Applied Ergonomics 78 (2019) 76–85

Contents lists available at ScienceDirect

Applied Ergonomics

journal homepage: www.elsevier.com/locate/apergo

A usability assessment of riding lawn-mowing equipment with varying levels of design standards compliance

Yulin Deng[a], James Shirley[a], Tyler Rose[a], Laura Geary[a], David Feltner[a], Karen Chen[a], Jeffery Hoyle[b], Mohini Dutt[b], David B. Kaber[c,*]

[a] Edward P. Fitts Department of Industrial & Systems Engineering, North Carolina State University, 400 Daniels Hall, 111 Lampe Dr., Raleigh, NC, 27695-7906, United States

[b] The Ergonomics Center of North Carolina, United States

[c] Herbert Wertheim College of Engineering, Department of Industrial and Systems Engineering, University of Florida, 303 Weil Hall / P.O. Box 116595, Gainesville, FL, 32611, United States

ARTICLE INFO

Keywords:
Riding lawn-mowing equipment
Standards compliance
Usability assessment
Human factors
Safety

ABSTRACT

The use of riding lawn equipment (RLE) is related to a significant number of accidents every year. To provide basis for product design and enhance user performance and safety, a usability and performance assessment of modern riding lawn-mowing tractor designs and features was conducted in a real-world test environment. Five current commercially available RLEs were tested with response measures including task performance time and accuracy, physiological workload, system usability scores (SUS), and subjective rankings of RLE models. This data was used to identify sensitivity of responses to variations in RLE design features and functionality. The data was also used to assess the validity of new tractor design standard conformance tool, the RLEval methodology. This tool made comprehensive evaluation of RLE models compliance with over 70 specific design standards and was applied by human factors experts. Experiment results revealed sensitivity of all response measures to design differences among the five RLE models, except the objective workload measures. Response measures including task performance, SUSs and subjective rankings showed partial agreement with the RLEval scores. In general, the study results demonstrated a comprehensive experimental methodology for usability and performance evaluations of RLEs as well as merit of using the RLEval as preliminary method to compare design features. Some aspects of the usability experimentation and the RLEval method appear to be complementary.

1. Introduction

Riding lawn mowers are gaining increasing popularity in the U.S. due to their efficiency and accuracy in operation. Riding lawn mowers currently account for two-thirds of all mowers produced and this figure is expected to increase (Deneen and Gross, 2006). However, some researchers have pointed-out that riding mowers may be potentially more dangerous than walk-behind mowers because of their larger size and complexity of operation (Hammig and Jones, 2010). In 2016, an estimate of over 14,000 persons visited a hospital emergency room due to riding lawn mowers, also known as riding lawn equipment (RLE; Consumer Product Safety Commission, 2016). Hammig et al. (2009) noted that of the hospital visits attributable to RLEs between 2002 and 2007, 40% were due to rollovers or operators falling off of the tractor. Driving on improper slopes was a main contributor; however, poor seating design and awkward placement of tractor hand and foot controls were also identified as substantial issues. According to previous research, a number of riding lawn mower incidents have been related to design elements, selection, placement, and/or operation of power mower controls (Heasly et al., 1989). Consequently, better riding lawn mower design is expected to reduce hazard exposure and improve operator safety (Patel et al., 2000; Yadav and Tewari, 1998). This expectation was the motivation for the present empirical assessment of the design of contemporary RLE products. Although the safety issues related to riding lawn mowers have been widely discussed in earlier literature, to date a limited number of studies have systematically evaluated riding lawn mower designs. Moreover, among the studies that provide insights into the design of RLE products, few were conducted in the past decade. Given the rapidly changing technology, there is a need for up-to-date reference for contemporary RLE designs. The

* Corresponding author. The Ergonomics Center of North Carolina, Director, Occupational Safety and Ergonomics Program, Edward P. Fitts Department of Industrial & Systems Engineering, North Carolina State University, 400 Daniels Hall, 111 Lampe Dr., Raleigh, NC, 27695-7906, United States.
E-mail address: dbkaber@ncsu.edu (D.B. Kaber).

https://doi.org/10.1016/j.apergo.2019.02.003
Received 20 November 2017; Received in revised form 3 November 2018; Accepted 9 February 2019
Available online 21 February 2019
0003-6870/

Table 1
Identification of major RLE functional features and overall RLEval score.

Model	Forward pedal		Reverse pedel		Blade Enagement		Deck Height		Parking Break		RLEval Score
	Ankle Action	Leg Action	Next to forward	Behind Forward	Longitudinal	Transverse	Left side	Right Side	Auto-Disengege	Manual-Disengage	
Model 1	✗			✗	✗			✗	✗		80
Model 2	✗			✗	✗			✗	✗		78.46
Model 3	✗		✗			✗	✗			✗	82.22
Model 4	✗		✗			✗	✗			✗	82.22
Model 5		✗		✗		✗	✗			✗	75.16

present study aims to fill this research gap by providing a systematic research on the evaluation of contemporary RLEs. It is one of the first investigations to undertake an empirical assessment of RLE designs, making use of objective task performance data and subjective response measures. Results of this study are expected to provide reference for RLE manufacturers to improve safety, reliability and quality of lawn tractors.

1.1. Current evaluation frameworks

Previous studies involving riding lawn tractor evaluations have used both subjective and objective methods to measure RLE and human activity outcomes. Methods have included physiological measures, usability assessment, and performance measures.

1.1.1. Physiological and workload measures

Heart rate variability has been identified as a promising measure of operator workload and physiological strain level (Jorna, 1993; Stanton et al., 2004; Young et al., 2015). Usually lower operator workload is considered a result of enhanced system design features (Mouloua, 2018). In a study of ride-on tractor operation, a ratio of heart rate under working vs. resting conditions was used as the index of operator workload with increases during tractor use (Syuaib et al., 2003). Other research found that heart rate may not be an appropriate measure for assessing seating discomfort during tractor driving (Mehta and Tewari, 2000), motivating the use of other measures.

1.1.2. Subjective measure of usability

Usability rating scales have been commonly used in ergonomic studies (Bangor et al., 2008). The system usability scale (SUS), for example, was developed by Brooke (1996) and it provides an effective and easy to use method by which to evaluate effectiveness, efficiency and user satisfaction level with systems (Bangor et al., 2008). Different tractor designs may influence the perceived usability of a product and user experiences with the system. Therefore, the SUS may be another useful measurement tool for application in the study of RLE usability.

1.1.3. Performance

In a previous ergonomic evaluation of riding tractor designs, lane keeping error and task time were used to assess operator performance (Syuaib et al., 2003). Task time has been widely used to evaluate machinery designs and, in general, superior designs have been found to lead to shorter task completion times (Hannaford et al., 1991; Kim and Singhose, 2010). A greater number of lane keeping errors were considered as an indicator of degraded task performance.

1.1.4. RLEval

The RLEval (Deng et al., 2017) is a new methodology for assessing the degree of RLE design conformance with existing guidelines and for comparison of various products. This tool scores modern RLEs in terms of a range of design features on the basis of existing ISO and ANSI safety, usability, and functionality standards, including ISO 15077, ISO 15079, ISO 4253, and ANSI 71.1. The tool yields an RLE usability classification result (Deng et al., 2017). The standards under examination contained 74 different design guidelines for assessing five categories of device functionality: (1) Foot Pedals, (2) General Controls, (3) Hand Controls, (4) Seating, and (5) Steering Wheel. Some examples of design guidelines include the following:

- "Provisions must be made to prevent unintentional brake release"
- "An indicator of blade rotation shall be provided on mowers"

In this study, RLE design features were both subjectively and objectively assessed relative to such design guidelines. Based on the evaluation, overall score for each model was calculated, which was determined as the total percentage of RLE features that conformed with guidelines. Subsequently, an overall usability classification is assigned based on the overall compliance of the RLE features (i.e., "Good" greater than 90%; "Fair" between 75% and 90%; and "Poor" less than 75%). In our previous study, five human factors experts applied the tool to five residential RLEs. Subjective evaluation criteria were analyzed by each expert, individually. The objective evaluation was carried out by a single group of two evaluators who conducted two measurements of each numeric criteria and recorded a mean value to determine design compliance. The models were further classified into three categories: "Good", "Fair", and "Poor" according to the overall score they received. Given its quantitative and comprehensive nature, the RLEval was chosen as an additional tool to evaluate the usability of contemporary tractors models. The overall scores of the five tractors assessed are presented in Table 1.

1.2. Objectives

The present study sought to compare contemporary RLE models through performance observations and to assess the correspondence of results with RLEval outcomes. Experiments were conducted to evaluate the usability and performance of current RLEs in mock residential use tasks. The study also provided a basis for determining the validity of the RLEval tool for future design inspections.

2. Methodology

2.1. Participants

Ten (10) participants (2 female, 8 male), age (21–64) [mean = 52.9, SD = 12.42] were recruited. Participants were either homeowners or lessees of a property and had at least 3 months of prior experience using an RLE, and no professional lawn service persons were included. These inclusion criteria were based on the study objective to understand the performance of average, residential RLE owners. This study was approved by the North Carolina State University Institutional Review Board.

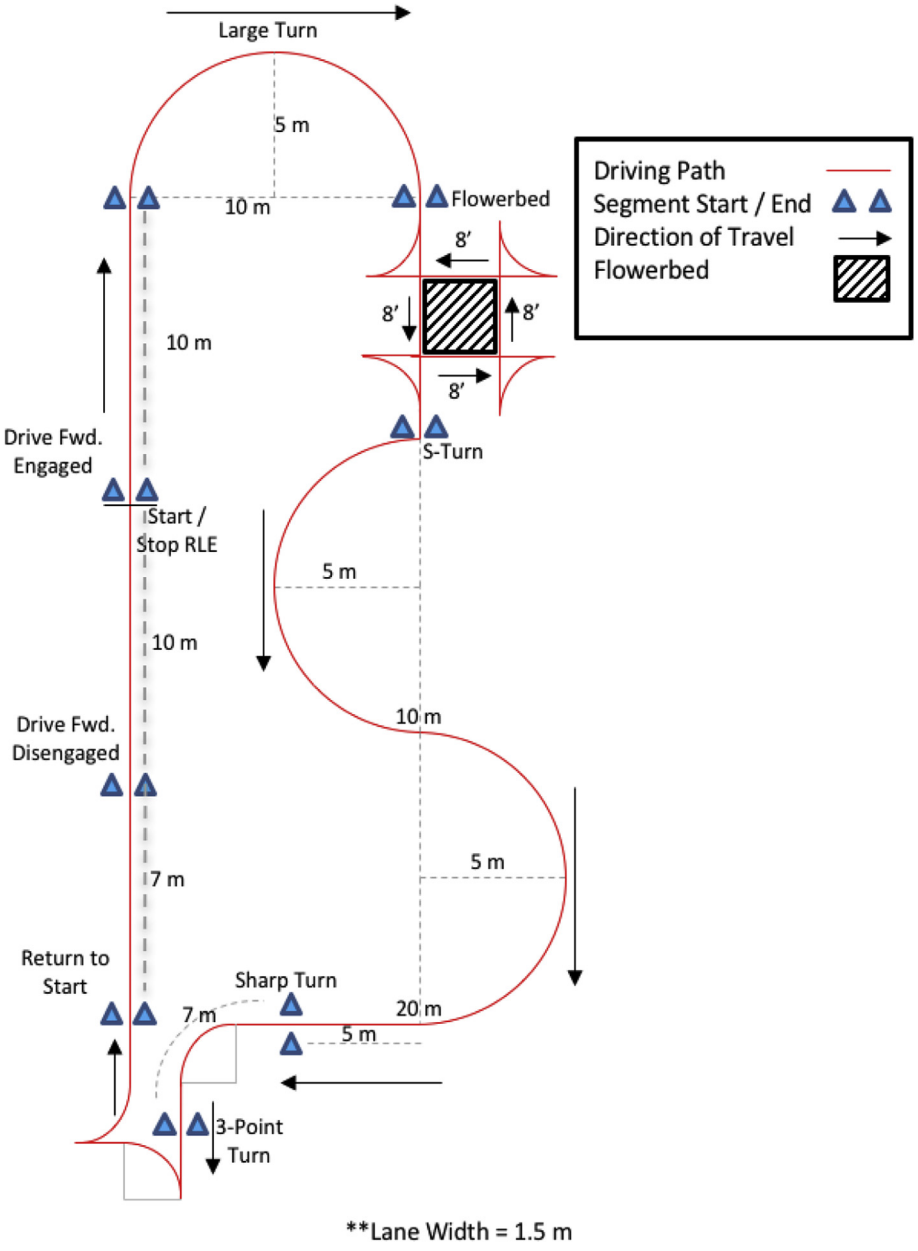


Fig. 1. Experiment track diagram.

2.2. Setup and apparatus

2.2.1. Track design

The experiment was designed around nine RLE use tasks representative of typical vehicle and mower operation. Each task was to be performed in a single segment of a large test track (Fig. 1). The specific vehicle control tasks included: 1. Turn RLE on; 2. Drive RLE straight with mower blades disengaged; 3. Drive RLE straight with mower blades engaged; 4. Large turn; 5. Flower bed maneuver; 6. SCurve; 7. Sharp turn; 8. 3-point turn; and 9. Turn RLE off. The test track was laid-out in a large outdoor parking lot. The straight segment, S-Curve, and sharp turn were designed based on a previous lawn equipment experiment (Syuaib et al., 2003). The flowerbed and 3-point turn were included to create additional tasks representative of common real-world mowing scenarios. These tasks required participants to use all five categories of device functions (i.e., foot pedals, general controls, hand controls, seating, and steering wheel). Therefore, any usability issues related to these device functions were expected to appear in task performance results. In addition, the complexity of these control tasks was intended to promote sensitivity of the performance response analysis for revealing any differences among various RLE designs.

2.2.2. Equipment

Five current and production RLE models were selected for this study and analyzed using the RLEval tool. The majority of models were selected from among the tractors sold at local home improvement retailers. We assessed RLEs made by multiple name-brand and top-rated manufacturers in order to ensure a broad range of designs. All RLEs were new.

To ensure safe testing, the track was thoroughly swept of debris. Two cameras (GoPro Inc., San Mateo, CA) were mounted above the front wheels of the tractors (one on the left side and one on the right) on each vehicle for capturing user control activity. Additionally, a Polar H7 Bluetooth Smart Heart Rate Chest Transmitter and accompanying RCX3 watch were used to capture participant heart rate activity (Polar Electro Oy Inc., Lake Success, NY). Additional participant safety

precautions included complimentary water and shelter under a canopy between test trials.

2.3. Experiment design

The experiment followed a randomized complete block design with subject as a blocking factor in order to account for variability among participants in comparison of the RLEs. Each participant performed two test trials with each RLE (i.e., the study design was replicated) and two unique 5 × 5 Graeco-Latin squares (Montgomery, 2005) were used to randomize the orders of presentation of RLEs across test trials. This approach also served to address the potential for any condition carry-over effects. The Graeco-Latin squares allowed for two initial training trials with one RLE model, one testing trial with each non-training RLEs under a unique randomization, and one replicated testing trial with each non-training RLE under another unique randomization.

Based on this experiment design, an Analysis of Variance (ANOVA) model was constructed to assess the effect of the RLE model on the various response measures. The statistical model was structured as follows:

$$X_{ijklr} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \alpha_i\delta_l + E_{ijklr} \quad (1)$$

Equation (1): Statistical model. Where μ = grand mean; α_i = RLE model effect (i = 1 … 5); β_j = participant effect (j = 1 … 10); γ_k = trial number effect (k = 1 … 8); l = track segment effect (1 … 6); $\alpha_i\delta_l$ = Interaction between tractor models and track segment; Replication (r = 1,2); E_{ijkr}: Error.

2.3.1. Independent variables

Two independent variables were manipulated including: (1) the RLE (5 models); and (2) the simulated lawn mowing task (9 tasks). Table 1 (in Section 1.1.4) presents an inventory of major design features of the RLE models and tested in the experiment. The table allows for identification of some key differences among the models. For example, the blade engagement is longitudinal for some models and transverse for others.

2.3.2. Dependent variables

The current study sought to investigate four types of response measures, including: participant workload in RLE use, perceived system usability, ranking of tractors and vehicle control task performance.

3.3.2.1. Workload. Participant heart rate (HR) data (bpm) was collected via the HR watch and accompanying chest strap (Polar Electro Oy, Inc.) throughout each trial. Participants wore the chest-band and watch during the entire experiment in order to prevent possible inconsistencies in data collection due to re-donning the chest strap. This data was time-synced to the GoPro camera recordings for analysis of the workload demands of each control task. HR data was recorded during the course of test trials. The data recording started at the beginning of a trial and ended the moment participants completed the tractor operation tasks. HR data recorded during the usability testing and resting periods was not included in any statistical analyses. Raw HR test values were normalized using participant baseline data. Prior to training and experiment trials, participant resting HR was measured. The resting HR was later used as a baseline for comparison with the average HR for each trial. The percent increase in average HR during each experiment trial was calculated relative to the resting HR and was statistically analyzed and reported.

3.3.2.2. System usability score (SUS). Perceived system usability data was collected by using the system usability questionnaire (Brooke, 1996), which was administered after each trial. The questionnaire consisted of 12 usability related questions and participants used a 5-point scale rating scale, ranging from 1 = “Strongly disagree” to 5 = “Strongly agree”, to indicate their level of agreement with each statement (see Appendix A). The ratings were used to compute an overall SUS score that were statistically analyzed. The overall SUS score was calculated as the sum of score contributions from the 12 items, as instructed by Brooke (1996).

3.3.2.3. Ranking of RLE models. Participants also ranked the five RLE models from 1 to 5 at the end of the entire experiment. Rank 1 was used to identify the most superior performing RLE. Participants were allowed to take notes on the characteristics of each RLE after each test trial, which served as a reference for them in the ranking process.

3.3.2.4. Task performance. Two response measures were used to assess participant overall task performance, including task time and accuracy.

Task time was measured as the time between the front bumper of a tractor crossing a set of stationary cones, identifying the beginning of a track segment (particular control task), to the time when the front bumper crossed the next set of cones, identifying the beginning of the next segment (and control task). Task time was determined from the GoPro camera video recordings (available for each side).

Task accuracy was measured by the number of vehicle control errors committed in each track segment. An error was recorded for every instance where the tires of a tractor touched lane markings (Syuaib et al., 2003). Markings were painted on the parking lot pavement and extended 23″ from the centerline of any tractor (The maximum mower cut deck width was 46”). The task error information was also recorded by GoPro cameras.

2.4. Hypotheses

We formulated the following hypotheses regarding RLE design conformance with existing guidelines (i.e., RLEval scores) and any relation to response measures recorded during the empirical testing:

H1. Increased RLE conformance with existing design standards will result in superior task performance, indicated by decreased task completion time, and fewer lane keeping errors.

H2. Increased RLE conformance with existing design standards will result in lower task workload, indicated by decreased work/rest heart rate ratio.

H3. Increased RLE conformance with existing design standards will result in increased SUS (perceived usability) scores.

H4. Increased RLE conformance with existing design standards will result in higher subjective performance rankings for a tractor.

2.5. Procedure

Consenting participant demographics (height, weight, age, gender, and experience with RLEs) were initially collected using a paper-based questionnaire. Participants were then instructed to don the HR monitor chest strap (Polar Electro Smart Heart Rate Chest Transmitter) and to wear the accompanying wrist watch.

To prevent any participant inspection prior to training and testing, each RLE model was covered with black tarps, and was only unveiled immediately before a training or test trial in which a participant was to make use. As part of training, all participants were provided with a brief orientation on RLE controls and functions, as well as the test track and segments. Each participant received a minimum of one training trial in which they performed the identified vehicle control tasks in the defined segments of the test track. To pass training, a participant was required to complete all nine tasks within the proper segments. Additionally, the participant was required to complete the track in under 3.5 min of driving time, based on pilot testing with experienced RLE operators. The training protocol was intended to ensure participants were all at a

Fig. 2. Screen capture from GoPro Camera Recordings.

comparable level of skill before beginning testing. If participant training performance remained unacceptable after four trials, his or her participation was terminated.

After training, all four remaining experimental RLE were evaluated according to the predetermined randomization orders for each participant. When a tractor was unveiled for testing, the GoPro cameras were attached to each side of the RLE body. Participants were given vehicle control task instructions. When ready, they began the test trials.

Fig. 2 presents two screen captures from Go Pro Camera Recordings. Only partial images of the tractors were shown to avoid revealing the make and model of RLEs.

Once complete and the RLE was turned off, the experimenters terminated the camera and HR recordings simultaneously. The participant was escorted to the shade canopy to complete the system usability scale (SUS). For replication, all four test models were tested a second time following another unique randomization order. Each trial took between 90 s and 3 min. Breaks between trials lasted approximately 2 min. In order to reduce the effect of fatigue on participant performance, prior to each trial, a 14-item fatigue questionnaire was administrated (Chalder et al., 1993). This scale has been widely used and demonstrated to be effective for assessing physical and mental fatigue symptoms (Shahid et al., 2011; Morriss et al., 1998). A fatigue score of 3 or less indicates that a person is not fatigued (Chalder et al., 1993). On this basis, no fatigue symptoms were observed for any participants.

Once all test trials were complete, participants were asked to rank order all RLE test models without any coaching or assistance by experimenters, with the models being uncovered and on display. Participants were allowed to refer to any notes that they made on each model during breaks/rest periods. The entire experiment took about 2.5 h per participant.

2.6. Statistical analysis

All data was screened for any outlying observations attributable to equipment issues or participant failure to follow instructions. Diagnostics were conducted on all response measures to assess normality assumption (assessed by Shapiro-Wilk's test) and constant variance assumptions (assessed by Bartlett's test) of the ANOVA model. In the case of assumption violations, data transformations were applied. If unsuccessful, response observations were ranked and a nonparametric equivalent of the ANOVA was conducted.

ANOVA models were developed to test the effect of the RLE type and control task on the various categories of dependent variables, with a significance level of $\alpha = 0.05$. Trial number was initially included in the ANOVA models as a co-variate, but it was not significant in any response, thus it was removed from the model. When appropriate, Tukey's post hoc tests were conducted to identify differences among the RLE models and task types.

Among the response measures, the HR measure, system usability score, and ranking data met the parametric test assumptions. All subtask times, save the Large Curve data, satisfied the ANOVA normality assumption. Subtask accuracy data did not meet test assumptions and transformations were unsuccessful. The response observations were, therefore, ranked for nonparametric analysis.

3. Results

The RLEval results for the various tractor models were determined in an earlier phase of this research project and have already appeared in print (Deng et al., 2017). Below, we present the overall scores for each of the models tested in the present study and provide a general classification of the models based on the scores (Fig. 3). According to our

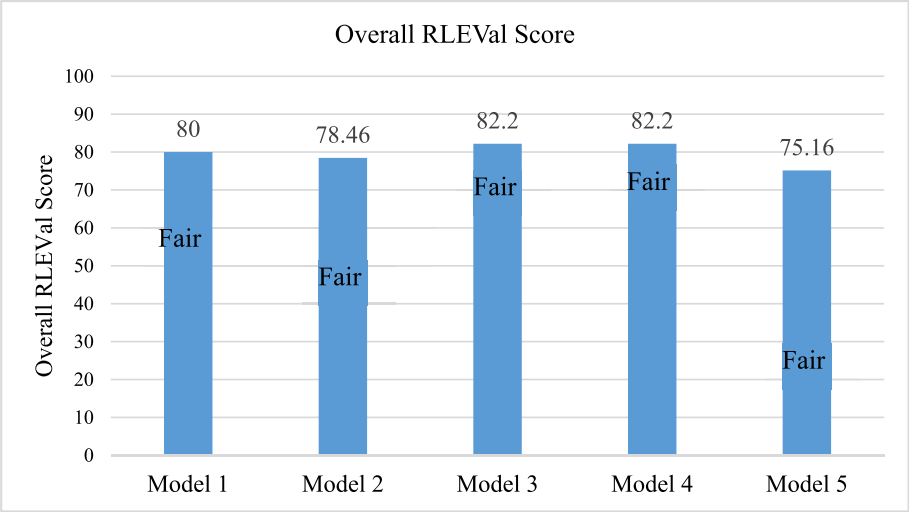


Fig. 3. Overall RLEval score and classification.

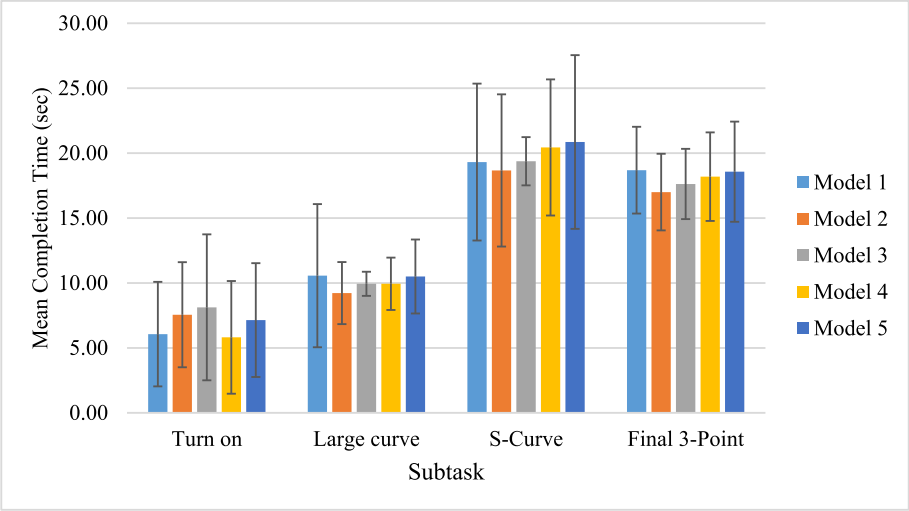


Fig. 4. Effect of RLE model on subtask completion times.

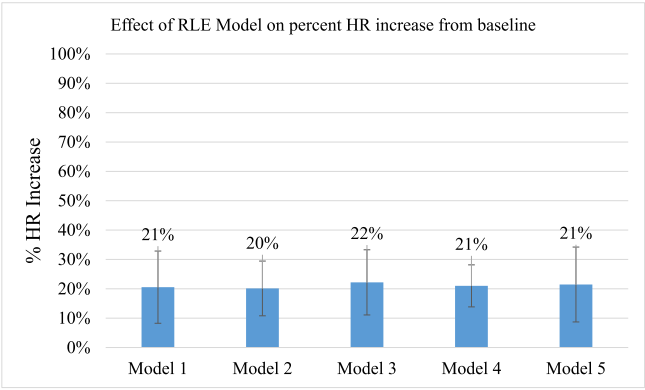


Fig. 5. Effect of RLE Model on percent Heart Rate Increase from Baseline.

previous assessment, all of the five RLE models received "Fair" classification.

3.1. Task performance

3.1.1. Task time

Among the nine subtasks observed in the experiment, four were significantly affected by the RLE model, including: Turn On the tractor ($F(4, 79) = 3.50$, $p = 0.012$); Large Curve negotiation ($F(4,79) = 8.96$, $p < 0.001$); S-Curve negotiation ($F(4,79) = 11.90$, $p < 0.001$); and 3-Point Turn ($F(4,79) = 2.70$, $p = 0.038$). Fig. 4 presents the trend of the RLE model effect for these tasks.

Turn On. Post-hoc comparison Tukey's HSD test to least square means (LSMeans) indicated that participants using Model 2 took significantly longer to turn the mower on than Model 4.

Large Curve. Tukey's HSD test indicated that participants using Model 3 took significantly longer to complete the Large Curve segment than Model 1 or Model 2. Additionally, participants using Model 4 or Model 5 took significantly longer than Model 2.

S-Curve. Tukey's HSD test indicated that participants on the Model 3 took significantly longer to complete the S-Curve segment than Model 1 or Model 2. Additionally, participants using Model 4 or Model 5 took significantly longer to complete this section than Model 2.

3-Point Turn. Tukey's HSD test indicated that participants on Model 5 took significantly longer to complete the 3-Point Turn segment than Model 2.

3.1.2. Task accuracy

Task accuracy was measured in terms of the number of errors committed during each vehicle control subtask. Since the Flowerbed subtask consisted of 4 identical 3-point turns, the subtask was analyzed at a 'per turn' level. Nonparametric analysis was conducted using the Wilcoxon ranked sums test. Among all the subtasks analyzed for accuracy in performance, only two were significantly affected by the RLE model, including: Flowerbed negotiation ($\chi^2_4 = 9.61$, $p = 0.048$) and S-Curve negotiation ($\chi^2_4 = 14.55$, $p = 0.006$).

Flowerbed: Pairwise comparisons using the Wilcoxon Method indicated Model 1 to produce significantly fewer errors than Model 2 and Model 3. Additionally, Model 5 produced significantly fewer errors than Model 2.

S-Curve. Pairwise comparisons using the Wilcoxon Method indicated that Model 3 and Model 4 produced significantly fewer errors than Model 1 and Model 2.

3.2. Workload

The RLE model was not found to have a significant effect on the percent increase in normalized HR ($F(4,79) = 0.09$, $p = 0.98$) from baseline to tractor testing (Fig. 5).

3.3. System usability score

The main effect of RLE model was found to be significant ($F(4,79) = 4.92$, $p = 0.0197$) in the system usability score (Fig. 6). Tukey's post-hoc test revealed Model 2 to produce the highest usability score and Model 5 to produce the lowest. All other scores were statistically comparable.

3.4. Ranking

The main effect of RLE model was found to be significant ($F(4, 39) = 0.8958$, $p = 0.0409$) on the average ranking response (Fig. 7). Tukey's post-hoc test revealed Model 4 to produce the highest average ranking (superior performance) compared to all other models, while Model 5 had the lowest average ranking. All other rankings were statistically comparable.

3.5. Correlation analyses

To provide a basis for evaluating the various research hypotheses, it was also necessary to conduct correlation analyses on the RLEval scores

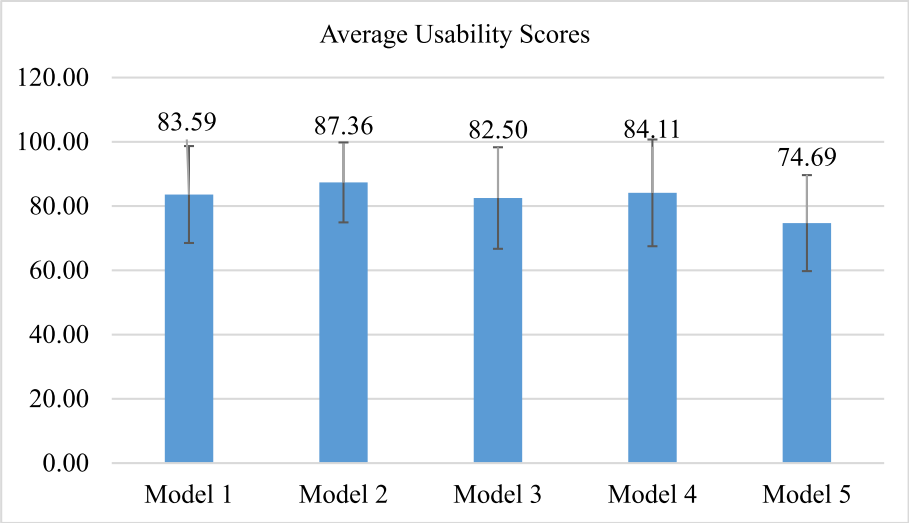


Fig. 6. Average system usability score for RLE models.

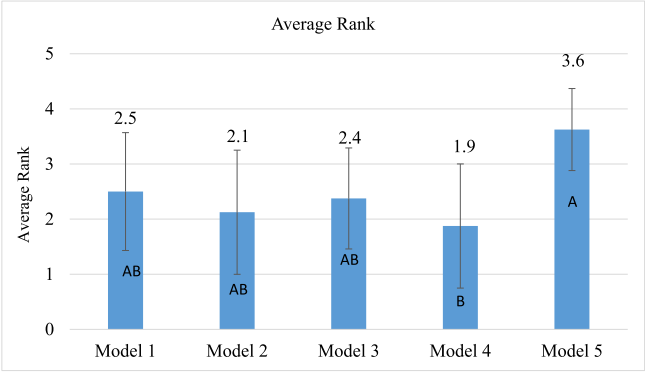


Fig. 7. Average ranking of RLE models.

for each of the tractor models with the various observational responses that were collected during the experiment and that were found to be significantly influenced by the RLE model. This set of experiment responses included: Subtask time for Turn On, Large Curve, S-Curve and 3-point turn; subtask accuracy for Flowerbed negotiation and S-Curve; the SUS score; and the RLE ranking.

Since the number of observations for each response were small, non-parametric correlation analyses (Spearman's rho) were applied to the data.

Ranking. Results revealed significant negative associations between the RLEval score and ranking ($r = -0.3893$, $p = 0.013$). The results showed that higher RLEval scores are associated with lower ranking values, which indicates superior performance.

Subtask Errors. Results revealed significant negative associations between the RLEval score and subtask accuracy for S-Curve ($r = -0.3921$, $p = 0.0123$). The negative correlation indicates that as the RLEval score increased, there was a tendency for fewer S-Curve errors.

All other potential possible correlations were insignificant.

4. Discussion

4.1. Task performance

According to the pairwise statistical comparisons, there were substantial performance differences among the RLE models. Trends indicated Model 2 to outperform other models in terms of task completion time. However, this finding was counter to our task time expectation (H1), as Model 2 also produced a lower RLEval score, as compared to Model 1, Model 3, and Model 4. Furthermore, the correlation analyses on the RLEval scores and subtask times were not significant. Regarding task errors, there was some evidence to suggest that the Model 3 and Model 4 outperformed Model 1 and Model 2. Related to this, the correlation analyses on the RLEval scores and subtask accuracy levels were significant for the SCurve subtask, indicating that models with higher RLEval scores tended to produce fewer errors. Therefore, the findings on task errors were in-line with the expectation (H1) that greater design guideline conformance would be associated with superior performance.

In general, it appeared that there may have been a speed-accuracy tradeoff in some of the RLE performance outcomes. This tradeoff was most clearly demonstrated in the SCurve task responses. Model 2 produced a faster average time than Model 3 and Model 4, but also made more errors than those models. This suggests that participants were more focused on completing the task quickly while using Model 2 rather than focusing on completing it accurately. In this respect, participants using the two models with the highest RLEval scores, Model 3 (= 82.22) and Model 4 (= 82.22) generally performed as well as, or better than, the other models in regard to task accuracy.

4.2. Workload

Results of the experiment revealed no significant effect of RLE on changes in HR from post-experiment at-rest baselines to test data collection while participants were riding on the RLEs, thus the hypothesis about workload (H2) was not supported. A possible reason for this finding is that any differences among the functional features of the RLE models was not substantial enough to alter user physiology or stress state in operation. This finding speaks to the general level of competitiveness in the current designs of RLEs. Moreover, given that all participants were experienced users of RLEs, the test track workload might not have been sufficiently high for stress state differences to emerge in vehicle control. Consequently, no correlation analysis was conducted on the RLEval scores and the workload responses.

4.3. System usability score

The experiment results revealed the SUS scores to be a sensitive measure for identifying usability differences among the RLE models. However, the experiment results were not supportive of our expectation (H3). As might be expected, the correlation analysis on the RLEval and SUS scores showed no significant associations. Model 3 and Model 4 produced the highest RLEval scores but Model 2 produced the highest

SUS score. Model 5 produced the lowest SUS score, which was consistent with its RLEval score. The reason for these findings is likely that users considered various aspects of tractor design in making usability scores that are not addressed by the design standards covered in the RLEval methodology. According to the users' comments, factors beyond the common design standards (e.g., noise, position of cup holders, smoothness of motion) might also affect their usability evaluations. This finding is actually useful in that the RLEval and SUS scores may be complementary in nature when applied for assessing the designs of RLEs.

4.4. Ranking

Hypothesis 4 was partially supported by the study results. The RLEs with highest (Model 4) and lowest rankings (Model 5) also had highest and lowest design standard conformance respectively, according to the RLEval method. It is worth noting that Model 3 and Model 4 produced the same level of design standard conformance but Model 3 was ranked lower in terms of performance by experiment participants. Related to this, the correlation analysis on the RLEval scores and the tractor performance ratings showed significant association. This again could be attributed to the focus of the RLEval methodology on specific design features covered by the standards and not all the features that users might have considered in making performance ratings.

4.5. Participant comments

As mentioned in the experiment procedure, participants were instructed to write down their comments on the advantages and disadvantages of each RLE model as part of the usability assessment process. These comments were expected to reveal features participants most likely considered during the usability evaluation process. Among the various features of RLE models, totally five features were mentioned in participant comments, including: (1) steering wheel, (2) forward/reverse acceleration pedals, (3) seat, (4) brake pedals, and (5) mower engagement/disengagement function. Among these features, the forward pedal (mentioned by 10 participants) and reverse pedal (commented by 9 out-of 10 participants) were most frequently mentioned. Half of participants (5) mentioned the brake pedal and steering wheel. Seating and mower engagement/disengagement function were each commented by one participant. Although these comments were specific to the RLE models tested and the required control tasks, it can be generally concluded that the features most frequency mentioned were more likely to influence participant subjective assessments of usability. Therefore, these features should be a focus of manufacturer future design enhancements in order to improve user experience.

5. Limitations

Although some of the response measures observed in the present experiment appeared sensitive to the RLE and task manipulations, including some statistically reliable effects emerging for various types of responses, the small sample size for the study likely limited the range of results. A larger sample size might reveal additional differences in measures among RLEs (e.g., HR) and certainly further promote the generalizability of results to the broader RLE user population.

Moreover, we only tested five RLE models across four different manufacturers. Given the variety of available commercial RLE models and their various features, there is a need to test a broader range of vehicles in order to develop a clearer understanding of how current design guidelines actual relate to user performance and usability outcomes.

In addition, given the scope of this study, the effects of individual design elements were not assessed by the usability evaluation measures. The reason for this approach is that completion of the identified tasks required using multiple RLE design features, and the ranking and SUS reporting required participants to take multiple features into consideration. Detailed assessment of individual RLE design features and usability evaluation results can be a direction of future study.

Finally, the present experiment was conducted on a pave surface with little tilt instead of a lawn with grass. This approach was taken in order to ensure consistency of course conditions and limit extraneous variables, such as matting of turf, mud accumulation, etc. However, the testing track was not representative of actual mowing environment conditions. It is possible that RLE model performance differences might be greater when tested on more difficult terrain (e.g., hills, knolls, valleys, banks, etc.) than our test track.

6. Conclusions

This study empirically assessed the performance, workload and usability of RLEs in mock residential use tasks. Response measures included task performance time and accuracy, physiological workload, system usability scores, and subjective rankings of RLE models. Differences were identified among five commercially available RLE models produced by major manufacturers in terms of all response measures, except the objective workload measures. The study found that several measures were potentially complementary in evaluating various aspects of RLE design. The study also assessed the validity of the new RLEval methodology for assessing the degree of RLE conformance with existing design standards and making predictions of RLE performance and usability outcomes. Response measures including task performance, SUS scores and subjective ranking showed partial agreement with the RLEval scores, suggesting that RLE design conformance with standards is only part of a complete design usability assessment, which may also require user observations. In general, the study results demonstrated some merit of using the RLEval as an evaluation tool for RLEs, but also revealed some limitations of the method.

Future work

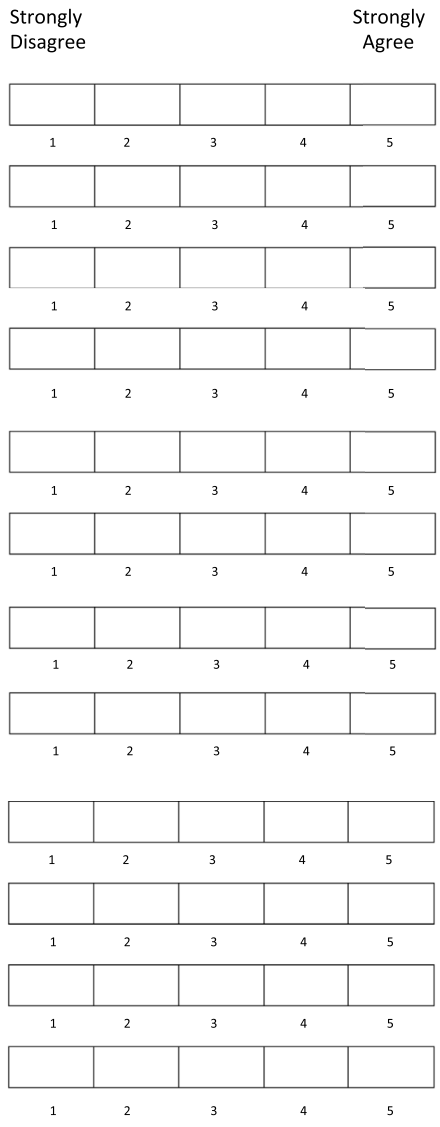
Based on the findings of the study, there appears to be a need to add additional usability evaluation criteria to the RLEval tool, possibly based on the ISO and ANSI standards for similar heavy machinery. This action might serve to increase the utility of the RLEval for tractor design assessment and usability and performance predictions. Additional experiments should be conducted the further validate the RLEval tool by making use of a larger participant sample size and a longer and more rigorous test track meant to simulate longer mowing periods and actual terrain.

Ackowledgement

This research was supported by a grant from a major RLE manufacturer and the Edward P. Fitts Department of Industrial & Systems Engineering. The opinions and conclusions are those of the authors and do not necessarily reflect the views of either of these organizations.

Appendix A

System Usability Scale



Comments:

References

Bangor, A., Kortum, P.T., Miller, J.T., 2008. An empirical evaluation of the system usability scale. Int. J. Human–Comput. Interact. 24 (6), 574–594.

Brooke, J., 1996. SUS-A quick and dirty usability scale. Usabil. Eval. Ind. 189 (194), 4–7.

Chalder, T., Berelowitz, G., Pawlikowska, T., Watts, L., Wessely, S., Wright, D., Wallace, E.P., 1993. Development of a fatigue scale. J. Psychosom. Res. 37 (2), 147–153.

Consumer Product Safety Commission, 2016. Selected Findings from the National Electronic Injury Surveillance System. (Bethesda, MD).

Deng, Y., Shirley, J., Rose, T., Feltner, D., Chen, K., Hoyle, J., Dutt, M., Kaber, D., 2017. Development of a usability and functionality assessment tool for riding lawn equipment. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 61. pp. 2015–2019 No. 1.

Deneen, M.A., Gross, A.C., 2006. The US market for power lawn and garden equipment. Bus. Econ. 41 (2), 62–67.

Hammig, B., Childers, E., Jones, C., 2009. Injuries associated with the use of riding mowers in the United States, 2002-2007. J. Saf. Res. 40 (5), 371–375. https://doi.org/10.1016/j.jsr.2009.07.005.

Hammig, B., Jones, C., 2010. Paediatric injuries incurred by being run over by a riding lawn mower: United States, 2002–2008. Int. J. Inj. Control Saf. Promot. 17 (3), 205–207.

Hannaford, B., Wood, L., McAffee, D.A., Zak, H., 1991. Performance evaluation of a six-axis generalized force-reflecting teleoperator. IEEE Transact. Sys., Man, Cybern. 21 (3), 620–633.

Heasly, C.C., Perse, R.M., Malone, T.B., Fleger, S.A., 1989. Riding mower control placement guideline development. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 33. SAGE Publications, pp. 474–478 No. 6.

Jorna, P.G.A.M., 1993. Heart rate and workload variations in actual and simulated flight. Ergonomics 36 (9), 1043–1054.

Kim, D., Singhose, W., 2010. Performance studies of human operators driving double-pendulum bridge cranes. Contr. Eng. Pract. 18 (6), 567–576.

Mehta, C.R., Tewari, V.K., 2000. Seating discomfort for tractor operators–a critical review. Int. J. Ind. Ergon. 25 (6), 661–674.

Montgomery, D.C., 2005. Design and Analysis of Experiments. S.L. John Wiley.

Morriss, R., Wearden, A., Mullis, R., 1998. Exploring the validity of the Chalder Fatigue scale in chronic fatigue syndrome. J. Psychosom. Res. 45 (5), 411–417.

Mouloua, M. (Ed.), 2018. Automation and Human Performance: Theory and Applications. Routledge.

Patel, R., Kumar, A., Mohan, D., 2000. Development of an ergonomic evaluation facility for Indian tractors. Appl. Ergon. 31 (3), 311–316.

Shahid, A., Wilkinson, K., Marcu, S., Shapiro, C.M., 2011. Chalder fatigue scale. In: STOP, THAT and One Hundred Other Sleep Scales. Springer, New York, NY, pp. 97–98.

Stanton, N.A., Hedge, A., Brookhuis, K., Salas, E., Hendrick, H.W. (Eds.), 2004. Handbook

of Human Factors and Ergonomics Methods. CRC press.
Syuaib, M.F., Moriizumi, S., Shimizu, H., Ishizuki, K., 2003. Ergonomic evaluation of ride on tractor operation between beginner and skillful operator comparative analyses of physiological strain, technical performance and viewing point. Jpn. J. Farm Work Res. 38 (3), 143153. https://doi.org/10.4035/jsfwr.38.143.
Yadav, R., Tewari, V.K., 1998. Tractor operator workplace design—a review. J. Terramechanics 35 (1), 41–53.
Young, M.S., Brookhuis, K.A., Wickens, C.D., Hancock, P.A., 2015. State of science: mental workload in ergonomics. Ergonomics 58 (1), 1–17.