and P_{insp} 10 cm H₂O, respectively. The stop signal was given when participants had answered the above questions.

3. Ventilator monitoring values recognition

When the tested ventilator was turned on in a specific ventilation mode, participants had to inform the testers of the monitored values. When in the VC-IMV mode, participants had to

inform the tester of the following monitored values: plateau pressure (P_{plat}), peak inspiratory pressure (P_{peak}), minute volume (MV), and expired tidal volume (V_{Te}). When in the PC-CSV mode, participants had to inform the tester of the following monitored values: minute volume (MV), respiratory rate (RR), positive end-expiratory pressure (PEEP), and tidal volume (VT). The stop signal was given when participants had reported all required monitored values.

4. Ventilator setting parameters modification

With the tested ventilator running in a specific ventilation mode, participants had to reset the values of setting parameters. In the VC-IMV mode, participants reset the following values: FIO₂ 0.4, VT 400 ml, RR 15/min, and PEEP 6 cm H₂O. In the PC-CSV mode, the values to be set were FIO₂ 0.6, RR 18/min, PEEP 8 cm H₂O, and P_{insp} 8 cm H₂O. The stop signal was given when all setting values were changed and activated.

5. Ventilator mode modification

With the tested ventilator running in a specific ventilation mode, participants had to modify the ventilation mode from VC-IMV to PC-CSV or from PC-CSV to VC-IMV. The stop signal was given with the first insufflation in the new mode.

6. Alarm parameter recognition and resetting

With the tested ventilator running in a specific ventilation mode, participants had to inform the tester of the following alarm setting parameter values: minute volume (MV), respiratory rate (RR), and airway pressure (P_{aw}). After this, participants had to reset the value of the above parameters. The stop signal was given when the change was activated.

7. Respond to alarm

With the tested ventilator running in a specific ventilation mode, the tester changed 1 alarm setting value to trigger an alarm. The participants had to stop the alarm, report the alarm content, adjust the alarm to predefined values, and reset the alarm. In this study, the alarms were low pressure, high tidal volume, and apnea, and they were alternated in that order between ventilators. The stop signal was given when the alarm values had been adjusted to the required levels.

Supplementary Table 1. Randomization table for device testing.

Participant number	Ventilator type		
1	Boaray 5000D	Servo I	Evital 4
2	Boaray 5000D	Evital 4	Servo I
3	Servo I	Boaray 5000D	Evital 4
4	Servo I	Evital 4	Boaray 5000D
5	Evital 4	Servo I	Boaray 5000D
6	Evital 4	Boaray 5000D	Servo I
7	Servo I	Evital 4	Boaray 5000D
8	Boaray 5000D	Servo I	Evital 4
9	Servo I	Evital 4	Boaray 5000D
10	Evital 4	Servo I	Boaray 5000D
11	Servo I	Evital 4	Boaray 5000D
12	Evital 4	Servo I	Boaray 5000D
13	Evital 4	Boaray 5000D	Servo I
14	Servo I	Boaray 5000D	Evital 4
15	Evital 4	Servo I	Boaray 5000D
16	Boaray 5000D	Servo I	Evital 4

This table details the randomization table for device testing by the respiratory therapists.

References:

- Needham D, Bronskill S, Calinawan J et al: Projected incidence of mechanical ventilation in Ontario to 2026: Preparing for the aging baby boomers. *Crit Care Med*, 2005; 33(3): 574–79
- Auriant I, Reignier J, Pibarot M et al: Critical incidents related to invasive mechanical ventilation in the ICU: Preliminary descriptive study. *Intens Care Med*, 2002; 28(4): 452–58
- Gravenstein J: How does human error affect safety in anesthesia? *Surg Oncol Clin N Am*, 2000; 9(1): 81–95
- Rothschild J, Landrigan C, Cronin J et al.: The Critical Care Safety Study: The incidence and nature of adverse events and serious medical errors in intensive care. *Crit Care Med*, 2005; 33(8): 1694–700
- Bracco D, Favre JB, Bissonnette B et al: Human errors in a multidisciplinary intensive care unit: A 1-year prospective study. *Intens Care Med*, 2001; 27(1): 137–45
- Giraud T, Dhainaut J, Vaxelaire J et al: Iatrogenic complications in adult intensive care units: A prospective two-center study. *Crit Care Med*, 1993; 21(1): 40–51
- Donchin Y, Gopher D, Olin M et al: A look into the nature and causes of human errors in the intensive care unit. *Crit Care Med*, 1995; 23(2): 294–300
- Wright D, Mackenzie SJ, Buchan I et al: Critical incidents in the intensive therapy unit. *Lancet*, 1991; 338(8768): 676–78
- Crowley, John J, Kaye, Ronald D: Identifying and understanding medical device use errors. *J Clin Eng*, 2002; 27(3): 188–93
- Israelski E, Muto W: Human factors risk management as a way to improve medical device safety: A case study of the therac 25 radiation therapy system. *Jt Comm J Qual Saf*, 2004; 30(21): 689–95
- Perry S: An overlooked alliance: Using human factors engineering to reduce patient harm. *Jt Comm J Qual Saf*, 2004; 30(8): 455–59
- Wiklund ME, Kendler J, Strohlic AY: Usability testing of medical devices. Florida: Taylor & Francis, 2011
- Food and Drug Administration (US): Applying Human Factors and Usability Engineering to Medical Devices. U.S. Food and Drug Administration. Washington: Food and Drug Administration, 2016
- Uzawa Y, Yamada Y, Suzukawa M: Evaluation of the user interface simplicity in the modern generation of mechanical ventilators. *Respir Care*, 2008; 53(3): 329–37
- Templier F, Miroux P, Dolveck F et al: Evaluation of the ventilator-user interface of 2 new advanced compact transport ventilators. *Respir Care*, 2007; 52(12): 1701–9
- Gonzalez-Bermejo J, Laplanche V, Hussein F et al: Evaluation of the user-friendliness of 11 home mechanical ventilators. *Eur Respir J*, 2006; 27(6): 1236–43
- Vignaux L, Tassaux D, Jolliet P: Evaluation of the user-friendliness of seven new generation intensive care ventilators. *Intens Care Med*, 2009; 35(10): 1687–91
- Marjanovic N, L'Her E: A comprehensive approach for the ergonomic evaluation of 13 emergency and transport ventilators. *Respir Care*, 2016; 61(5): 632–39
- Morita PP, Weinstein PB, Flewwelling CJ et al: The usability of ventilators: A comparative evaluation of use safety and user experience. *Crit Care*, 2016; 20(1): 263–72
- Marshall S: Identifying cognitive state from eye metrics. *Aviat Space Environ Med*, 2007; 78(5 Suppl.): B165–75
- Ryu K, Myung R: Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. *Int J Ind Ergonom*, 2005; 35(11): 991–1009
- Zheng B, Jiang X, Tien G et al: Workload assessment of surgeons: correlation between NASA TLX and blinks. *Surg Endosc*, 2012; 26(10): 2746–50
- May JG, Kennedy RS, Williams MC et al: Eye movement indices of mental workload. *Acta Psychol (Amst)*, 1990; 75(1): 75–89
- Mazur LM, Mosaly PR, Hoyle LM et al: Subjective and objective quantification of physician's workload and performance during radiation therapy planning tasks. *Pract Radiat Oncol*, 2013; 3(4): e171–77
- Benedetto S, Pedrotti M, Minin L et al: Driver workload and eye blink duration. *Transp Res F Traffic Psychol Behav*, 2011; 14(3): 199–208
- Veltman J, Gaillard A: Physiological workload reactions to increasing levels of task difficulty. *Ergonom*, 1998; 41(5): 656–69

27. Görges M, Staggers N: Evaluations of physiological monitoring displays: A systematic review. *J Clin Monit Comput*, 2008; 22(1): 45–66
28. Lund A: Measuring usability with the USE questionnaire. *Usability Interface*, 2001; 8(2): 3–6
29. Martynov P, Mitropolskii N, Kukkola K et al: Testing of the assisting software for radiologists analysing head CT images: Lessons learned. *BMC Med Imaging*, 2017; 17(1): 59
30. González-Landero F, García-Magariño I, Lacuesta R, Lloret J: PriorityNet App: A mobile application for establishing priorities in the context of 5G ultradense networks. *IEEE Access*, 2018; 14(8): 14141–50
31. Lestantri ID, Putrima, Sabiq A, Suherlan E: Developing and pilot testing M-health care application for pregnant and toddlers based on user experience. *J Phys*, 2018; 978(1): 012067
32. Dunn OJ: Multiple comparison using RANK sums. *Technometrics*, 1964; 6(3): 241–52
33. Li X: [The distribution of left and right handedness in Chinese people.] *Acta Psychologica Sinica*, 1983; 15(3): 268–76 [in Chinese]
34. Hodges E, Griffiths A, Richardson J et al: Emergency capnography monitoring: Comparing ergonomic design of intensive care unit ventilator interfaces and specific training of staff in reducing time to activation. *Anaesthesia*, 2012; 67(8): 850–54
35. Marjanovic NS, Simone AD, Jegou G et al: A new global and comprehensive model for ICU ventilator performances evaluation. *Ann Intensive Care*, 2017; 7(1): 68