Usability and Use Safety in the Context of Medical Devices

Maxim Javdoschin

January 6, 2023

Abstract

Background: Usability and use safety are a crucial aspect of medical devices because people's lives may depend on it. Studies to evaluate and compare a device's usability are therefore unavoidable and seeing as new devices are released constantly, have to be performed often.

Explanation of Methods: Three usability studies of critical care ventilators are summarized, explained and compared to illustrate different means and ways to conduct usability studies. The studies of Morita, Marjanovic and Spaeth are common in some but different enough in other ways to better explain different results based on the priorities for usability studies.

Results of the Studies: The chosen metrics and statistical proceedings allowed the three teams to directly compare ventilator results through different means of visualization. Through those means, each group was able to notice significant differences in performance in some of the tested areas and evaluate the over-performing and under-performing ventilators.

Discussion: The three studies focus on different ventilators but have some common evaluation metrics such as the established SUS Questionnaire and the NASA-TLX metric for workload evaluation. Depending on demographics, there are also different things to account for. Experienced operators have less problems when faced with the tasks but biases through their experience may overshadow results, as can be seen in Morita's data evaluation. Naive operators, which are only chosen by Spaeth, may give more raw and unbiased data, but require an introduction into the usage of critical care ventilators and are not the ones currently using the devices, so their results may vary from skilled technicians, which is illustrated through Spaeth's graphics.

Conclusion: There are many established metrics and statistical proceedings developed for usability evaluation, but self-measured data through other means like eye-tracking devices may show more praxis oriented results. Regardless, all of the measurement means follow the defined components and will yield the wanted results of usability and use safety.

1 Background

The usability and use safety of technical devices is often as crucial of a thing to get right in their design as it is to have all necessary functions. While the functionality can easily be tested by the manufacturers or the customers, usability and use safety are harder to determine, because the ones more involved in the development process have a much better overview over the devices.

Usability is defined by the five components learnability, efficiency, memorability, errors, and satisfaction. These components compute to the easiness, with which a device can be handled by a first-time user, the quickness of task completion, the re-establishment of proficiency (not evaluated in the studies mentioned down below), how many error the users make during this process, and how pleasant the design of the device is to use [14].

To ensure the best usability and use safety, usability tests are conducted. The evaluation is part of the completion process prior to regulatory approval, but these evaluations are conducted in a vacuum and only needing to meet minimal requirements. Adding to this is the confidentiality of the information gathered [13]. To address this issue, individual teams of researchers conduct usability tests on multiple devices to not only make the information of these tests accessible, but also compare similar devices of different manufacturers to put the results into perspective. These studies are usually held with a group of volunteers who are given tasks to complete using the tested devices. Outside of that, usability tests can differ greatly from one another through different requirements the volunteers have to meet to on-boarding processes.

Plinio P. Morita, Nicolas S. Marjanovic, Johannes Spaeth and their respective groups all conducted studies to evaluate usability and use safety on critical care ventilators [13] [12] [17]. Especially in critical patient care many devices are used to help sustaining or monitoring a patients health. Ventilators are a fundamental technology in that area. Because of their vital role, any errors in their functionality and use are detrimental to the life of patients. The following will describe some different ways usability tests can be performed by examining the three different papers (by Morita, Marjanovic and Spaeth respectively) documenting usability tests on critical care ventilators.

2 Explanation of Methods

2.1 Ventilators

The three teams of scientists all sampled their ventilators from market leading manufacturers from the time of writing. They chose four to six ventilators as samples with some overlapping models and/or manufacturers as listed in the table below [Table 1].

In addition to the chosen ventilators, Marjanovic and Spaeth hooked the devices up to a lung simulator. Marjanovic used the ASL5000 lung simulator (Ingmar, Pittsburgh, PA, USA) and Spaeth the SMS Manley Lung Simulator

Ventilators		
Morita	**Marjanovic**	**Spaeth**
Covidien Puritan Bennett 980	Covidien Puritan Bennett 980	
Maquet SERVO-U	Maquet SERVO-U	Maquet Flow-i
Dräger V500	Dräger V500	Dräger Perseus and Primus
Hamilton G5	Hamilton S1	
	Philips V680	
	General Electrics R860	
	Carefusion Avea (reference device)	
		Avance CS

Table 1: The ventilators used by the three teams sorted by manufacturer.

(BC Group International Inc., St Charles, MO, USA).

2.2 Participants

Usability studies can not be performed outright without any preparation. You need to know, how the devices compare and what tasks would be sufficient to achieve the best possible data. An additional step could be to perform a pilot study. Morita opted to perform such a study with 13 participants to establish priori knowledge. He used the three metrics of use safety, system usability and workload for power and significance at the 5% level, as established by Jacob Cohen [3], on a test-group of 13 participants. He deducted, that a minimum of 48 participants would be required to show substantial differences between the selected models of ventilators.
All of his 48 participants were respiratory therapists (RTs), including the 13 chosen for the pilot study. In North American hospitals, it is their responsibility responding in emergency situations, initiating and managing ventilators and providing airway management in high-risk areas [1]. As such, RTs are the primary users of critical care ventilators and the prime subjects for usability studies concerning these ventilators.
Additionally to selecting RTs as a subject sample, surveys were used to determine RTs with the experience with familiar ventilators from the same families and further care was put into avoiding RTs with ties to the manufacturers, as to not influence the results by any biases.

Marjanovic chose 20 senior ICU (Intensive Care Unit) physicians from five different ICUs for the evaluation. Everyone of the ICU physicians only tested 3-4 devices (unlike Moritas test, where every RT tested every device) but the device order was randomized and spaced out, so that each device was tested 11 or 12 times. For the selection process particular attention was paid to the fact,

that none of the participants were familiar with the used devices and also had work experience with the Carefusion Avea, which the group used as a reference device for the ergonomics evaluation.

The biggest difference were Spaeth's chosen participants. 28 volunteers participated in his study. 22 of those were not form a professional workforce in the critical care department but rather medical students that were blinded against the intention of the study. They had no prior experience with any type of critical care ventilator and so were, in the case of the study, considered naive operators (NOs). They were chosen as to avoid habituation or coping strategies experienced users may form. As, due to missing data, a sample size calculation based on empirical values could not be performed, the sample size was calculated based on pre-existing literature, and was estimated to disclose 95% of usability issues [5].
A separate group for experienced operators was formed, comprised of six anaesthesia residents, that also volunteered for the study. The data arising from this group were merely to draw a comparison to the tests performed by the NOs. As further precautions, operators where denied participation, if they indicated uptake of sedatives, alcohol or drugs in the past 10 h before study or had a sleep period of less than 5 h.

2.3 Demographics

Out of the 48 participants Morita used as test subjects 34% were male (n = 16) and 66% female (n = 32). Furthermore, 68% of the participants were in the age range of 25 to 45 years old (n = 33) and 63% had at least 5 years of experience as a RT. A balance of participants experience levels could not be achieved, but the experience levels did not affect the results in any significant way. Only some prior experience with the PB980 and G5 had a minor effect on the PSSUQ score of the PB980.

Marjanovic did not integrate an analysis of his chosen demographic into his paper.

Spaeth's NOs ages were all between 24 and 35 years old, with eleven of those being female and the rest nine male. The reaction times for both demographics were recorded, with the NOs having a reaction time median of 333ms (within a range of 211-432ms). The six EOs ranged from 30-43 years old with a distribution of two female and 4 male participants. Their reaction time median was at 309ms (within a range of 249-364ms). Two NOs of the 22 were exempt from the study because of familiarity with one of the devices, which they were not aware of until visual inspection.

2.4 Tasks and Scenarios

During Moritas study, the participants were to complete 16 different tasks, which were based on the international standard: ISO 80601-2-12:2020, Medical electrical equipment — Part 2-12: Particular requirements for basic safety and

Tasks		
Morita	**Marjanovic**	**Spaeth**
ventilator parameter set-up/start ventilation, adjust alarm limits	alarm control	set VCV mode
Activate expiratory and inspiratory pause	mode recognition	set tidal volume
read/ adjust respiratory rate	identify humidification system on the screen	set vent. frequency
leak test	ventilator setting reading	set PEEP
wean from pressure control to mandatory ventilation	power on the ventilator	start VCV
return to previous mode	start ventilation	set P max
standby	set inspiratory flow	set inspiratory flow
	ventilator mode modification	set I:E ratio
	set cycling to 60%	set end-insp. pause
	non-invasive ventilation mode activation	open alarm menu
	ventilator extinction	set alarm limit of MV
		end VCV
		set PCV mode
		set inspiratory pressure
		set vent. frequency
		start PCV
		quit alarms
		switch to manual vent.
		read aloud minute vent.
		show emergency O2 supply

Table 2: These are the tasks through which the usability was tested by the three teams in operation order.

essential performance of critical care ventilators [18]. These tasks were completed during seven different scenarios, which were meant to mirror operating procedures in clinics. They were designed by authors and vetted through by other RTs to ensure the accuracy. Both the tasks and scenarios (typical clinical scenarios as well as time-sensitive scenarios) were presented in the same order to all participants, aside from randomized alarm tasks. Furthermore, every task had a 10-minute time limit, in which it was to be completed or else counted as a failure. This amounted to a total of 160 minutes per ventilator for every

participant.
To increase realism, exploration-based training was used in the study [11]. This meant before the start of the study, the participants were given time, to familiarize themselves with each ventilator. Throughout the execution of the tasks, an administrator was available to answer questions and demonstrate functions.

Marjanovics team gave eleven tasks to their sampled group. Four of those were mainly dedicated to monitoring and the other seven to setting, creating two testing groups. In each of the two groups, the tasks were randomized, and completion time per tasks was at a maximum of 120s. Everything above, as well as a wrong response or task abandonment, were counted as failures.

Spaeth was in the unique position of having NOs incorporated in his study. The NOs were, of course, completely unfamiliar with the handling of critical care ventilators and so had to be given a tutorial on the day of examination. The tutorial included the setting of ventilation parameters and alarm limits, interpretation of display information and ventilation curves, breathing-circuit principles, as well as how to switch from mechanical to manual ventilation. Teaching was performed using non-specific schemes of ventilators and was conducted in the same manner for each NO.
All operators were then able to inspect each ventilator for a maximum of one minute. After that, they all performed 20 tasks, that were designed to simulate typical operating steps. The ventilators were presented to the operator in a randomized order and tasks were given verbally by the investigator. These tasks were all read aloud word by word, so that the testing conditions for all operators would be the same. The operators where then able to perform them at their own discretion. In order to avoid premature completion, tasks that contained numbers had them changed.
No time limit was set for task completion. The operators were asked, to complete the tasks quickly, but a task would only be considered complete at the operators decision. The investigator could be asked assistance of a total of three times during a task. If an operator would choose not to, but face obstacles, assistance was given every 60 seconds (up to 180 seconds).

2.5 Measurements

The three major measures of interest for Moritas team were use safety, system usability and workload. These variables were measured through observation on the investigators and verbal telling on the operators part.
Use safety was measured with the proportional inverse of tasks in which a participant had a use error or close call (UE/CC). The UE/CCs were collected through already well-established observations techniques for the usability of medical devices [19]. In this case, use errors are defined as an action or inaction, that directly leads to a compromise in safety or undesirable/unintended treatment of a patient. Close calls on the other hand are instances of usability issues, that are recovered from in time by the user. UE/CCs were observed and documented by two human factors experts during the execution of the tasks. After completion, the observers compared the ratings with each other and clarify any

issues. If agreement could not be reached, a third human factors professional would resolve it through video review.
System usability was evaluated through the UE/CC metric, as well as a Post-Study System Usability Questionnaire (PSSUQ) [10]. The PSSUQ is a 16-question evaluation for participants of usability tests. The topics described through questions can be given scores from 1 to 7 (lower number indicates better score).
Lastly, Moritas team evaluated workload using the National Aeronautics and Space Administration Task Load Index (NASA-TLX) [6]. It is a workload assessment divided into six sub-scales: Mental, Physical and Temporal Demand (dependent on users perception) and Own Performance, Effort and Frustration (dependent on interaction between the subject and task) [12]. It already has been used extensively in evaluation of healthcare devices. The output from the instrument is a score between 0 and 100, with lower scores being perceived as less workload [13].

Marjanovics team used the System Usability Scale (SUS) to measure the devices usability. It measures a devices effectiveness, efficiency and satisfaction through a 10-question evaluation with scores on a five-point Likert scale. The positions of the negative items are then subtracted from 1 and the positives from 5, after which the sums of the resulting scores are multiplied by 2.5 leaving an overall number score from 0 to 100 (highest score meaning easy to use) [2]. Like Moritas team, Marjanovic also employs the NASA-TLX metric to measure the mental workload of the subjects. Physiological parameters were also recorded during this study. Through a eye-tracking system (SMI ETG 1, SensoMotoric Instruments GmbH, Teltow, Germany) changes in pupil diameter were recorded and a biometric belt (Hexoskin, Montréal, Canada) was used to measure heart and respiratory rate as well as thoracic volume variations. Whenever one of these systems triggered, data were recorded and later used for evaluation by numerical integration of the triggers.

Spaeth also incorporated an eye-tracking system into his data collection process. A pair of eye-tracking glasses (Tobii Pro Glasses 2, Tobii AB, Stockholm, Sweden) were used to measure the gaze direction of the operators by sensing infra-red light reflected from the pupil. The gaze points and gaze fixation were then mapped on a two-dimensional image of the respective ventilators interface. This was used to determine causes of confusion in tasks, that took the operators longer to complete. These tasks were calculated by the difference between the start of vocalization of the task and the time to first fixation, which was illustrated through a heat map.
Further usability was evaluated through the SUS questionnaire, that Marjanovic also employed. The participants were asked to immediately fill out the questionnaire after test completion for each ventilator as to not give them much time to think about the tasks.

2.6 Data Analysis

The sessions for each of Moritas participants lasted a maximum of eight hours, during which their task completion is being observed and documented by the UE/CC metric. After completion of the last training period, the participants then filled out the PSSUQ and NASA-TLX questionnaires and lastly, they voiced their opinions in an interview.
The statistical analysis was performed using the statistics software SPSS Version 22.0 (IBM Corp, Armonk, NY, USA). Differences in the ventilators were explored through repeated measures analysis of variance (ANOVA) [7] and any two ventilators were then compared with pairwise post-hoc t tests [9]. The results then underwent Bonferroni corrections, however, in studies of this type, Bonferroni corrections can be overly conservative, which is why both, the results with and without corrections was considered [16].

Marjanovic categorized his evaluation process into four dimensions. The first dimension, tolerance to error, was evaluated through the completion of the tasks. This was the primary reason Marjanovic choose to primarily focus on NOs, as skilled physicians would have little to no problems operating easy-to-use devices. The second dimension are the bench testings, which are for exploration of the technical determinants of efficiency: tidal volume accuracy, pressurization accuracy, triggering, and asynchrony management . Then as the third dimension, there is the efficiency evaluation, which was the purpose for including the eye tracking system. Lastly, engagement during use of the device had to also be evaluated through psycho-cognitive scales and physiological parameter measurements.
In this case, parameters were calculated over 10-20 cycles. To calculate the error, the mean ± standard deviation of the calculated parameters was used and subsequently the median ± interquartile for the evaluation of the dimensions precision. Nonparametric Friedman and Wilcoxon signed-rank test as well as analysis of variance (ANOVA) were used for data comparison. The statistical analysis was then performed using the software MedCalc 12.7.4. for Windows (MedCalc software, Ostend, Belgium).
For the ergonomics evaluation Marjanovic used the reference device (Carefusion Avea), which also had the best success rate, followed by the PB980.

To compare the results between ventilators for his study, Spaeth opted to using the chi-squared test and as a post hoc for comparison of assistances Fisher's exact test [8]. Descriptive statistics were used for analysing processing times and TFF, and Friedman's test [15] followed by Dunn's multiple comparison test [4] were applied to compare the results of the SUS questionnaires.

3 Results of the Studies

3.1 Ventilator Comparison

Morita has summarized his results from ventilator analysis in a table. This includes the percentage of tasks with UE/CCs, perceived workload as deducted

by the NASA-TLX scale, and the mean and standard deviation for the system usability calculated by the PSSUQ. Through ANOVA, statistically significant differences could be seen on all three variables for the different ventilators. After being presented with the results, each ventilator was compared to one another, after applying the Bonferroni correction, based on the three metrics mentioned above. The SERVO-U outperformed each other ventilator in terms of better perceived usability, as well as outperforming two ventilators in safety and lower workload. The G5 was the only other ventilator having statistically significant advantages in the better perceived usability and lower workload categories in its comparison with the PB980. Morita displayed this in two different tables, with one showing the winning ventilator for each of the three metrics and the other detailing the contrast of the post hoc t tests with and without the Bonferroni correction.

Marjanovic's teams categorized their results in four segments, that represent the technical determinants of efficiency, as mentioned.

In terms of tidal volume, significant differences could be measured between the ventilators. All tested devices, except the S1, were within the 10% error range, with the PB980 having the lowest error but being outperformed by the SERVO-U in terms of precision response, of witch it was statistically the best. The V500 and V680 also had a relatively low error but suffered in the response to respiratory mechanics modifications.

The data for pressurization accuracy shows, that high positive expiratory pressure (PEEP) accuracy was similar between all devices, except for the V500. Concerning the 10% error range for pressure support, Marjanovic seems to contradict himself. In his description of the measured data it is stated, that three of the six ventilators, the SERVO-U, PB980, and S1, are over the 10% error range, although in his modelling of the box-plot for pressure support variation as well as in his description of said box-plot it can be observed, that the PB980 does not exceed the 10% error range.

As already mentioned, the data for the three aforementioned analysis criteria (tidal volume variation, PEEP variation, and pressure support variation) was modelled in box-plots for each one of the ventilators. For each of the three aforementioned criteria the data was modelled in six box-plots, one for each ventilator. Here again Marjanovic makes a mistake by mislabelling the data on each container for the V500 as data for the V300, which was not included in his study.

The next segment is the evaluation of the triggering. In terms of inspiratory triggering, no differences could be observed between the devices and triggering delay was also below 150ms for all devices. Every device, except the PB980, triggering delay exceeded 200ms in obstructive conditions. Here, the biggest difference was in the triggering pressure, with the performance of the devices all varying from one another. The S1 and R860 had the highest triggering pressure, meaning the maximal pressure drop to trigger inspiration is higher for those devices. This data was displayed using three graphs per ventilator, each one detailing the normal, the restrictive, and the obstructive respiratory mechanics through a curve.

The last of the technical determinants of efficiency is the asynchrony management. Here, the mean asynchrony indexes under standard ventilatory modes were at 31%. The non-invasive modes recorded a lower mean at 14.5% for all devices. Overall, the R860 and SERVO-U superseded the other ventilators wile using non-invasive ventilation algorithms with the only asynchrony indices under 10%. This data is also illustrated through box-plots, two for each ventilator to display standard and non-invasive asynchrony indices.
Additionally, Marjanovic displayed an ergonomics evaluation of each ventilator in a table. The evalutaion concludes, that in terms of time to power on, the Servo-U, although exceeding in most other areas, was proven to be the worst, with a total of 36% of gathered participants being able to power the device on. Subsequently the Servo-U had the highest task failure rate of all the ventilators. On all the tables and graphs mentioned in this section, Marjanovic labelled the data corresponding to one of the ventilators as data for the V300. Seeing as this ventilator was not included in his study and all data from the tested and similarly named V500 is not displayed, it can be concluded, that this was a labelling error and all data labelled as such is corresponding to the V500. All of the information written in this section that mentions the V500 ventilator is gathered from the graph labelled V300.

Spaeth measures his differences between ventilator performances in a graph, where he details the frequencies of assistances given. Every action shows two bars for the NOs and experienced operators respectively. In the graphs it is noticeable, that for two of the four tested ventilators, the Perseus and Primus, the EOs required no assistance. Respectively, the NOs also had a low assistance frequency for those two ventilators, relative to the other two. Also notable is, that for the Flow-i and Avance, the EOs required even more assistance in some cases, than the NOs.
Overall, the number of falsely executed tasks was low, with the Flow-i having the highest failure rate at 20 out of 520 tasks falsely executed. For the processing times of the individual tasks, the data was displayed through box-plots. Here, the Perseus and Primus again outperform the Avance and Flow-i, like with the assistance frequencies.
Other than the other two studies, Spaeth also used an eye tracking device. The resulting data was displayed through a table, as well as a heat map on a picture of each device. The eye-tracking-data affirmed the processing time data for the ventilators.
Lastly, the data evaluated by the SUS score reflected the above mentioned data by affirming the Perseus and Primus for best usability for the NOs and EOs respectively.

4 Discussion

Although the different teams had varying samples of ventilators they choose to incorporate in their study, the methods and data gathering are not altered by the taken samples. Only Morita's team mentions the reason behind sampling

Used statistical procedures	**Morita**	**Marjanovic**	**Spaeth**
PSSUQ	X		
NASA-TLX	X	X	
SUS		X	X
ANOVA	X	X	
UE/CC	X		
chi-squared + Fisher's exact test			X

Table 3: This table details the metrics used for the ventilator comparison and overlaps in usage between the groups.

from the ventilators as they did, with those being the market leading devices during time of the study in their region, it can be assumed, that the other teams followed the same principle, seeing as the point of the study is to determine the best ventilator for usability and user experience.

4.1 Evaluating Samples

The first major differences in performing the study can immediately be observed with the chosen participants.
Morita effectively had two groups of participants: one group of thirteen he chose to perform a pilot study with, to have a basis of knowledge for following samplings, and a group of a total of 48 experienced respiratory therapists. His reasoning for choosing RTs being, that they are the primary users of critical care ventilators. The only concern when gathering the participants was, to not allow any with ties to the manufacturers of the four chosen ventilators.
Following the same thought process, Marjanovic also used experienced participants in the form of ICU physicians.
allow any with ties to the manufacturers of the four chosen ventilators. Following the same thought process, Marjanovic also used experienced participants in the form of ICU physicians. Spaeth however, opted for completely naive operators with no prior experience with critical care ventilators, using a few experienced participants as contrast. This lead to a completely different preparation, execution and data evaluation process. This lead to a lengthy preparation phase, where each participant had to be familiarized with the ventilators first but it also allowed for a more raw and completely unbiased evaluation, which the other two groups could not guarantee fully, because of the inherent skill, the EOs posses through their trade as well as undisclosed familiarities with similar devices.

In addition to that, when gathering participants, Spaeth incorporated more precautions on established metrics, to make sure, the participants are eligible for the study, which the other groups either did not do or record.

4.2 Evaluating Demographics

Two of the three teams paid attention to the demographics of their chosen participants. Morita and Spaeth both recorded the gender and age of the participants, although it was concluded, that it had no significant effect on the gathered data, nor had Morita's participants with prior knowledge of some of the devices.

4.3 Evaluating Tasks and Scenarios

Before the tasks were given, Morita's group of participants were given a little time to familiarize themselves with each ventilator, before then receiving the tasks and having to complete them in a ten-minute time limit. An operator was also present to answer questions that might arise. Marjanovic was able to have a much more concise time limit of 120s per task, because of his gathered sample of skilled ICU technicians. This time limit paired with the use of a lung simulator allowed him to more closely mirror real critical conditions in his study. Completely to the contrary, Spaeth did not give his participants a time limit. However, he still recorded the processing times of each task completion, unlike Morita, who used his limit only to have another parameter for task failure.

4.4 Evaluating Measurements and Data Analysis

As there are already well documented metrics for the evaluation of usability, it is not uncommon to see the same metrics used by different studies. Such is the case with the NASA-TLX, for workload evaluation, and the SUS Questionnaire, to measure device usability.

Both, Morita's and Marjanovic's teams used the NASA-TLX to evaluate the mental workload of their chosen devices. Spaeth's team evaluated the mental workload of their devices through data gathered by the eye-tracking device, more specifically through the recorded time of first fixation. For the system usability, Spaeth used the mentioned SUS Questionnaire, also used by Marjanovic, which the participants filled out after completing all tasks on a ventilator. Morita's participants were also given a questionnaire after completion, although he chose the PSSUQ instead of the SUS, which has more items. Furthermore, Morita used the UE/CC metric based on well-established observation techniques as well as a post test interview, to further evaluate the usability of his chosen devices.

The resulting data were used by all teams to directly compare the ventilators. Morita also modelled the data in graphs for further visualization.

4.5 Evaluating Ventilator Comparisons

Through different means of data gathering, the teams also have different means of data display.

As already mentioned, Morita displayed his data based on the UE/CC, PSSUQ and NASA-TLX metrics as well as the data calculated in the post-hoc test (both with and without Bonferroni correction) in a table comparing

each ventilator with one another. These tables were his only means of data display, as it was his main focus to evaluate the differences in usability based on these metrics.
Marjanovic uses a combination of box-plots and graphs to display data. His team focused on the technical determinants of efficiency as an area of comparison. If the data is a percentage, such as it is the case with tidal volume accuracy, pressurization accuracy and asynchrony management, box-plots, and for time measurement of triggering evaluation,a function-graph is used as the form of display.
Although box-plots are the common means of visualizing percentage data, Spaeth instead opted for bar-graphs to display the frequency percentage of assistance given during the tasks. As his primary focus was to compare the NOs data to EOs, the bar-graphs offer a better visualization for individual tasks. His team also had the eye-tracker as a unique form of measurement. The data for time of first fixation is displayed in a table, comparing each ventilator and the raw eye-tracking heat-map is displayed in images of the ventilators.

5 Conclusion

As stated through the three different studies, there are multiple different means and forms of measurement one can incorporate in order to evaluate usability and user experience on different devices.

Morita and Marjanovic chose to test the usability with experienced operators and tried to mirror real-life conditions as close as possible. Marjanovic even included a lung simulator and had his participants perform the tasks under concise time restraints. And although Spaeth chose to go into a different direction by utilizing naive operators, his comparisons with the few experienced technicians showed, that comparative results do not deviate much from the groups.

From established usability metrics to self calculated data there are many things needing to be weighed when conducting studies to evaluate a devices usability depending on what the conductor is looking for. But no matter the differences in choice between these methods and metrics, they all seem to harken back to one of the defining components of usability. As long as one follows the components by their choice of methods, one will yield distinguishing results that are applicable to individual devices.

References

[1] AARC. What is an rt? ; https://www.aarc.org/careers/what-is-an-rt/, 08.2021.

[2] John Brooke. Sus: A retrospective. *Journal of Usability Studies*, 8(2):29–40, 2013.

[3] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Erlbaum, Hillsdale, NJ, 2. ed. edition, 1988.

[4] O. J. Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 2012.

[5] Laura Faulkner. Beyond the five-user assumption: benefits of increased sample sizes in usability testing. *Behavior research methods, instruments, & computers : a journal of the Psychonomic Society, Inc*, 35(3):379–383, 2003.

[6] Sandra G. Hart and Lowell E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In Peter A. Hancock and Najmedin Meshkati, editors, *Advances in Psychology : Human Mental Workload*, volume 52, pages 139–183. North-Holland, 1988.

[7] Hae-Young Kim. Analysis of variance (anova) comparing means of more than two groups. *Restorative Dentistry & Endodontics*, 39(1):74–77, 2014.

[8] Hae-Young Kim. Statistical notes for clinical researchers: Chi-squared test and fisher's exact test. *Restorative Dentistry & Endodontics*, 42(2):152–155, 2017.

[9] Michael H. Kutner, Christopher J. Nachtsheim, John Neter, and William Li. *Applied linear statistical models*. McGraw-Hill/Irwin series Operations and decision sciences. McGraw-Hill Irwin, Boston, Mass., 5. ed. edition, 2005.

[10] James R. Lewis. Psychometric evaluation of the pssuq using data from five years of usability studies. *International Journal of Human–Computer Interaction*, 14(3-4):463–488, 2002.

[11] Regina Lipkens and Steven C. Hayes. Producing and recognizing analogical relations. *Journal of the Experimental Analysis of Behavior*, 91(1):105–126, 2009.

[12] Nicolas S. Marjanovic, Agathe de Simone, Guillaume Jegou, and Erwan L'Her. A new global and comprehensive model for icu ventilator performances evaluation. *Annals of intensive care*, 7(1):68, 2017.

[13] Plinio P. Morita, Peter B. Weinstein, Christopher J. Flewwelling, Carleene A. Bañez, Tabitha A. Chiu, Mario Iannuzzi, Aastha H. Patel, Ashleigh P. Shier, and Joseph A. Cafazzo. The usability of ventilators: a comparative evaluation of use safety and user experience. *Critical care (London, England)*, 20:263, 2016.

[14] Nielsen Norman Group. Usability 101: Introduction to usability, 21.09.2021.

[15] Dulce G. Pereira, Anabela Afonso, and Fátima Melo Medeiros. Overview of friedman's test and post-hoc analysis. *Communications in Statistics - Simulation and Computation*, 44(10):2636–2653, 2015.

[16] T. V. Perneger. What's wrong with bonferroni adjustments. *BMJ*, 316(7139):1236–1238, 1998.

[17] J. Spaeth, T. Schweizer, A. Schmutz, H. Buerkle, and S. Schumann. Comparative usability of modern anaesthesia ventilators: a human factors study. *British journal of anaesthesia*, 119(5):1000–1008, 2017.

[18] Technical Committee: ISO/TC 121/SC 3. Iso 80601-2-12:2020; https://www.iso.org/standard/72069.html, 02.2020.

[19] Michael E. Wiklund, Jonathan Kendler, and Allison Y. Strochlic. *Usability testing of medical devices*. CRC Press, Boca Raton, second edition edition, 2011.