Heliyon 6 (2020) e03291

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

# Evaluation of algorithmic, textual and pictorial forms of representation of standard operating procedures for error reduction in complex systems

Claudius Hammann [a,*], Christoph Krause [c], Anna Feldhütter [b]

[a] Technical University of Munich, Department of Mechanical Engineering, Institute of Nuclear Technology, Bolzmannstreet 15, 85748 Garching, Germany
[b] Technical University of Munich, Department of Mechanical Engineering, Institute of Ergonomics, Bolzmannstreet 15, 85748 Garching, Germany
[c] Technical University of Munich, Department of Aerospace and Geodesy, Institute of Flight System Dynamics, Bolzmannstreet 15, 85748 Garching, Germany

ARTICLE INFO

*Keywords:*
Safety engineering
Aerospace engineering
System engineering
Standard operating procedures
Work error
Introduction form
Human reliability
Complex system evaluation
Evidence based
Flight research simulator

ABSTRACT

A new concept has been developed to compare different ways of presenting instructions for action for evaluation procedures. The representation forms algorithm (A), image (I) and text (T) are examined with regard to the number of top events, error frequencies, execution times and subjectively perceived workload. For this purpose, a study was carried out with $n = 93$ test persons in the research flight, in which the test persons had the task of landing a passenger aircraft using the autopilot with different representation forms. Possible work errors 14 with 11 different steps in the representation. Results of positive work-task landing plane: algorithm 58 % (1.7/14 errors $\sigma = 1.5$), text 62 % (1.5/14 errors $\sigma = 1.1$), image 93 % (0.8/14 errors $\sigma = 1.1$).

## 1. Fundamentals and introduction

Human error can lead to serious consequences, especially when it concerns the operation of highly complex systems such as the surveillance of nuclear power plants or operation of an aircraft; examples of this include the Three Mile Island accident in 1979 and the Chernobyl disaster in 1986, both caused by human error [1]. Preventing accidents and ensuring that safety systems are in place are therefore still key issues in today's industry. Various studies have shown that humans play a major role in the cause of accidents [2]. In general, it is repeatedly stated that between 70% and 80% of all malfunctions in complex systems are due to human failure [3]. Improving human reliability is therefore an important challenge to ensure the functionality and safety of a system [2].

In addition to other requirements from the field of human factor engineering, such as the layout of operating elements, the ergonomic design of the working environment, and the design of ergonomic interfaces, the availability of clear and easy to understand instructions is an essential component in the avoidance of human error [4]. Due to the increasing complexity, there is a greater number of possible mistakes an employee can make, especially since not all scenarios can be practiced regularly anymore. Well-structured and clear instructions are indispensable to be able to react to all possible causes. At present, various industry-specific approaches for action instructions can be found in practice. To date, there are no standardized evaluation methods to investigate the efficiency of various instructions for action or the influence they have on security systems.

### 1.1. Aim of this research

The aim of this study is to develop the first possible approaches to this problem and provide initial insights into the quality of three common forms of presentation (image, algorithm and text) used for instructions for actions. An experimental design will be created to evaluate different forms of instructions for dealing with complex systems. Subsequently, the performance of these three different instructions for action will be compared with each other. The purpose of the instructions is to increase safety and security systems, serve a wide range of applications, and reduce training times and costs. The overall objective of this study is to investigate whether there are significant differences in human reliability depending on the type of presentation used for instructions for operating highly complex systems.

* Corresponding author.
*E-mail address:* claudius.hammann@tum.de (C. Hammann).

https://doi.org/10.1016/j.heliyon.2020.e03291
Received 3 April 2019; Received in revised form 8 July 2019; Accepted 21 January 2020

### 1.2. Hypothesis

#### 1.2.1. Frequency of top events

Definition of top event: Airplane crash landing on the ground or airplane not landing on the runway.

$H_0$ (Top event): When using the different instructions, there is no significant difference in the number of top events observed.

$H_1$ (Top event): When using the different instructions, there is a significant difference in the number of observed top events.

$$H_{0\ (top\ event)}\ :\mu_1=\mu_2=\mu_3\ vs\ H_{1\ (top\ event)}\ :\ \mu_i\neq\ \mu_j\ for\ at\ least\ one\ pair\ of\ means$$

#### 1.2.2. Number of errors

Definition of error: Any human action which exceeds acceptance limits.

Definition of unknown error: Human action, performed for the first time.

Definition of known error: Human action has already been performed.

$H_0$ (*Error*): There is no significant difference in the number of observed errors when using the different instructions.

$H_1$ (*Error*): When using the different instructions, there is a significant difference in the number of errors observed.

$$H_{0\ (error)}\ :\mu_1=\mu_2=\mu_3\ vs\ H_{1\ (error)}\ :\ \mu_i\neq\ \mu_j\ for\ at\ least\ one\ pair\ of\ means$$

#### 1.2.3. Time needed to implement the action

$H_0$ (*Time*): When using the various instructions, there is no significant difference in the time required to implement the action.

$H_1$ (*Time*): When using the various instructions, there is a significant difference in the time required to implement the action.

$$H_{0\ (time)}\ :\mu_1=\mu_2=\mu_3\ vs\ H_{1\ (time)}\ :\ \mu_i\neq\ \mu_j\ for\ at\ least\ one\ pair\ of\ means$$

#### 1.2.4. Subjective workload

Definition of workload: Entirety of all external conditions and requirements in the working system that could influence a person physically and/or psychologically.

$H_0$ (Workload): When using the various instructions, there is no significant difference in the workload perceived by the subjects.

$H_1$ (*Workload*): When using the different instructions, there is a significant difference in the workload perceived by the subjects.

$$H_{0\ (workload)}\ :\ \mu_1=\mu_2=\mu_3\ vs\ H_{1\ (workload)}\ :\mu_i\neq\ \mu_j\ for\ at\ least\ one\ pair\ of\ means$$

## 2. Test settings with aircraft cockpit for test persons

### 2.1. Environment and experimental scenario

The research flight simulator (Figure 1) represents a realistic complex system and is therefore suitable for testing different presentation of instructions. It has a three-channel viewing system with a viewing angle of over 180° and Full HD resolution for each channel. The cockpit is equipped with an active sidestick from Wittenstein, which includes an electronic control charging system for force feedback as well as variable hard and soft stops. The SimulinkIMULINK flight system dynamics model running in the background enables simple data recording during test flights in the simulator. Thus, actual states of switches, controllers, levers, buttons and displays as well as all inputs to them can be recorded over the entire test period. The data obtained from the Simulink model forms the basis for the subsequent evaluation to assess the effectiveness of the various instructions for action.

The scenario starts at a fixed airporne position a feq miles from Munich Airport. The aim of the scenario is to land the aircraft safely on the runway, with the help of the automatic flight control system. To successfully complete this task, 14 manual steps must be performed in the simulator cockpit in the correct order and at specified times. To do this, certain display values must be read and compared from different displays, controller inputs must be changed, levers and knobs must be correct, and certain time restrictions must be observed.

**Figure 1.** Research flight simulator.

In addition, some steps may only be carried out when a certain precondition has been reached. The flight scenario used is basically like an approach to an airport in combination with landing by autopilot, as is also done in real flight practice.

Figures 2, 3, and 4 show the relevant displays and control elements, in order to illustrate the action steps for performing the tasks, which is described in more detail below.

If the 14 steps are executed in the order shown above, the scenario will end with a "perfect" landing on the runway, under consideration of certain temporal deviations as well as no temporal deviations. This automatisation makes it possible to check the effectiveness of the instructions for action, as only a limited number of outcomes are possible.

### 2.2. Introduction forms

Due to the given flight simulation scenario, the steps to be taken for a successful landing are identical for all three methods of instructions. The instructions only differ in the form of the representation.

1. Algorithmic representation (Figure 11): The diamond-shaped elements describe the stateof certain levers, buttons, controllers and displays of the system, in front of which the test person are seated. The yellow, rectangular elements describe the action steps that need to be performed by the test person when the initialization state is reached. As a localization aid for the core elements of the algorithm, a corresponding description of the relevant system elements for the respective action step is added in blue boxes on the right.
2. Image instruction (Figure 12): The objective of this method is to imitate the simple and intuitive form of presentation of the action

**Figure 2.** Cockpit.


**Figure 3.** Center console.


**Figure 4.** Keypad front.

instructions presented above, i.e. to create an action instruction that relies solely on images and symbols in order to depict the 14 predefined action steps. The left side shows the actual (current) states of certain levers, buttons, controllers and displays of the system, the right side describes the target state.

3. Textual instruction for action (Figure 13): This type of instruction is based on the one used in aviation. There are two columns: left column describing the initialization state and right one specifying the action to be performed.

### *2.3. Data acquisition from cockpit with error types*

With Simulink from MathWorks, all variables were recorded with 100 [Hz], in order to be able to evaluate them afterwards. The aim of the test is to perform the necessary steps as quickly and precisely as possible. The errors were defined according to [5]. An step is only to be executed when the initialization condition are fullfilled. In the case that a respondent needs so long for action step one that the initialization condition for action step two is already falsified during the conversion of this action one, the result for action step two would be falsified. The time at which these two conditions are fulfilled is the time $t_{PreInit}$ for each action step. The time at which the required action was successfully performed corresponds to the time $t_{succ}$ . For the evaluation, the difference between $t_{succ}$ and $t_{PreInit}$ is now calculated for each action step. This then corresponds to the difference $t_{succ} - t_{PreInit}$.

A metric is defined in the [s] to answer the question about the temporal differences between the occurrence of the required action and the execution of the action. The difference between the time at which an action step becomes necessary and the time at which the action is successfully implemented $t_{succ} - t_{PreInit}$ is produced for each action step. To distinguish the different error types, a split is made in: Operation too late (Time difference between the required operation and actual performing is too long); Operation too early (Action was performed too early); Exessivley the control (Value set too high for implementing the action); Wrong operating element operated (Instead of operating the required operating element, another operating element was used) and No operation (The complete action step was not executed; for example, skipping an instruction step).

### *2.4. Workload acquisition by questionnaire*

The NASA-RTLX [6] is used in this study to investigate the perceived workload of the test person when operating the highly complex system. This is the best-known version of the NASA-TLX, which is easier and quicker to perform and evaluate because it does not weigh the individual characteristics against each other. The comparability of the questionnaires NASA-TLX and NASA-RTLX is given in [6]. This is sufficient to collect the total workload relevant for the study during the work task. Thus, after the study, the total load between the three instructors can be compared and possible differences in the subjective perception of these can be revealed [7]. Figure 5 shows the process of the trial for the proband. Six parameters are collected, each with a possible response of 20 steps:

- Mental Demand
- Physical Demand
- Temporal Demand
- Performance
- Effort
- Frustration

## 3. Results

Data and overview of the test population:

- The study included n = 93 test persons who are either employees or students of the Technical University of Munich (sex: 73 males, 20 females).
- Each of the 93 test persons was randomly assigned to either the group algorithm (n = 31), image (n = 31), or text (n = 31).
- The Age was between 18 and 61 (arithmetic mean = 29.04, years, standard deviation = 9.32 years).
- 79 test persons had no experience with simulators (PC, etc.).
- 24 test persons had flight experience (6 test persons have already operated a flight simulator).
- Technical affinity with ATI questionnaire is 4.721 (scale from 1 to 6). There are no significant differences in the technical affinity between the three groups (algorithm = 4.771, image = 4.633, text = 4.756).

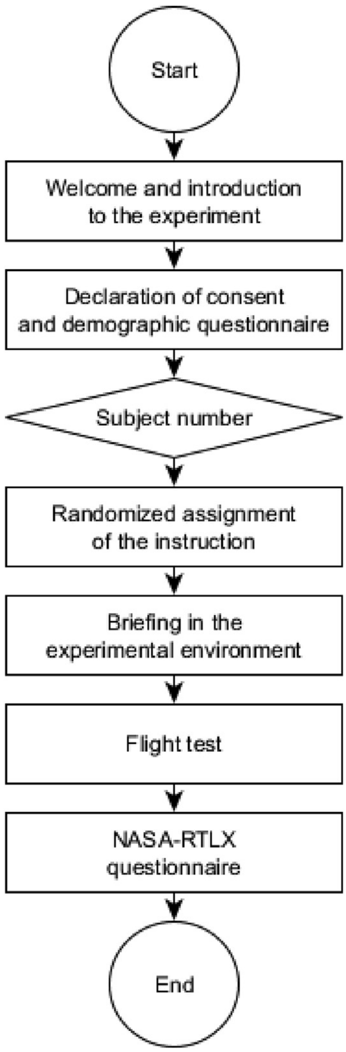


**Figure 5.** Test standard procedure.

### *3.1. Frequency of top event*

A successful landing depends both on the correctly executed steps in the simulator and on the correct chronological sequence. Therefore, a distinction is made between landing on the ground, but not on the runway, (correct sequence of steps in the simulator) and landing on the airport runway (correct sequence of steps in the simulator and appropriate time).

#### *3.1.1. Frequency of top event: No landing on the ground*

Figure 6 shows how many of the test persons managed to prevent the top event ("no landing") and how many caused the occurrence of a top event.

The chi-square test was conducted between the type of instruction and successful landing. No expected cell abundances were less than 5. There is no association between the type of introduction forms and a successful landing on the runway with $\chi^2(2) = 11.586$, $p = .003$, $V = 0.353$.

Since $p = .003$ is smaller than $\alpha = .05$, there is a no association between the three different type of introduction form and successful landing on the ground. With a V of 0.353, the choice of action instruction has an average effect on the number of landings.

$H_{0\ (top\ event\ landing)}$ : *There is no association between the type of instruction and seccessful landing on the ground*

#### *3.1.2. Frequency of top event: No landing on the runway*

Figure 7 shows how many of the test persons managed to prevent the top event (no landing or landing next to the runway) and managed to land on the runway.

The chi-square test was conducted between the type of instruction and successful landing on the runway. There was no expected cell frequency that was less than 5. There was a statistically significant relationship between the type of instruction and successful landing, $\chi^2(2) = 15.040$, $p = .001$, $V = 0.402$.

Since $p = .001$ is smaller than $\alpha = .05$, there is a significant difference in the frequency of runway landings between the three samples. With a V of 0.402, the choice of the action instruction has a medium effect on the number of landings on the runway.

$H_{0\ (top\ event\ runway)}$ : *There is no associtation between the type of instruction and successful landing on runway*

### *3.2. Number of errors*

The following section describes the results for the three hypothesis pairs in terms of the number of errors made. Figure 8 shows a box plot on the X-axis as the representation of the action instruction and on the y-axis the number of errors depending on known, unknown, and all errors.

#### *3.2.1. Number of all errors*

The mean number of all errors (include the known and unknown error) was in the group with the algorithm introduction form 1.74 errors ($\sigma = 1.527$), with the image form Ø 0.81 errors ($\sigma = 1.138$) and the text form with Ø 1.55 errors ($\sigma = 1.502$).

Statistical tests:

- Normal distribution of the three samples of $p < .05$ with significant Kolmogorov-Smirnov test result for all three samples.
- The effective strength with $\eta^2 = .79$ corresponds to an average effect.
- Variance homogeneity with Levene's tests with $p = .382$ showed that a variance homogeneity of the three samples can be assumed.

An alternative hypothesis $H_{1\ (top\ event\ all)}$ : *least one mean is differnet form the others* is assumed. The number of errors differs significantly between the display formats for the total number of errors.

#### *3.2.2. Number of unknown errors*

In case of the analysis only the unknown errors (A single action step that was executed for the first time) the results shows that the group with the algorithm form in mean Ø .35 errors ($\sigma = 1.124$), the group with the image form Ø .55 errors ($\sigma = .850$) and the group with the text form Ø 1.26 errors ($\sigma = 1.064$).

Statistical tests:

- Normal distribution of the three samples of $p = .002$ with significant Kolmogorov-Smirnov test result for all three samples.
- The effective strength with $\eta^2 = .113$ corresponds to an average effect.
- Variance homogeneity with Levene's tests with $p = .082$ showed that a variance homogeneity of the three samples can be assumed.

Alternative hypothesis $H_{1\ (top\ event\ unknown)}$ : *at least one mean is differnt form the others* at is assumed. The number of errors differs significantly between the representation forms for the total number of errors.

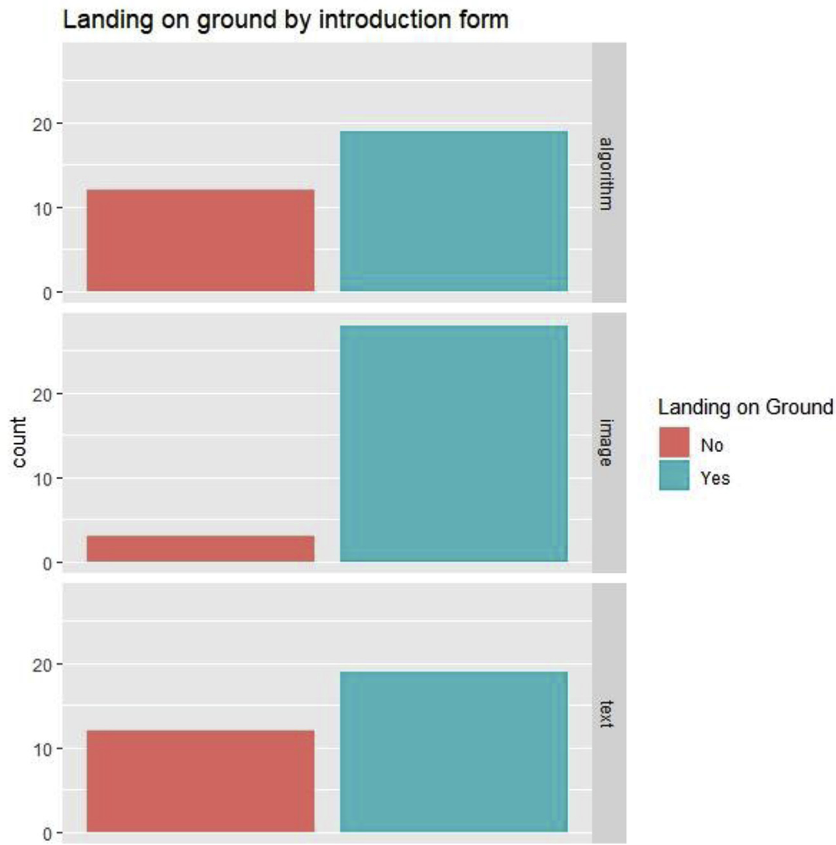


**Figure 6.** Face sheet with number of landings on ground.

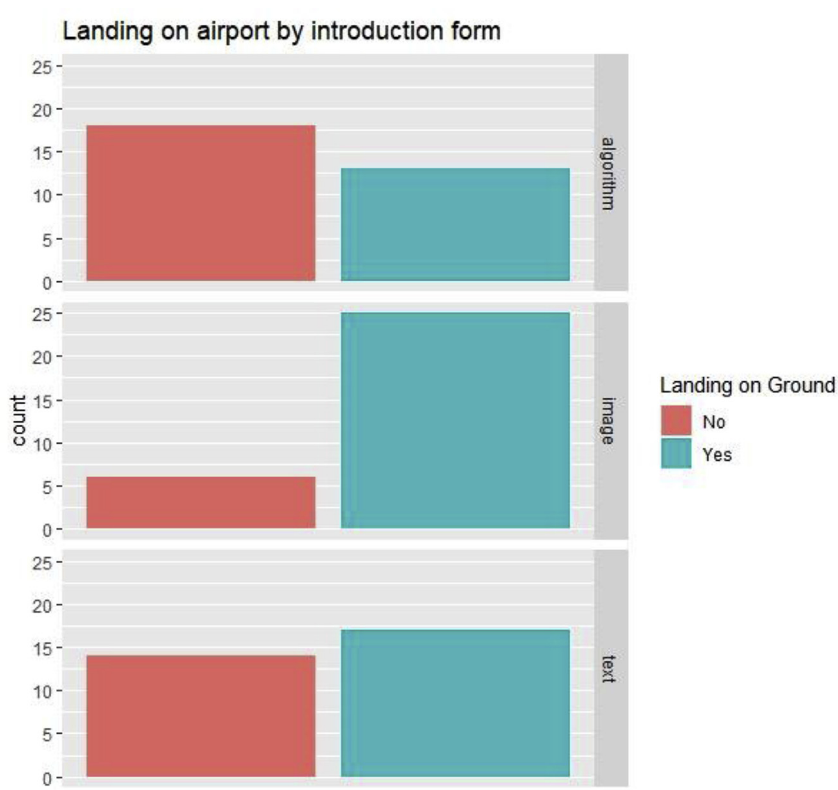


**Figure 7.** Face sheet with number of landings on airport runway.

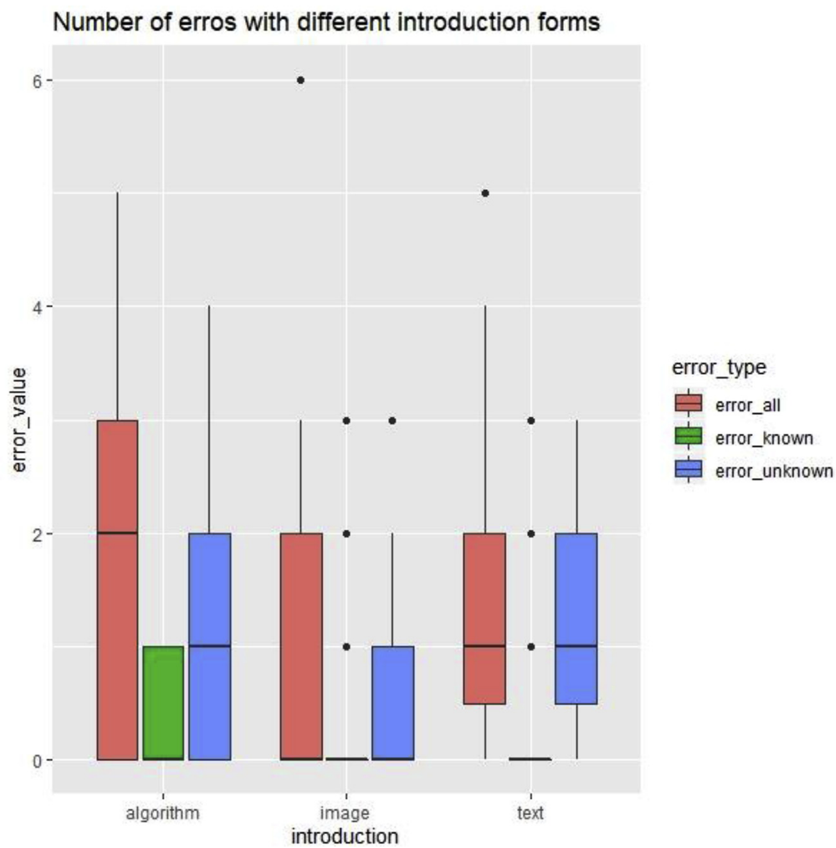


**Figure 8.** Number of errors, divided into all known and unknown errors according to display form.

### 3.2.3. Number of known errors

If analyzed only the known errors (Steps that have already been carried out before and now are to be carried out again) the mean was in the group with the algorithm form Ø .39 errors ($\sigma = .715$), the group with the image form Ø.26 errors ($\sigma = .514$) and the group with the text form Ø .29 errors ($\sigma = .693$).

Statistical tests:

- Normal distribution of the three samples of $p < .001$ with significant Kolmogorov-Smirnov test result for all three samples.
- The effective strength is smaller with $\eta^2 = .007$.
- Variance homogeneity with Levene's tests with $p = .392$ showed that a variance homogeneity of the three samples can be assumed.

Null hypothesis $H_{0\ (top\ event\ known)}$ is assumed. The number of errors does not differ significantly between the representation forms for the total number of errors.

### 3.2.4. Results of human error probability (HEP)

In order to make the results of the three hypotheses more understandable and comparable with the values in the literature, the statistical evaluation, the HRP and HEP values for the various instructions for action were calculated for the three hypothesis pairs. Figure 9 shows the error probabilities of the different instructions. The calculated error probabilities show that the probability of making an error is more than twice as high for algorithmic and textual instructions than for known instructions. The difference between unknown and known steps in the pictorial instruction is much smaller with a rate of below 4% (see Table 1).

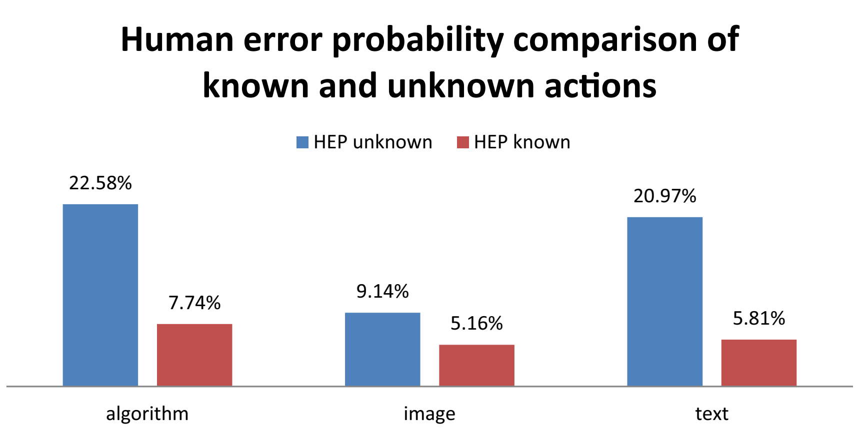


**Figure 9.** Comparison of HEP for unknown actions and known actions.

**Table 1.** Description of the type of errors.

| Error type | Description of error |
|---|---|
| Operation too late | Time difference between the required operation and actual performing is too long |
| Operation too early | Action was performed too early |
| Exessivley the control | Value set too high for implementing the action |
| Exessivley the control | Value set too low for implementing the action |
| Wrong operating element operated | Instead of operating the required operating element, another operating element was used |
| No operation | The complete action step was not executed; for example, skipping an instruction step |

**Table 2.** Number of samples for time analysis.

| Introduction form | n for all<br>mean [s]/standard deviation [s] | n for unknown<br>mean [s]/standard deviation [s] | n for known<br>mean [s]/standard deviation [s] |
|---|---|---|---|
| Algorithm | 242<br>15.289/12.718 | 132<br>18.989/15.077 | 110<br>10.848/6.925 |
| Image | 297<br>12.612/9.799 | 162<br>13.777/11.239 | 125<br>11.215/7.532 |
| Text | 275<br>15.531/15.080 | 150<br>19.293/17.658 | 135<br>11.016/9.494 |

### 3.3. Time for operation

The following results were measured from the times it took the test persons to execute one action step at a time. It should be noted that the calculation is only done for probed events with a complete data set. Table 2: Number of samples for time analysis shows the number of analyzed individual action steps executed by the test persons with the average time in seconds for each action instruction with the mean value and standard deviation.

#### 3.3.1. Time for all operations

The times for all operations (known and unknown operations) required to perform the required actions were statistically significantly different for the three different instructions for all actions, Welch-Test F(2, 502.587) = 5.608, p = .004. To find significant time difference between all possibilities combinations of algorithm, text or image are calculated with Games-Howell post-hoc test. Significant different results are only between algorithm and image form with p = 0.020 (2.676, 95%-CI[.335, 5.018]) and on a further combination between text and image with p = 0.018 (2.919, 95%-CI[.397, 5.440]). The one way ANOVA results performed single-sector ANOVA without repeated measurement showed that the times required to execute known action steps did not differ statistically significant for the three different action instructions, ANOVA $F(2, 367) = .063,\ p = .939,\ \eta^2 < .001$. The null hypothesis

$$H_{0\ (time)}\ :\ There\ is\ no\ time\ difference\ between\ the\ type\ of\ instructions$$

is therefore assumed. Correspondingly, there is no time difference between the different type of introductions.

#### 3.3.2. Time for unknown operations

The result for unknown operations time was for all three introduction form are different. The Welch Test resulted F(2, 271.174) = 8.305, with an significance level of $p < .001$. With the Games-Howell post-hoc find ony significant differences between the algorithm and image form with p = .003 (5.212, 95%-CI[1.482, 8.943]). Further find relation between the text and image form with p = .004 (5.517, 95%-CI[1.530, 9.503]).

#### 3.3.3. Time for known operations

If analyzed only the known operations, there are no statistically significantly different between any other combinations of the three group (algorithm, image or text).

### 3.4. NASA-RTLX workload

Results of test subjects' workload with RTLX score:

- Algorithm: Ø = 8.935, σ = 2.809
- Image: Ø = 7.817, σ = 2.910
- Text: Ø = 9.978, σ = 2.551

Figure 10 shows the answers to the questionnaire (possible answers between 0 and 20 points) with regard to effort, frustration, mental requirement, performance, physical requirement, and time requirement, depending on the different introduction forms.

The null hypothesis was rejected based on ANOVA test results.

$$H_0\ :\ There\ is\ not\ workload\ difference\ between\ the\ type\ of\ instructions$$

is assumed by ANOVA test (F(2, 90) = 4.751, p = .011, $\eta^2$ = .095).

Significant difference between:

- Image and algorithm (1.118, 95%-CI[-.553, 2.789])
- Algorithm and text (2.161, 95%-CI[.490, 3.832])

No significant difference between

- Significant difference in workload between Levene's test (p = .615) and Tukey post-hoc with medium effect strength $\eta^2$ = .095 only for text and image with p = .008 (2.161, 95%-CI[.490, 3.832])

## 4. Discussion

The significant results of the Chi2 test with p = .003 and p = .001, respectively, allow the conclusion that the different results for the frequencies of landings and landings on the runway are not accidental. $H_{0\ (top\ event\ landing)}\ :\ \mu = \mu_0$ is discarded and $H_{1\ (top\ event\ landing)}\ :\ \mu \neq \mu_0$ is assumed. This applies to the landing and landing on the runway of the airport. Therefore, the following conclusion is made: Presentation forms

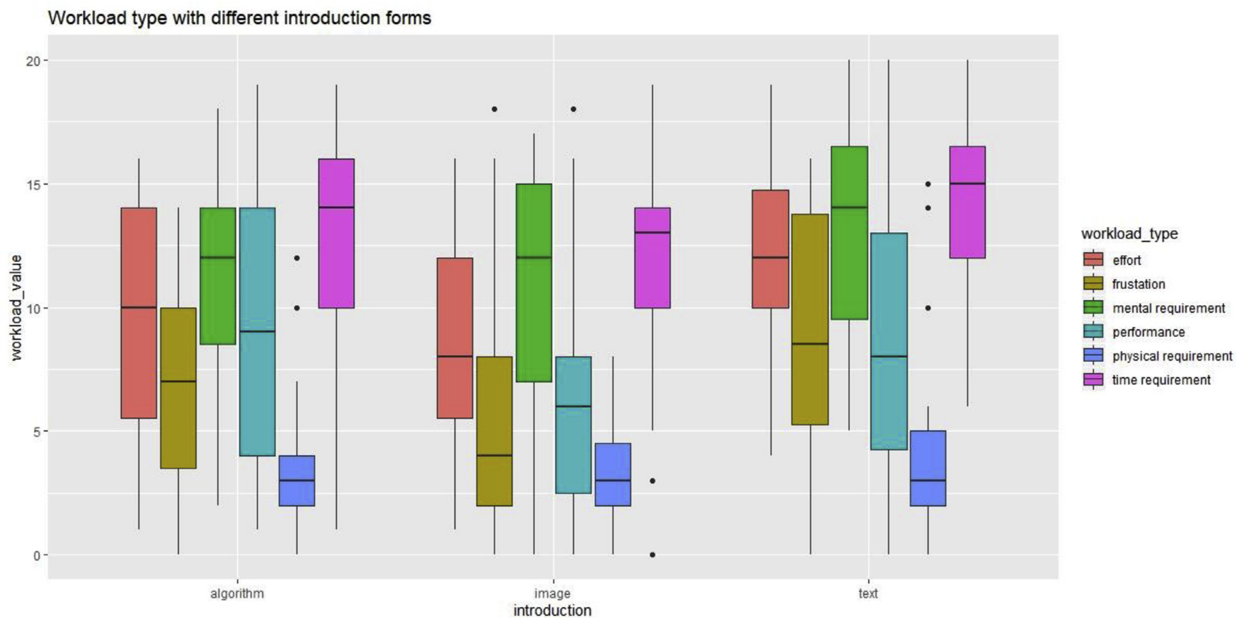


**Figure 10.** Boxplot with workload types depending on the introduction form.

have an influence on the avoidance of top events. The most successful landings (landings next to the runway/landing on runway/total test persons) are visualized (29/26/31), followed by the text (19/13/31) and the algorithm (18/13/31).

The results of the ANOVA on error frequencies show that there is a significant difference between the 11 usable, considered action steps with regard to the error frequencies between the three action instructions. The individual frequencies differ between the three samples to the extent that it appears that these results are not of a random nature. However, the post hoc tests conducted showed that this significant difference can only be demonstrated between algorithmic and pictorial instructions for action. A similar picture can be seen when considering the unknown action steps. The conducted post hoc tests show that there are significant differences between pictorial and textual instructions as well as between pictorial and algorithmic instructions. Here, there is a clear tendency that the pictorial action instruction causes fewer errors in unknown action steps than both algorithmic and textual action instructions. If only the known steps of action are considered, a different picture emerges: error frequencies do not differ significantly between the types of representational form. Thus, none of the three instructions used seems to be superior with regard to error avoidance. An explanation for this could be the fact that the pictorial representation is the only one that visually supports finding the required operating elements by visually describing the position of the operating elements. The other instructions only contain this description in text form. In addition, the examination of the prerequisites for carrying out the ANOVA showed that the assumption of a normal distribution of the samples cannot be assumed on the basis of the Kolmogorov-Smirnov test performed. Since the ANOVA of large samples ($n \geq 30$) is very robust against a violation of the normal distribution, this can be performed and interpreted as in the present case [8].

The results of the Welch ANOVAs on the execution times show that there are significant differences in the execution times of the action steps between the three action instructions, across all action steps. In addition, there is also a significant difference in the execution times of the unknown action steps. For the known action steps, however, no significant difference was found. It should be noted in advance that, in contrast to the previous hypotheses, the sample sizes were no longer identical. Only complete data sets were included in the evaluation, which meant that only 22 test subjects were included in the algorithmic HW, 27 in the pictorial HW, and 25 in the textual HW. This procedure, the elimination of incomplete data sets, can be viewed critically. There are various ways of dealing with missing values. These range from the elimination of incomplete data sets to imputation, which is equivalent to estimating missing values [9]. For the following reasons, an estimation of missing values was dispensed with, and the procedure of eliminating incomplete data sets was chosen. It is also conspicuous that the algorithmic instruction on average achieves better execution times in the known steps than the textual instruction. However, this must be put into perspective insofar as the algorithmic instruction with 22 test subjects had the smallest sample for this test, the worst test subjects had already been eliminated. However, since no significant difference between algorithmic and textual instructions could be determined, this fact plays a rather subordinate role. The comparatively high standard deviations across all hypothesis tests are a further sign that the inter- and intraindividual differences were very large. This makes it more difficult to interpret the results unambiguously.

The results of the evaluation of the RTLX questionnaire have shown that there is a significant difference in perceived stress between textual and visual presentation. However, a significant difference between the textual and algorithmic, as well as the algorithmic and pictorial presentation form was not found.

The perceived load during the work task was shown to be the highest with the textual instruction for action. This is followed by the algorithmic instruction, which was evaluated as somewhat less stressful than the textual form. However, since the difference between the RTLX scores achieved was not significant for these two forms of presentation, a precise statement cannot be made about the causes of the difference. The pictorial instruction for action yielded the lowest value and is thus the "least stressful" type of presentation. It was shown that there are substantial differences with regard to the perceived stress. Since the success of the accomplished task must also be evaluated in the RTLX, this may also be a reason for the good results of the pictorial action instruction. This was underlined by the results of the "top event" hypothesis. The pictorial instruction also resulted in the most successful landings, which may be related to the lower perceived load. The use of the algorithmic form of representation tended to be classified as less stressful than the textual form of representation. With regard to the RTLX scores evaluated with a single-sector ANOVA, it is questionable whether the ANOVA requirement for a dependent variable with at least interval scaling is fulfilled. The problem here is whether the RTLX score represents such an interval-scaled variable. To achieve this, the difference between two different intervals would have to be interpreted in exactly the same way [10].

In research practice, it has become established to assume multilevel questionnaire items as being interval-scaled [11]. The implementation of a single-step ANOVA is thus a tried and tested means of evaluating the RTLX scores [12].

## Declarations

### *Author contribution statement*

C. Hammann: Conceived and designed the experiments; Wrote the paper.

A. Feldhütter: Analyzed and interpreted the data.

C. Krause: Contributed reagents, materials, analysis tools or data.

### *Funding statement*

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

### *Competing interest statement*

The authors declare no conflict of interest.

### *Additional information*

No additional information is available for this paper.

## Appendix

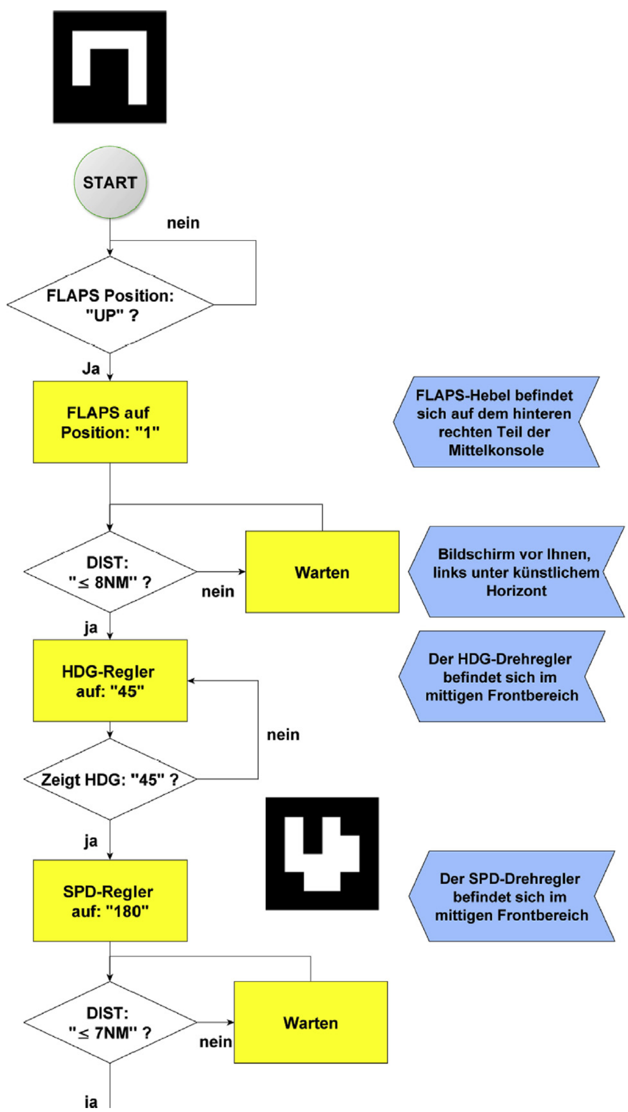


**Figure 11.** Algorithm introduction form.

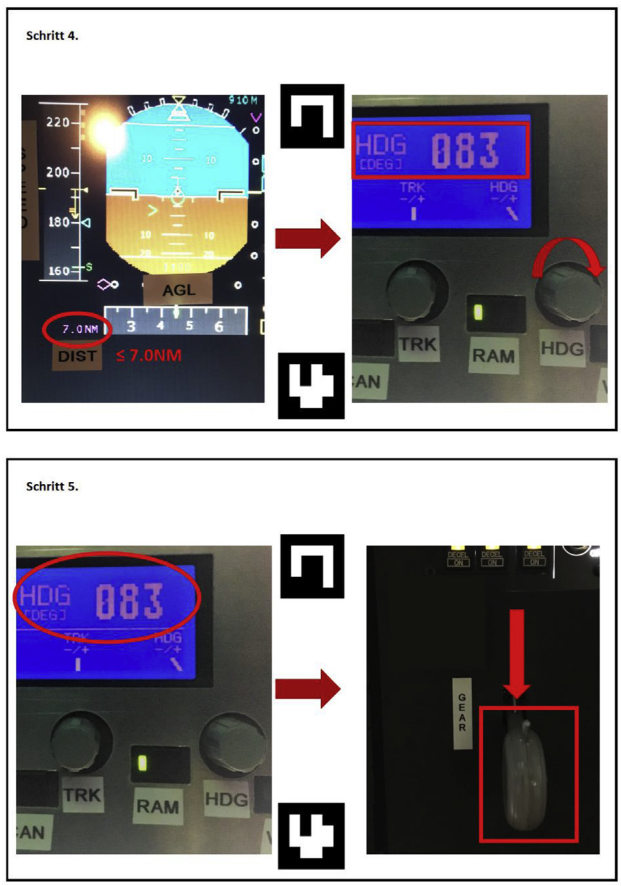


Figure 12. Image introduction form.

**Chronologische Checkliste für Landeanflug**

| Initialisierungszustand | Auszuführende Handlung |
|---|---|
| 1. **FLAPS** (Hebel Mittelkonsole hinten rechts) auf Position: **„UP“** | **FLAPS** auf Position: **„1“** |
| 2. **DIST** (Bildschirm vor Ihnen, links unter künstlichem Horizont) zeigt: **„≤ 8.0NM“** | **HDG**-Regler (mittiger Frontbereich) auf: **„45“** |
| 3. **HDG-Regler auf: „45“** | **SPD**-Regler (mittiger Frontbereich) auf: **„180“** |
| 4. **DIST** (Bildschirm vor Ihnen, links unter künstlichem Horizont) zeigt: **„≤ 7.0NM“** | **HDG**-Regler auf: **„83“** |
| 5. **HDG**-Regler auf: **„83“** | **GEAR**-Hebel (runder Hebel rechts neben zentralem großen Bildschirm) auf Position: **„unten“** |
| 6. **DIST** (Bildschirm vor Ihnen, links unter künstlichem Horizont) zeigt: **„≤ 6.0NM“** | **APR**-Button (mittiger Frontbereich): “**drücken**” |
| 7. **APR**-Button: **„aktiv“** | **FLAPS** auf Position: **„2“** |
| 8. **FLAPS** auf Position: **„2“** | **SPD**-Regler auf: **„150“** |
| 9. **SPEED** (Bildschirm vor Ihnen, links neben künstlichem Horizont) zeigt: **„≤ 155“** | **FLAPS** auf Position: **„3“** |
| 10. **FLAPS** auf Position: **„3“** | **SPD**-Regler **„140“** |
| 11. **SPEED** (Bildschirm vor Ihnen, links neben künstlichem Horizont) zeigt: **„≤ 144“** | **FLAPS** auf Position: **„FULL“** |
| 12. **AGL** (Bildschirm vor Ihnen, unter künstlichem Horizont) zeigt: **„≤ 100“** | **A/THR**-Button (mittiger Frontbereich): “**drücken**” |
| 13. **A/THR**-Button: **“inaktiv”** | **SCB-LEFT** und **SCB-RIGHT** Hebel (Doppelhebel in zentraler Mittelkonsole) auf: “**0**” |
| 14. **SCB-LEFT** und **SCB-RIGHT** Hebel: **“0”** | **SPOIL**-Hebel (Hebel Mittelkonsole hinten links) **auf: „zurück“** |
| 15. **SPEED** (Bildschirm vor Ihnen, links neben künstlichem Horizont) zeigt: **“≤ 10”** | **End.** |

**Figure 13.** Text introduction form.

## References


[1] James T. Reason, Joachim Grabowski, Menschliches Versagen. Psychologische Risikofaktoren und moderne Technologien, Spektrum Akad. Verl, Heidelberg, 1994 (Spektrum Psychologie).

[2] Human factors and reliability engineering for safety and security in critical infrastructures, in: Fabio de Felice, Antonella Petrillo (Eds.), Decision Making, Theory, and Practice, Springer International Publishing (Springer Series in Reliability Engineering), Cham, 2018.

[3] Heiner Bubb, Menschliche Zuverlässigkeit - Ergonomische Anforderung und Methoden der Bewertung. VDI 4006 Blatt 1 und 2, 2002. Online verfügbar unter, https://www.researchgate.net/publication/291661476_Menschliche_Zuverlassigkeit_-_Ergonomische_Anforderung_und_Methoden_der_Bewertung_VDI_4006_Blatt_1_und_2. (Accessed 15 October 2018).

[4] Balbir S. Dhillon, Human Reliability, Error, and Human Factors in Engineering Maintenance. With Reference to Aviation and Power Generation, CRC Press, Boca Raton, Fla, 2009. Available online at, http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10308731.

[5] A.D. Swain, H.E. Guttmann, Handbook of Human-Reliability Analysis with Emphasis on Nuclear Power Plant Applications. Final Report, Sandia National Labs, 1983 (NUREG/CR-1278).

[6] J. Byers, A. Bittner, S. Hill, Traditional and raw Task Load Index (TLX) correlations: are paired comparisons necessary, in: Taylor und Francis (Ed.), Advances in Industrial Ergonomics and Safety I. Proceedings of the Annual International

Industrial Ergonomics and Safety Conference, Held in Denver, Colorado, June 10-14, 1992, CRC Press, 1989, pp. 481–489 [Place of publication not identified].
[7] W.F. Moroney, D.W. Biers, F.T. Eggemeier, J.A. Mitchell, A comparison of two scoring procedures with the NASA task load index in a simulated flight task, in: IEEE (Ed.), Proceedings of the IEEE 1992 National Aerospace and Electronics Conference@m_NAECON 1992. IEEE 1992 National Aerospace and Electronics Conference@m_NAECON 1992. Dayton, OH, USA, May 18-22, 1992, IEEE, 1992, pp. 734–740.
[8] Markus Bühner, Matthias Ziegler, Statistik für Psychologen und Sozialwissenschaftler. 3. Aufl, Pearson Studium (Always learning), München, 2012.
[9] Mario Gollwitzer, Michael Eid, Manfred Schmitt, Statistik und Forschungsmethoden. Lehrbuch. Mit Online-Materialien. Deutsche Erstausgabe, 3., korrigierte Aufl. Weinheim, Basel: Beltz Verlagsgruppe, 2013. Available online at, http://www.content-select.com/index.php?id=bib_view&ean=9783621278348.
[10] Günther Bourier, Beschreibende Statistik. Praxisorientierte Einführung - mit Aufgaben und Lösungen. 13. Auflage. Wiesbaden, Springer Gabler, 2018. Available online at.
[11] Hans Dieter Mummendey, Ina Grau, Die Fragebogen-Methode. Grundlagen und Anwendung in Persönlichkeits-, Einstellungs und Selbstkonzeptforschung. 6., korrigierte Aufl. Göttingen, Hogrefe Verlag, 2014. Available online at, http://elibrary.hogrefe.de/9783840925771/U1.
[12] Julie A. Jacko, Human-computer Interaction. 12th International Conference, HCI International 2007, Beijing, China, July 2007; Proceedings, Part 2: Interaction Platforms and Techniques, Springer, Berlin, 2007 (Lecture notes in computer science, 4551). Available online at, http://www.springerlink.com/openurl.asp?genre=issue&issn=0302-9743&volume=4551.