Analysis report

Programmable content and a pattern-matching algorithm for real-time automatic authoring in AR maintenance applications

Abstract

This document presents the experimental results and analyses that contribute to the research validation of *"Programmable content and a pattern-matching algorithm for real-time automatic authoring in AR maintenance applications"*. The following packages, functions and themes have been used for data analysis within the R code presented in this document.

```r
# Functions to manage and analyse data
library(dplyr)
library(tidyr)
library(car)
library(scales)
# Functions to work with plots
library(ggplot2)
library(grid)
library(ggpubr)
# Functions to work with tables
library(knitr)
# Function to save plots
plot_save <- function(filePlot,filePath, fileDevice) {
    ggsave(filename = filePath, plot = filePlot, device = fileDevice,
           width = 15, height = 10, units = "cm", dpi = 640)
}
# Declare colour palettes
c03Palette = c("#1A406A","#7F7F7F","#0D1930")
c06Palette = c("#3F97C0","#1A406A","#9EBF43","#C2446F","#F2BC41","#5D4184")
c09Palette = c("#3F97C0","#1A406A","#9EBF43","#C2446F","#F2BC41","#5D4184",
               "#D32D40","#7F7F7F","#0D1930")
c12Palette = c("#D32D40","#F2BC41","#9EBF43","#3F97C0","#5D4184","#C2446F",
               "#791C24","#C77F3A","#617628","#2A386B","#402D55","#782A43")
# Declare plot theme
plotTheme <-
    theme(panel.background = element_rect(colour= "gray90", fill = "white"),
          strip.background = element_rect(colour = "gray90", fill = "white"),
          panel.grid.major = element_line(colour = "gray90", size = 0.35),
          panel.grid.minor = element_line(colour = "gray90", size = 0.175),
          axis.ticks = element_blank(),
          text = element_text(size = 11, family = "Times"))
```

This document is structured as follows. Its first section presents the validation methodology, including research objectives and contributions to validate, and the methods to do so. The application of such methods, along with their criteria and protocols, is presented in the second section. Based on the outcomes of these experimental protocols, the third section describes their quantitative and qualitative analysis. Its fourth section describes the analysis results and discusses their impact on the research validation. Finally, the fifth section draws the research's conclusions, along with the analysis assumptions made to infer those.

1. Methodology

The research presented in *"Programmable content and a pattern-matching algorithm for real-time automatic authoring in AR maintenance applications"* proposes a pattern-matching algorithm for ontology-based automatic authoring in AR maintenance applications. It matches programmable content formats with ontological individuals (attributes and relationships) through semantical analysis to automatically generate augmented content in real-time. The proposal's aim is to reduce AR deployment costs while maintaining augmented content's semantic understanding compared to other ad-hoc authoring proposals. Because the authoring proposal is automatic, it should be cheaper to deploy than other authoring approaches (e.g. expert-based or specific). Hence, this research's validation aims to evaluate content's semantic understanding through its impact on AR-supported maintenance operations.

In academia, a common approach [**refs**] to evaluate the impact of Augmented Reality in maintenance is analysing its effect on efficiency. Maintenance operations are often information-intensive human tasks [**ref**] and so, AR effect on efficiency can be quantitatively evaluated through **time** and **errors**. These measures can be considered a direct representation of efficiency for a certain level of effectiveness in human-related tasks [**ref**]. Thus, AR experiments normally involve **pre-determined maintenance tasks** with specific levels of effectiveness or quality. Besides, another relevant aspect of AR impact in maintenance is its user-perceived **usability**. Although it is a qualitative measure, it seems reasonable to believe that an AR human-interface may not have a positive effect on human-related tasks if users do not find it usable. Hence, AR proposals with a positive impact on maintenance efficiency should also have a positive perceived usability.

This research focuses on evaluating the abovementioned measures in two different maintenance operations for validating its proposal against two expected contributions:

1. To produce content automatically with similar semantic understanding of other ad-hoc authoring solutions.
2. To produce suitable content for diverse maintenance operations such as repair or remote diagnosis.

Future works: further analyse related AR deployment costs and authoring usability to demonstrate it is cheaper and easier to implement

In order to do so, this research validation conducts the two following experimental methods:

1. **Efficiency experiments:** to evaluate the proposal's impact on maintenance efficiency compared to other authoring solutions.
2. **Usability surveys:** to evalute the proposal's perceived usability compared to other authoring solutions.

These two evaluations should provide sufficient evidence to demonstrate the proposal's validity to fulfil the abovementioned contributions. For this purpose, the following section presents the experimental design and protocols for the two validation objectives described above.

2. Design

This section presents the experimental protocols for this research validation. According to similar research **[refs]**, the following criterions have been considered appropriate for evaluation by each of the abovemention methods:

	Quantitative	Qualitative
Experiments	Time and errors	
Surveys		Nielsen's usability

For these criterions to be appropriate for evaluating maintenance efficiency effects, the following assumptions are required:

- Time and errors can be a direct representation of efficiency if a consistent quality is assumed at the experimented maintenance operations. In order to ensure so, the study assumes pre-determined operations whose quality does not depend on the tester's performance.
- The interaction of the AR solution can affect efficiency if it is not compatible with the tester's manual operations or environment. So, the proposal's usability is an important measures towards the evaluation of maintenance operations' quality. Since there are no available quantitative measures to evaluate those, qualitative, subjective measures will be used.

Each set of criteria, along with the relevant experimental protocols to evaluate them, are presented in the two following subsections per each validation method and consequent objective. The third subsection presents the methods' cases of study, which comprise two operations and their maintained equipment. Finally, the fourth subsection describes the experimental protocols for conducting validation methods and analysing their results.

2.1. Stopwatch time and errors studies

The stop watch time and errors study aims to analyse the effect of the proposed authoring **solution** (PMAU) over maintenance efficiency on different **operations** compared to alternative solutions (ARAUM, SMAARRC, NONE). It is assumed that AR-improved semantic understanding of real-world objects increases efficiency of maintenance tasks. In such scenarios, it can be said that efficiency solely depends on time for similar levels of effectiveness (quality).

Time can be described by the number of seconds required by a tester to find, understand and complete a maintenance **task**. Quality, also declared as **errors**, can defined as the number of tasks completed by a tester that deviate in form or result of what was pre-determined. Besides, semantic understanding is assumed to affect maintenance efficiency through the authoring **solution** experimented and the **task** maintenance **operation** being conducted.

Based on previous definitions, it can be said that if errors (quality) are unvariable, then the effect of authoring solutions through semantic understanding over maintenance efficiency can be evaluated based on its effect on completion time. Such evaluation should be made over different maintenance operations to demonstrate the validity of this research contribution. If the assumptions above are correct, then it is reasonable to expect the following results:

- Errors do not vary with the use of different solutions for each maintenance operation.
- Times are reduced with the use of authoring solutions compared to non-AR solutions for each maintenance operation.
- Times do not vary significantly between authoring solutions for the same maintenance operation.

The study described above considers one response variable (**time**), two control variables to test assumptions (**errors** and **tasks**), and two independent factor variables (**solution** and **operation**). These variables are defined in the table below:

Variable	Type	Definition
Time	Response	Time taken by a tester to identify, understand and complete a maintenance task
Errors	Control	Tasks completed with form or result deviations from its pre-defined target
Step	Factor	Specific assignment to be undertaken by a tester as part of a maintenance operation
Solution	Factor	Authoring solution employed to generate augmented content support to conduct maintenance tasks
Operation	Factor	Nature of tasks being conducted which belong to a specific step in the maintenance process

Each factor variable has different levels. Their definitions are presented in the table below:

Factor	Level	Definition
Solution	PMAU	Use of this research proposal to generate AR support
Solution	ARAUM	Use of an ad-hoc authoring solution for maintenance repair
Solution	SMAARRC	Use of an ad-hoc authoring solution for maintenance remote diagnosis
Solution	NOAR	Use of non-AR solutions to support maintenance operations
Operation	Repair	Maintenance tasks aiming to return equipment to its working conditions
Operation	Diagnosis	Maintenance tasks aiming to identify the cause of an equipment's failure

The experiments aim to at test the proposed authoring solution against other ad-hoc and non-AR authoring solutions in two different maintenance operations. In order to simplify the evaluation process, the tasks experimented at the ad-hoc authoring solutions researches will be re-utilised in these experiments. These cases of study, which comprise different maintenance tasks and equipment, are presented in **Section 2.3**.

Each study experiment, one for each operation, will consist of a tester conducting operation's steps with two different authoring solutions. Besides, results from previous researches for non-AR support will be re-utilised to use them as baseline comparators. Therefore, testers will be grouped in six different groups according to the abovementioned procedure and factors. These groups are as follows:

	PMAU	ARAUM	SMAARCC	NOAR
Repair	A	B		C
Diagnosis	B		A	D

The reason to re-use testers on two different maintenance operations is for them to be able to compare the usability of two different authoring solutions. This is comparison is necessary because testers are **assumed** to have none or very little previous experience in maintenance or AR. Besides, the maintenance tasks (described in **Section 2.3**) can be **considered** sufficiently different for not expecting carry-over effects between experiments.

2.2. Usability surveys

Usability surveys aims to evaluate the perceived validity of the proposed authoring solution to enhance semantic understanding compared to alternative authoring methods. Usability refers to the ability of the authoring solution to deliver information appropriately to the user regarding the maintenance operation to be conducted. Besides, it is a feature perceived by users and so subject to opinion. Therefore, it is neccesary to use qualitative criteria for its evaluation. Based on similar research **[refs]**, the criteria utilised in these surveys is that presented by Nielsen in his 1993 book "Usability Engineering" **[ref]**. These usability criterions aim

to evaluate different aspects of the authoring solution regarding its formats and its impact on maintenance operations. The criterions are defined in the following table:

Criterion	Aspect	Scale
Ease-to-learn	Start, Finish, Intuitiveness	Likert Scale 1-5
Ease-to-use	Gestures, Text, Buttons, Images, Models, Holograms, Animations	Likert Scale 1-5
Accuracy	Overlay, Shaking, Occlusion, Visualisation, Latency	Likert Scale 1-5
Effectiveness	Efficiency, Confidence	Likert Scale 1-5
Satisfaction	Design, Feeling, Overall	Likert Scale 1-5

Each criterions includes a separate survey section with several statements for each aspect regarding the authoring solutions tested in experiments. Users are asked to determine their agreement with these statements in a Likert Scale (1-5). The results collected serve to evaluate the authoring solution's usability compared to other specific approaches. Besides, operational quality is also evaluated in terms of efficiency and confidence improvements. There are some assumptions to consider regarding these surveys:

- Errors are not evaluated in quality terms as they may be dependent on user expertise, which can vary for potential users of this solution.
- It is assumed that the quality is of consistent level for the stop watch time study if the results of the questionnaire provide a similar result to the experiments.

The protocols to collect and analyse experimental and survey data are described in **Section 2.4**. Instead, the following section presents the experimental cases of study along with the testing population's sample.

2.3. Cases of study and population samples

The cases of study comprises two maintenance operations (repair and remote diagnosis) to be conducted in two different complex-engineering assets. These cases of study were already presented and discussed in the two publications regarding the experimental alternative authoring solutions. In order to accomodate these cases of study to ontology-based information systems, the mapping procedure from **Cullot, Ghawi and Yétongon (2007)** was used.

2.3.1. Maintenance repair: Repont

The first case study is the same one utilised in **[ref]**. It represents maintenance repair operations in complex engineering assets for the Defence Industry. These are focused mainly in mechanical, electric and hydraulic systems and assembly and replacement procedures. The case-study equipment is a laboratory prototype of a gearbox for studying gear-wheels degradation that represent real-life conditions of asset-repair scenarios. The experiments described in **[ref]** focus on a specific repair operation composed of several assembly, disassembly and replacement steps involving mechanical components. These experiments aimed to analyse the effect of an ad-hoc tablet-based authoring solution called ARAUM, which aimed to simplify the generation of animations. The experimental repair scenario conducts an operation to replace a gearbox's component (brake wheel) when it has been worn away. This repair scenario includes the following instructions:

1. Unscrew and remove the transparent cover
2. Unscrew and remove brake support
3. Replace brake piece and re-screw brake support
4. Place back and re-screw the transparent cover

Additional data from ARAUM's database is also augmented for experiment purposes. Each 'operation' includes 'tools', 'items' and 'safety_precautions' that are displayed as text in the AR application. Each 'instruction' is also delivered by AR means through a textual description and an additional animation overlaid on top of the real-world object imitating the movement to be done for conducting the repair step. For testing

PMAU, ARAUM's databse was converted into an ontology. Few modifications have been done comparing with the database:

- 'System' and 'Component': instead of having the data for AR-tracking each object, PMAU utilises existing CAD models through model-based approaches. So, the information required is just the name of the system or component to track and the source of its CAD model.
- 'Assembly': PMAU fully automatises augmented content. So, there is no longer needed to identify a 'rendering model' and its movement. Instead, the geometrical relations ('Spatial' and 'Mating) between a component and its assembly pair are used to generate the disassembly animation. The ontology needs a knowledge base to be fully completed for experimentation. The instructions and additional descriptions above comprise the ontology individuas that conform the case-study's knowledge base.

2.3.2. Maintenance remote diagnosis: Remont

The second case study is the same one utilised in [**ref**]. It describes remote maintenance diagnosis operations for complex engineering assets in the Aerospace Industry. The focus of these operations is purely in mechanical systems. In this case study, AR aims to develop effective communication-support tools for enhancing remote diagnosis in 'decision-to-fly' scenarios. The case-study equipment is an aircraft's fuel hatch prototype with unidentified imperfections that are the diagnosis target. The experiments described in [**ref**] focus on a diagnosis operation that comprises inspection, measurement and repair of mechanical components. These experiments aimed to analyse the effect of an ad-hoc HoloLens-based authoring solution called SMAARRC, which aimed to simplify the understanding of complex messages. The experimental diagnosis scenario conducts an operation to identify several defects that the fuel hatch has and resolve them if necessary. This diagnosis scenario includes the following instructions:

1. Open front panel
2. Inspect interior and exterior of panels
3. Apply patch in left porthole crack
4. Photograph patch final result

As a communication-support AR tool, the case-study database represents the elements declared for the specific remote diagnosis messages. The database does not store pre-identified information per se but the reported messages generated by the AR-supported communication. Therefore, the message elements and the augmentation methods for each of them declare the database structure. Each message element has different augmentation methods, which are the real-time authoring rules given to the expert as a desktop application to send messages to the technician for conducting the remote diagnosis using a head-mounted device.

In order to use the case-study data in PMAU validation, it seems necessary to transform that to an ontology. Few modifications have been done comparing with the database:

- 'Component': its name is used to also to identify the hologram to deploy.
- 'hasValueMeasure' and 'hasUnitMeasure': specific datatypes to declare the value and the unit of a measure in a standard international formats.
- 'nextIs' and 'previousIs': substitute 'identifier' linking message occurrence.
- 'Indicator': specific class to declare 3D location and rotation coordinates.

The ontology's knowledge base is conformed by the messages necessary to remotely send the instructions above. The SMAARRC case study involves additional modifications compared to ARAUM. The SMAARRC experiment considers a desktop application for the remote expert to send the messages to the AR-supported technician. This desktop application comprises a 3D model view where to generate the messages and the technician's live streaming. For PMAU experiments to replicate ARDRRC experiments, the expert application will comprise an ontology interface to send messages [**ref**], a 3D model view and the technician's live streaming.

2.3.3. Experimental sample

A total of 30 MSc students (24 males and 6 females) participated as testers in laboratory experiments. Their ages range from 22 to 29 years and they are all enroled in engineering-related MSc degrees. Although they have some basic knowledge in AR and maintenance due to their courses, they have no previous hands-on experience in any of them. So, they were given a short training on AR devices right before experimentation to avoid the presence of any learning curves. Testers were randomly allocated to one of the two groups (A (15) or B (15)) to avoid "carry-over" effects between maintenance procedures while using two different authoring solutions. Besides, the results from previous researches were used as baselines for the groups of NOAR solutions (groups C and D).

2.4. Experimental protocol

The protocol comprises the steps to collect and analyse experimental and survey data for validating this research proposal against its expected contributions. It implements the validation methods above in the context of the cases of study described above. The following list summarises this experimental protocol:

1. **Data collection** (30 testers per experiment):
    a. AR-maintenance introduction: to briefly describe testers the purpose of experiments as well as the use of AR solutions in maintenance operations.
    b. Efficiency experiments: to capture quantitative data on the effect on efficiency of different authoring solutions for diverse maintenance operations.
    c. Usability surveys: to capture qualitative data on tester's opinions regarding usability of the authoring solution proposed compared to other alternatives used within experiments.
2. **Data analysis** (45 testers per experiment):
    a. Errors effect study: to ensure the validity on the assumption that quality is kept among experiments. Results should reflect that there is no significant differences on the errors made by testers using different solutions in maintenance operations. Basic statistics and graphical analysis will be used for this matter.
    b. Time effect study: to analyse the correlation between the response variable (time) and considered factors (solution and operation). Results should reflect that the proposed authoring solution (PMAU) does not present significant differences on time compared to alternative authoring solutions (ARAUM and SMAARRC) in different maintenance operations. They should also reflect that these are significantly different to NOAR solutions. Experiments are set independently for each maintenance operation, and so the factors to consider in the analyses (Step and Solution). Due to the number of factors (2 - step and solution), a two-way ANOVA analysis will be used to tests these hypotheses for each experiment. Moreover, additional post hoc test comparisons (TukeyHSD test) will be used to evaluate existing interactions between factors' levels. Besides, homogeneity, normality, linearity and additivity assumptions will be tested to demonstrate the validity of the analyses results.
    c. Usability study: to quantitatively evaluate testers' opinions on the proposal's usability. Results should reflect that usability does not compromise the effectiveness of the supported maintenance operation. Due to the quantitative nature of these results, basic statistics and graphical analysis will be used for this matter.

This experimental protocol aims to validate the research proposal against its expected contributions. For this validation to be coherent and complete, there are few assumptions to consider:

- In order to keep consistency with previous researches [**refs**] the experiments were conducted in a laboratory environment in order to keep constant other factors (e.g. ergonomics or lighting conditions) that may affect the results. This enabled to reutilise results from previous research regarding the testing of NOAR solutions for the case study operations.
- Additional effects studied in previous researches are not considered in this protocol. The aim is to prove that the new authoring method achieves similar times to alternatives, so the contributions achieved with those should also be applicable to these new authoring method.

- Experimental sample size for the abovementioned statistical tests can be estimated **"a priori"**. Such estimation can be done using a F test for the most requiring analytical test (two-way ANOVA) using GPower software [**ref**]. With 12 factor groups (solution and step factor levels), a variance of 0.25 (partial eta squared), a type-I error of 0.1 (alpha) and a power of 0.9 (1 – beta), the resultant sample size is 51 people. That is quite close to the 45 sample size achieved: 30 testers from this research experiments and additional 15 testers results obtained from previous researches [**refs**]. Besides, these numbers are bigger compared to similar researches that achieved sample sizes of 30 testers [**refs**].
- As described above, testers are MSc students with none or very little experience in AR or maintenance. Although this ensures a baseline for measuring maintenance efficiency, further experiments should be require to corroborate these laboratory results in real-life working conditions.

The analyses and results of this experimental protocol are presented in the following sections.

3. Analysis

WHAT ABOUT POTENTIAL STUDIES ON THE CONTENT BEING GENERATED AND COMPARISON WITH OTHER AUTHORING SOLUTIONS?

3.1. Data pre-processing: collection and formatting

Each data set has been prepared in R-readable formats (long tables) for further treatment. These data sets can therefore be imported and transformed into data frames.

```
## 'data.frame':    90 obs. of  4 variables:
##  $ Tester   : Factor w/ 60 levels "1","2","3","4",..: 1 1 2 2 3 3 4 4 5 5 ...
##  $ Operation: Ord.factor w/ 2 levels "Repair"<"Diagnosis": 1 2 2 1 1 2 2 1 1 2 ...
##  $ Solution : Ord.factor w/ 4 levels "PMAU"<"ARAUM"<..: 2 1 1 2 2 1 1 2 2 1 ...
##  $ Errors   : int  1 1 0 1 0 1 0 0 2 1 ...
```

```
## 'data.frame':    360 obs. of  5 variables:
##  $ Tester   : Factor w/ 60 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 2 2 ...
##  $ Operation: Ord.factor w/ 2 levels "Repair"<"Diagnosis": 1 1 1 1 2 2 2 2 2 2 ...
##  $ Solution : Ord.factor w/ 4 levels "PMAU"<"ARAUM"<..: 2 2 2 2 1 1 1 1 1 1 ...
##  $ Step     : Factor w/ 8 levels "D1","D2","D3",..: 5 6 7 8 1 2 3 4 1 2 ...
##  $ Seconds  : int  105 151 164 118 152 58 41 15 151 67 ...
```

```
## 'data.frame':    1440 obs. of  6 variables:
##  $ Tester   : Factor w/ 30 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Operation: Ord.factor w/ 2 levels "Repair"<"Diagnosis": 1 2 1 2 1 2 1 2 1 2 ...
##  $ Solution : Ord.factor w/ 4 levels "PMAU"<"ARAUM"<..: 2 1 2 1 2 1 2 1 2 1 ...
##  $ Criterion: Ord.factor w/ 5 levels "Ease-To-Learn"<..: 1 1 1 1 1 1 2 2 2 2 ...
##  $ Aspect   : Factor w/ 24 levels "Animations","Buttons",..: 21 21 7 7 14 14 2 2 10 10 ...
##  $ Response : int  5 4 5 5 5 5 4 4 NA 5 ...
```

Modifications

- Number of errors per step is very low. Hence, errors data has been group for all steps per tester for further analyses.
- Data frames for errors and seconds are splitted according to Operation factor as this has an effect on other factors for hypotheses testing.
- Data frame for surveys may have missing values. Some question have not been responded by some testers, there are some NA values within the dataset that need to be removed on treatment.
- Data frame for surveys is splitted according to Criterion factor to simplify analyses.

3.2. Errors effect study

3.2.1. Exploratory analysis

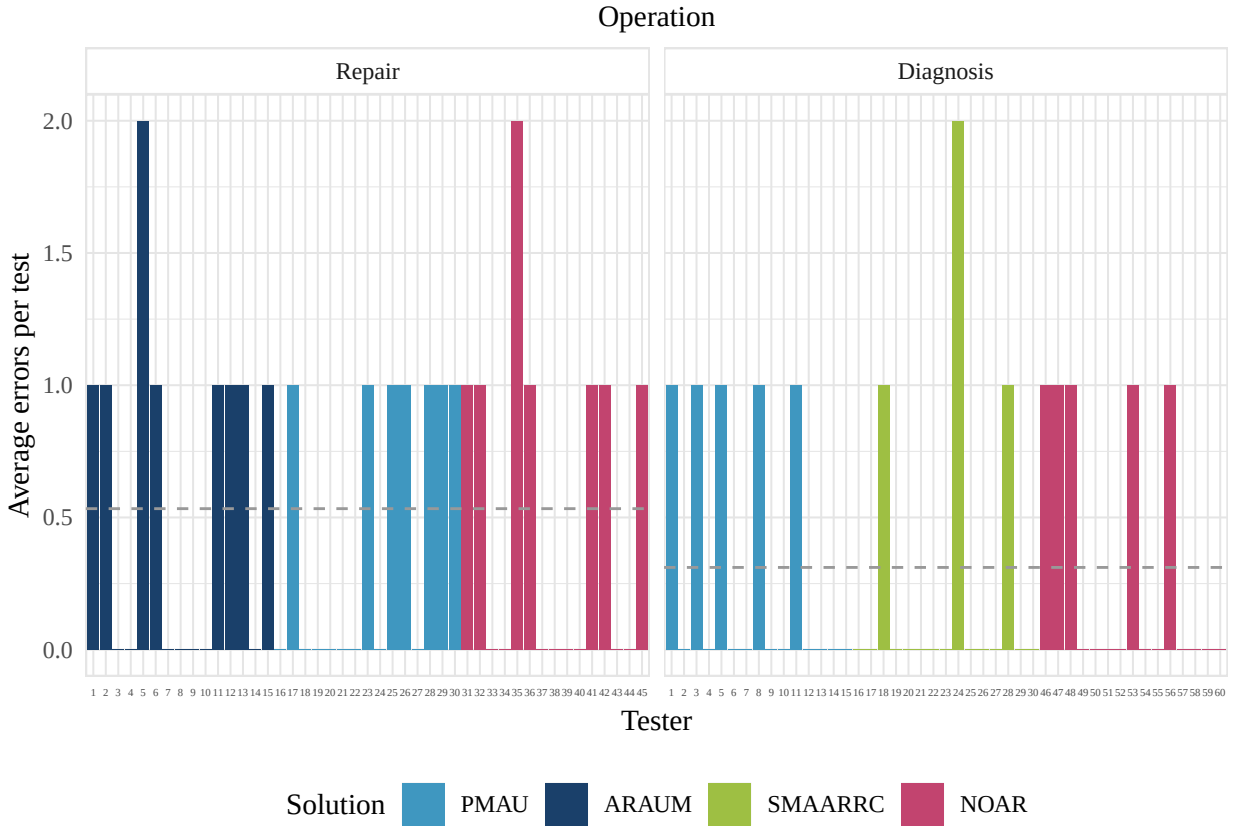
Present results overview with basic statistics. Summarise basic statistics.

```
##      Tester       Operation     Solution      Errors
##  1      : 2   Repair   :45   PMAU   :30   Min.   :0.0000
##  2      : 2   Diagnosis:45   ARAUM  :15   1st Qu.:0.0000
##  3      : 2                  SMAARRC:15   Median :0.0000
##  4      : 2                  NOAR   :30   Mean   :0.4222
##  5      : 2                               3rd Qu.:1.0000
##  6      : 2                               Max.   :2.0000
##  (Other):78
```

Analyse factors group average errors. Calculate mean and standard deviations per factor group (operation and solution).

Operation	Solution	count	mean	sd
Repair	PMAU	15	0.4666667	0.5163978
Repair	ARAUM	15	0.6000000	0.6324555
Repair	NOAR	15	0.5333333	0.6399405
Diagnosis	PMAU	15	0.3333333	0.4879500
Diagnosis	SMAARRC	15	0.2666667	0.5936168
Diagnosis	NOAR	15	0.3333333	0.4879500

Graphically analyse variances per factors group (solution and operation). Plot errors per tester as bar chart and average errors per operation as line.



3.2.2. Variance analysis

Analyse significance of variances on errors results per solution for repair operation. Calculate one-way for errors results per solution.

Repair operation:

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Solution     2  0.133  0.0667   0.186  0.831
## Residuals   42 15.067  0.3587
```

Remote diagnosis operation:

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Solution     2  0.044 0.02222    0.08  0.923
## Residuals   42 11.600 0.27619
```

Analyse significance of variances on errors results per operation. Calculate t-test for errors results per operation.

```
##
##  Welch Two Sample t-test
##
## data:  Errors by Operation
## t = 1.9085, df = 86.483, p-value = 0.05964
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.009230362  0.453674806
## sample estimates:
##    mean in group Repair mean in group Diagnosis
##               0.5333333               0.3111111
```

3.3. Time effect study

3.3.1. Exploratory analysis

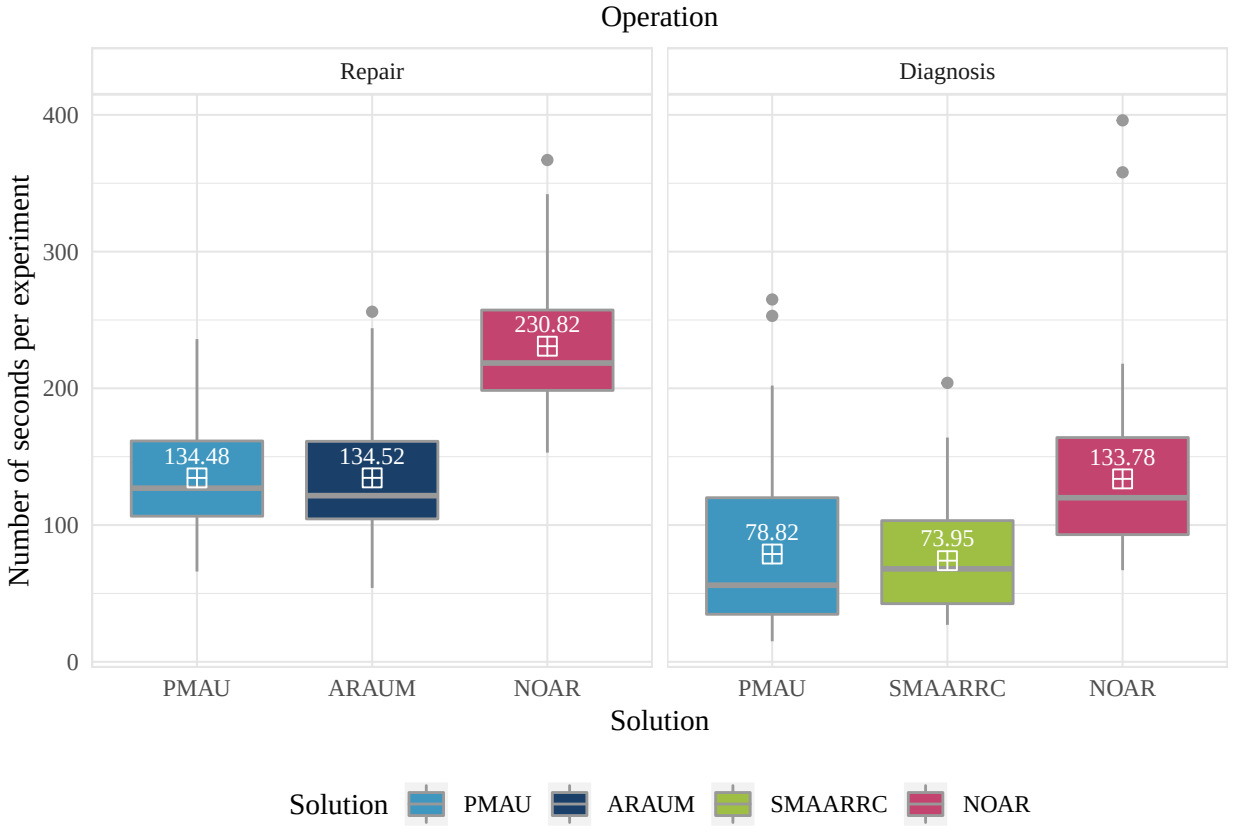
Present results overview with basic statistics. Summarise basic statistics.

```
##      Tester        Operation      Solution       Step       Seconds
##  1      : 8   Repair   :180   PMAU   :120   D1     :45   Min.   : 15.0
##  2      : 8   Diagnosis:180   ARAUM  : 60   D2     :45   1st Qu.: 80.0
##  3      : 8                   SMAARRC: 60   D3     :45   Median :119.0
##  4      : 8                   NOAR   :120   D4     :45   Mean   :131.1
##  5      : 8                                 R1     :45   3rd Qu.:176.2
##  6      : 8                                 R2     :45   Max.   :396.0
##  (Other):312                                (Other):90
```

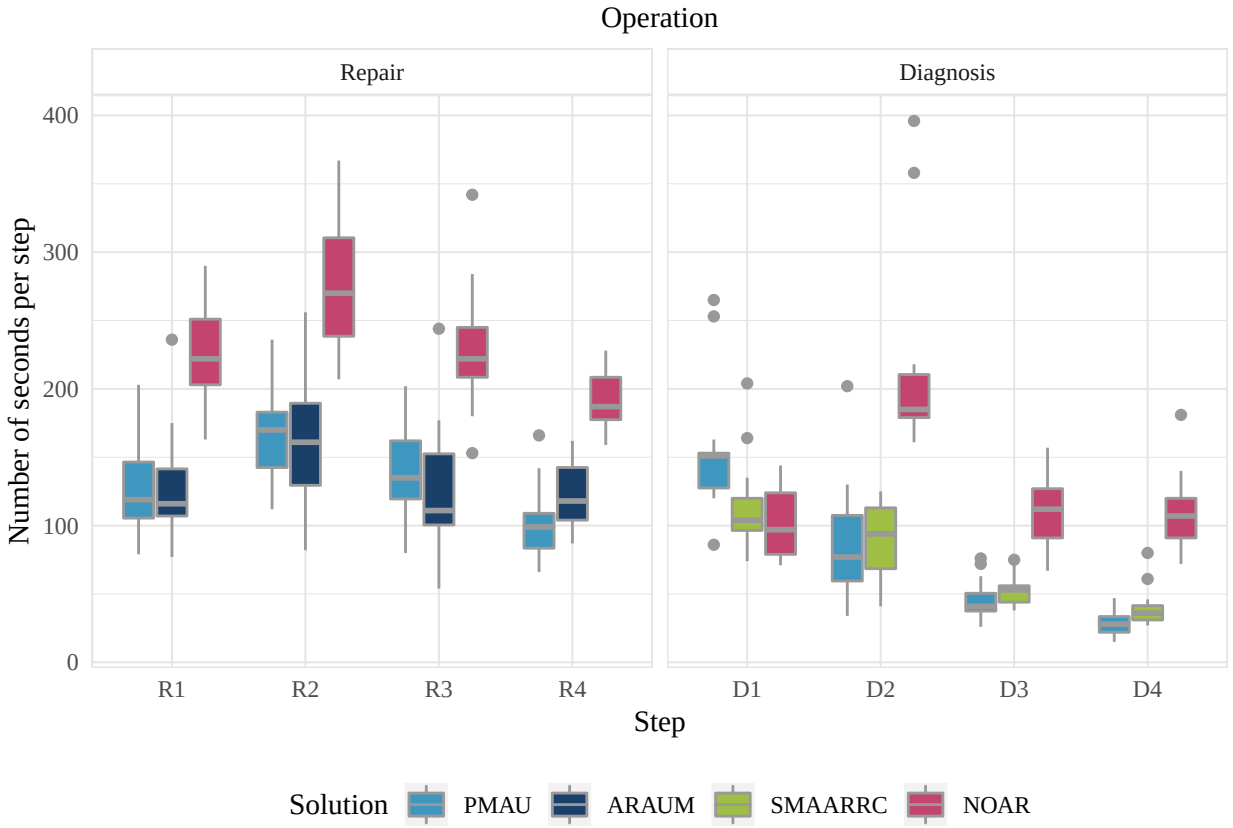
Analyse factors group average seconds. Calculate mean and standard deviations per factor group (operation and solution).

Operation	Solution	count	mean	sd
Repair	PMAU	60	134.48333	39.37154
Repair	ARAUM	60	134.51667	42.42740
Repair	NOAR	60	230.81667	48.69048
Diagnosis	PMAU	60	78.81667	57.75533
Diagnosis	SMAARRC	60	73.95000	37.54790
Diagnosis	NOAR	60	133.78333	61.03100

Graphically analyse variances per factors group (solution and operation). Plot average errors per test as box and whiskers plot per step, solution and operation.



Graphically analyse variances per factors group (solution and operation). Plot average errors per test as box and whiskers plot per step, solution and operation.



3.3.2. Correlation analysis in repair experiment

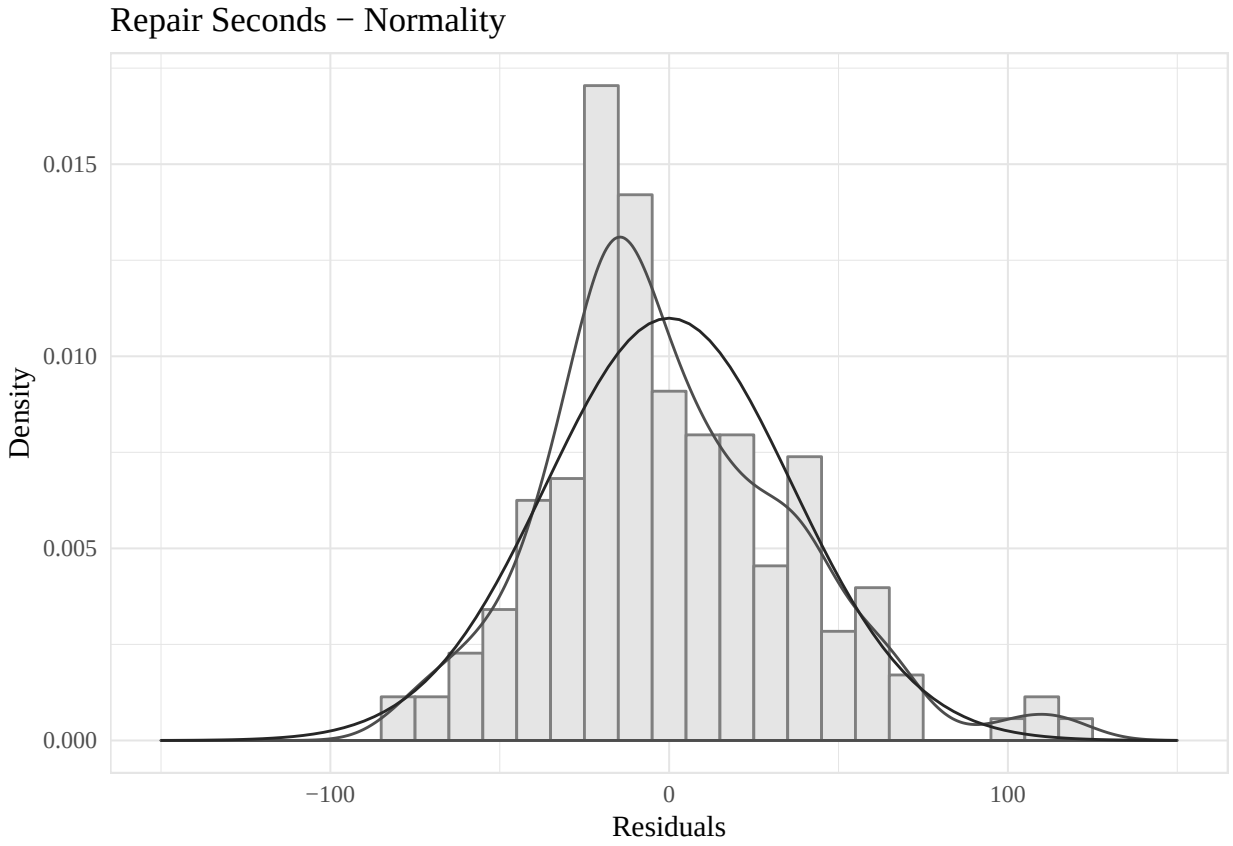
3.3.2.1. Assumptions testing: normality, linearity, homogeneity

Prepare data for in-depth analysis by removing outliers. Use subset function with boxplot stats to manually identify and remove outliers. Fit linear model and calculate residuals and predictors.

```
##     Tester Operation Solution Step Seconds
## 274     39    Repair     NOAR   R2     367
```

Graphically test normallity plotting histogram of residuals. Plot residuals and normal distribution for graphical testing.

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```

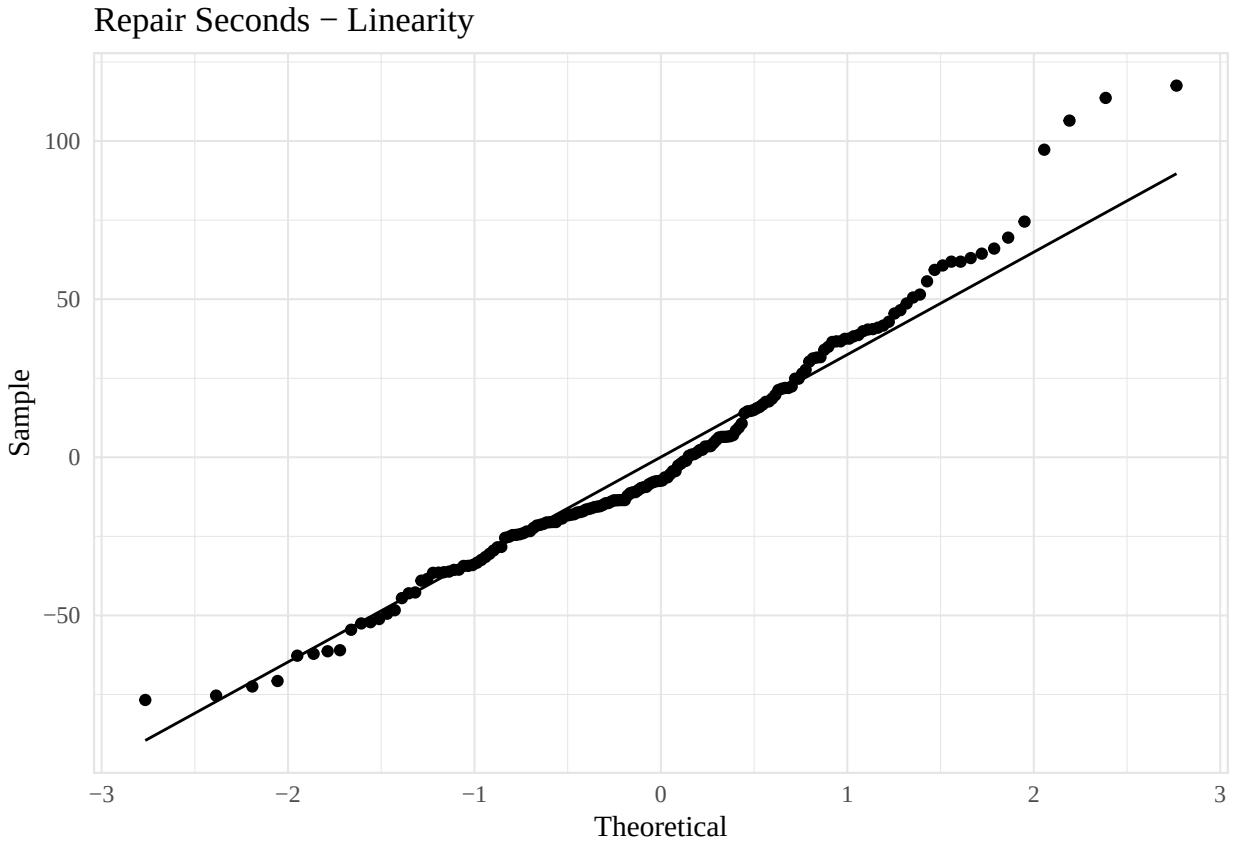



```
## Warning: Removed 2 rows containing missing values (geom_bar).
```

Test normality with shapiro test. Reject null hypothesis with a significance threshold of p-value < 0.05.

```
##
##  Shapiro-Wilk normality test
##
## data:  secondsRClean$Residuals
## W = 0.97318, p-value = 0.00177
```

Graphically test linearity plotting a predicted quantiles versus sample quantiles. Plot residuals and samples and check against diagonal for graphical testing.



Test homogeneity assumption with Bartlett test. Reject null hypothesis with a significance threshold of p-value < 0.05.

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Seconds by interaction(Step, Solution)
## Bartlett's K-squared = 19.7, df = 11, p-value = 0.04963
```

3.3.2.2. Anova analysis

Test correlation between response (seconds) and effects (step, solution) for repair operations. Conduct two-way anova.

```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## Step            3  86309   28770   20.12 3.42e-11 ***
## Solution        2 371076  185538  129.79  < 2e-16 ***
## Step:Solution   6  11061    1843    1.29    0.265
## Residuals     168 240168    1430
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test differences between factor groups means using Tukey HSD test. Reject null hypotheses with a significance threshold of p-adj-value < 0.05.

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Seconds ~ Step:Solution, data = secondsR)
```

```
##
## $`Step:Solution`
##                           diff         lwr         upr     p adj
## R2:PMAU-R1:PMAU     38.066667   -7.697203  83.8305364 0.2084755
## R3:PMAU-R1:PMAU     12.866667  -32.897203  58.6305364 0.9986928
## R4:PMAU-R1:PMAU    -26.866667  -72.630536  18.8972031 0.7285119
## R1:ARAUM-R1:PMAU     1.066667  -44.697203  46.8305364 1.0000000
## R2:ARAUM-R1:PMAU    30.266667  -15.497203  76.0305364 0.5575761
## R3:ARAUM-R1:PMAU    -2.000000  -47.763870  43.7638697 1.0000000
## R4:ARAUM-R1:PMAU    -5.133333  -50.897203  40.6305364 0.9999999
## R1:NOAR-R1:PMAU     96.866667   51.102797 142.6305364 0.0000000
## R2:NOAR-R1:PMAU    147.200000  101.436130 192.9638697 0.0000000
## R3:NOAR-R1:PMAU    101.133333   55.369464 146.8972031 0.0000000
## R4:NOAR-R1:PMAU     64.200000   18.436130 109.9638697 0.0004067
## R3:PMAU-R2:PMAU    -25.200000  -70.963870  20.5638697 0.8018370
## R4:PMAU-R2:PMAU    -64.933333 -110.697203 -19.1694636 0.0003252
## R1:ARAUM-R2:PMAU   -37.000000  -82.763870   8.7638697 0.2455961
## R2:ARAUM-R2:PMAU    -7.800000  -53.563870  37.9638697 0.9999904
## R3:ARAUM-R2:PMAU   -40.066667  -85.830536   5.6972031 0.1497810
## R4:ARAUM-R2:PMAU   -43.200000  -88.963870   2.5638697 0.0841777
## R1:NOAR-R2:PMAU     58.800000   13.036130 104.5638697 0.0019728
## R2:NOAR-R2:PMAU    109.133333   63.369464 154.8972031 0.0000000
## R3:NOAR-R2:PMAU     63.066667   17.302797 108.8305364 0.0005723
## R4:NOAR-R2:PMAU     26.133333  -19.630536  71.8972031 0.7620034
## R4:PMAU-R3:PMAU    -39.733333  -85.497203   6.0305364 0.1585951
## R1:ARAUM-R3:PMAU   -11.800000  -57.563870  33.9638697 0.9994153
## R2:ARAUM-R3:PMAU    17.400000  -28.363870  63.1638697 0.9827918
## R3:ARAUM-R3:PMAU   -14.866667  -60.630536  30.8972031 0.9952800
## R4:ARAUM-R3:PMAU   -18.000000  -63.763870  27.7638697 0.9776432
## R1:NOAR-R3:PMAU     84.000000   38.236130 129.7638697 0.0000005
## R2:NOAR-R3:PMAU    134.333333   88.569464 180.0972031 0.0000000
## R3:NOAR-R3:PMAU     88.266667   42.502797 134.0305364 0.0000001
## R4:NOAR-R3:PMAU     51.333333    5.569464  97.0972031 0.0140417
## R1:ARAUM-R4:PMAU    27.933333  -17.830536  73.6972031 0.6770088
## R2:ARAUM-R4:PMAU    57.133333   11.369464 102.8972031 0.0031298
## R3:ARAUM-R4:PMAU    24.866667  -20.897203  70.6305364 0.8152152
## R4:ARAUM-R4:PMAU    21.733333  -24.030536  67.4972031 0.9159807
## R1:NOAR-R4:PMAU    123.733333   77.969464 169.4972031 0.0000000
## R2:NOAR-R4:PMAU    174.066667  128.302797 219.8305364 0.0000000
## R3:NOAR-R4:PMAU    128.000000   82.236130 173.7638697 0.0000000
## R4:NOAR-R4:PMAU     91.066667   45.302797 136.8305364 0.0000000
## R2:ARAUM-R1:ARAUM   29.200000  -16.563870  74.9638697 0.6128492
## R3:ARAUM-R1:ARAUM   -3.066667  -48.830536  42.6972031 1.0000000
## R4:ARAUM-R1:ARAUM   -6.200000  -51.963870  39.5638697 0.9999991
## R1:NOAR-R1:ARAUM    95.800000   50.036130 141.5638697 0.0000000
## R2:NOAR-R1:ARAUM   146.133333  100.369464 191.8972031 0.0000000
## R3:NOAR-R1:ARAUM   100.066667   54.302797 145.8305364 0.0000000
## R4:NOAR-R1:ARAUM    63.133333   17.369464 108.8972031 0.0005610
## R3:ARAUM-R2:ARAUM  -32.266667  -78.030536  13.4972031 0.4548437
## R4:ARAUM-R2:ARAUM  -35.400000  -81.163870  10.3638697 0.3087745
## R1:NOAR-R2:ARAUM    66.600000   20.836130 112.3638697 0.0001940
## R2:NOAR-R2:ARAUM   116.933333   71.169464 162.6972031 0.0000000
## R3:NOAR-R2:ARAUM    70.866667   25.102797 116.6305364 0.0000493
## R4:NOAR-R2:ARAUM    33.933333  -11.830536  79.6972031 0.3739598
```

```
## R4:ARAUM-R3:ARAUM  -3.133333  -48.897203  42.6305364 1.0000000
## R1:NOAR-R3:ARAUM   98.866667   53.102797 144.6305364 0.0000000
## R2:NOAR-R3:ARAUM  149.200000  103.436130 194.9638697 0.0000000
## R3:NOAR-R3:ARAUM  103.133333   57.369464 148.8972031 0.0000000
## R4:NOAR-R3:ARAUM   66.200000   20.436130 111.9638697 0.0002199
## R1:NOAR-R4:ARAUM  102.000000   56.236130 147.7638697 0.0000000
## R2:NOAR-R4:ARAUM  152.333333  106.569464 198.0972031 0.0000000
## R3:NOAR-R4:ARAUM  106.266667   60.502797 152.0305364 0.0000000
## R4:NOAR-R4:ARAUM   69.333333   23.569464 115.0972031 0.0000813
## R2:NOAR-R1:NOAR    50.333333    4.569464  96.0972031 0.0178705
## R3:NOAR-R1:NOAR     4.266667  -41.497203  50.0305364 1.0000000
## R4:NOAR-R1:NOAR   -32.666667  -78.430536  13.0972031 0.4348966
## R3:NOAR-R2:NOAR   -46.066667  -91.830536  -0.3027969 0.0468862
## R4:NOAR-R2:NOAR   -83.000000 -128.763870 -37.2361303 0.0000007
## R4:NOAR-R3:NOAR   -36.933333  -82.697203   8.8305364 0.2480510
```

3.3.3. Correlation analysis in remote diagnosis experiment

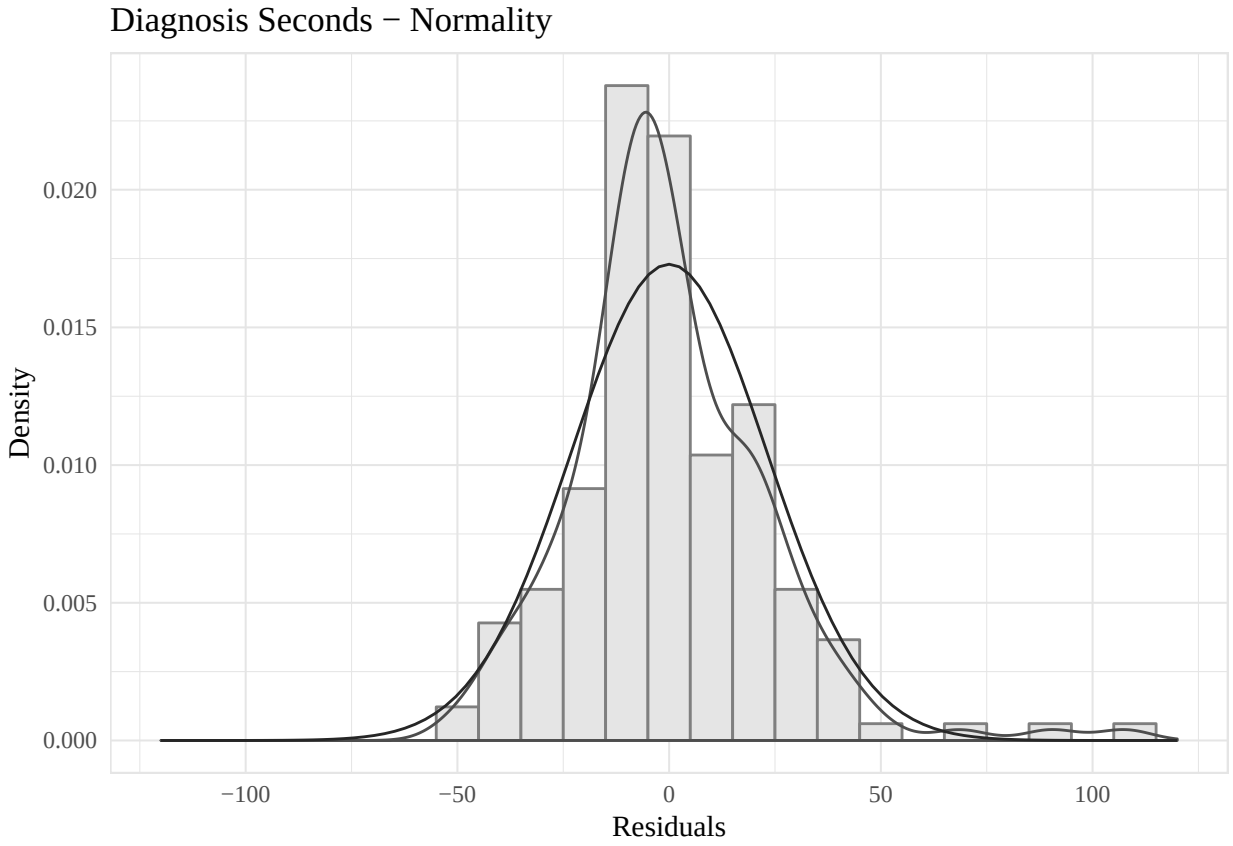
3.3.3.1. Assumptions testing: normality, linearity, homogeneity

Prepare data for in-depth analysis by removing outliers. Use subset function with boxplot stats to manually identify and remove outliers. Fit linear model and calculate residuals and predictors.

```
##     Tester Operation Solution Step Seconds
## 41       6 Diagnosis     PMAU   D1     253
## 105     14 Diagnosis     PMAU   D1     265
## 306     47 Diagnosis     NOAR   D2     396
## 314     49 Diagnosis     NOAR   D2     358
```

Graphically test normallity plotting histogram of residuals. Plot residuals and normal distribution for graphical testing.

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```

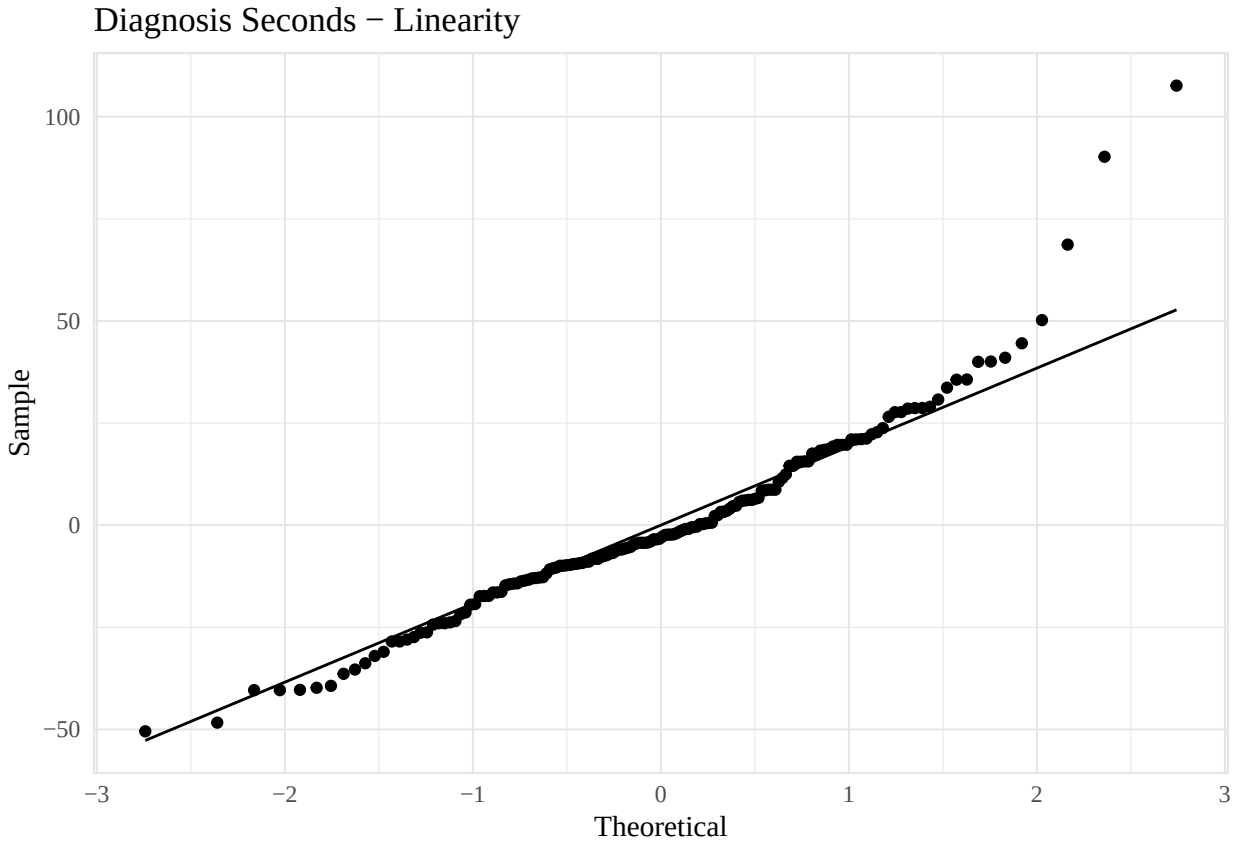



```
## Warning: Removed 2 rows containing missing values (geom_bar).
```

Test normality with shapiro test. Reject null hypothesis with a significance threshold of p-value < 0.05.

```
##
##  Shapiro-Wilk normality test
##
## data:  secondsDClean$Residuals
## W = 0.94081, p-value = 2.445e-06
```

Graphically test linearity plotting a predicted quantiles versus sample quantiles. Plot residuals and samples and check against diagonal for graphical testing.



Test homogeneity assumption with Bartlett test. Reject null hypothesis with a significance threshold of p-value < 0.05.

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Seconds by interaction(Step, Solution)
## Bartlett's K-squared = 49.86, df = 11, p-value = 6.632e-07
```

3.3.3.2. Anova analysis

Test correlation between response (seconds) and effects (step, solution) for repair operations. Conduct two-way anova.

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## Step            3 176247   58749   53.08 <2e-16 ***
## Solution        2 132501   66250   59.86 <2e-16 ***
## Step:Solution   6 137561   22927   20.71 <2e-16 ***
## Residuals     168 185940    1107
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test differences between factor groups means using Tukey HSD test. Reject null hypotheses with a significance threshold of p-adj-value < 0.05.

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Seconds ~ Step:Solution, data = secondsD)
```

```
##
## $`Step:Solution`
##                                diff         lwr         upr     p adj
## D2:PMAU-D1:PMAU        -64.666667 -104.933946 -24.3993870 0.0000206
## D3:PMAU-D1:PMAU       -107.200000 -147.467280 -66.9327203 0.0000000
## D4:PMAU-D1:PMAU       -124.066667 -164.333946 -83.7993870 0.0000000
## D1:SMAARRC-D1:PMAU     -39.000000  -79.267280   1.2672797 0.0673504
## D2:SMAARRC-D1:PMAU     -63.466667 -103.733946 -23.1993870 0.0000325
## D3:SMAARRC-D1:PMAU    -100.066667 -140.333946 -59.7993870 0.0000000
## D4:SMAARRC-D1:PMAU    -112.866667 -153.133946 -72.5993870 0.0000000
## D1:NOAR-D1:PMAU        -50.600000  -90.867280 -10.3327203 0.0028253
## D2:NOAR-D1:PMAU         59.733333   19.466054 100.0006130 0.0001293
## D3:NOAR-D1:PMAU        -41.000000  -81.267280  -0.7327203 0.0418387
## D4:NOAR-D1:PMAU        -44.200000  -84.467280  -3.9327203 0.0183008
## D3:PMAU-D2:PMAU        -42.533333  -82.800613  -2.2660536 0.0284246
## D4:PMAU-D2:PMAU        -59.400000  -99.667280 -19.1327203 0.0001458
## D1:SMAARRC-D2:PMAU      25.666667  -14.600613  65.9339464 0.6143815
## D2:SMAARRC-D2:PMAU       1.200000  -39.067280  41.4672797 1.0000000
## D3:SMAARRC-D2:PMAU     -35.400000  -75.667280   4.8672797 0.1455439
## D4:SMAARRC-D2:PMAU     -48.200000  -88.467280  -7.9327203 0.0058737
## D1:NOAR-D2:PMAU         14.066667  -26.200613  54.3339464 0.9912928
## D2:NOAR-D2:PMAU        124.400000   84.132720 164.6672797 0.0000000
## D3:NOAR-D2:PMAU         23.666667  -16.600613  63.9339464 0.7270790
## D4:NOAR-D2:PMAU         20.466667  -19.800613  60.7339464 0.8727124
## D4:PMAU-D3:PMAU        -16.866667  -57.133946  23.4006130 0.9643077
## D1:SMAARRC-D3:PMAU      68.200000   27.932720 108.4672797 0.0000052
## D2:SMAARRC-D3:PMAU      43.733333    3.466054  84.0006130 0.0207446
## D3:SMAARRC-D3:PMAU       7.133333  -33.133946  47.4006130 0.9999857
## D4:SMAARRC-D3:PMAU      -5.666667  -45.933946  34.6006130 0.9999987
## D1:NOAR-D3:PMAU         56.600000   16.332720  96.8672797 0.0003915
## D2:NOAR-D3:PMAU        166.933333  126.666054 207.2006130 0.0000000
## D3:NOAR-D3:PMAU         66.200000   25.932720 106.4672797 0.0000114
## D4:NOAR-D3:PMAU         63.000000   22.732720 103.2672797 0.0000388
## D1:SMAARRC-D4:PMAU      85.066667   44.799387 125.3339464 0.0000000
## D2:SMAARRC-D4:PMAU      60.600000   20.332720 100.8672797 0.0000944
## D3:SMAARRC-D4:PMAU      24.000000  -16.267280  64.2672797 0.7090935
## D4:SMAARRC-D4:PMAU      11.200000  -29.067280  51.4672797 0.9988157
## D1:NOAR-D4:PMAU         73.466667   33.199387 113.7339464 0.0000006
## D2:NOAR-D4:PMAU        183.800000  143.532720 224.0672797 0.0000000
## D3:NOAR-D4:PMAU         83.066667   42.799387 123.3339464 0.0000000
## D4:NOAR-D4:PMAU         79.866667   39.599387 120.1339464 0.0000000
## D2:SMAARRC-D1:SMAARRC  -24.466667  -64.733946  15.8006130 0.6832818
## D3:SMAARRC-D1:SMAARRC  -61.066667 -101.333946 -20.7993870 0.0000795
## D4:SMAARRC-D1:SMAARRC  -73.866667 -114.133946 -33.5993870 0.0000005
## D1:NOAR-D1:SMAARRC     -11.600000  -51.867280  28.6672797 0.9983684
## D2:NOAR-D1:SMAARRC      98.733333   58.466054 139.0006130 0.0000000
## D3:NOAR-D1:SMAARRC      -2.000000  -42.267280  38.2672797 1.0000000
## D4:NOAR-D1:SMAARRC      -5.200000  -45.467280  35.0672797 0.9999995
## D3:SMAARRC-D2:SMAARRC  -36.600000  -76.867280   3.6672797 0.1140375
## D4:SMAARRC-D2:SMAARRC  -49.400000  -89.667280  -9.1327203 0.0040919
## D1:NOAR-D2:SMAARRC      12.866667  -27.400613  53.1339464 0.9959074
## D2:NOAR-D2:SMAARRC     123.200000   82.932720 163.4672797 0.0000000
## D3:NOAR-D2:SMAARRC      22.466667  -17.800613  62.7339464 0.7879954
## D4:NOAR-D2:SMAARRC      19.266667  -21.000613  59.5339464 0.9118925
```

```
## D4:SMAARRC-D3:SMAARRC  -12.800000  -53.067280  27.4672797 0.9960878
## D1:NOAR-D3:SMAARRC      49.466667    9.199387  89.7339464 0.0040095
## D2:NOAR-D3:SMAARRC     159.800000  119.532720 200.0672797 0.0000000
## D3:NOAR-D3:SMAARRC      59.066667   18.799387  99.3339464 0.0001643
## D4:NOAR-D3:SMAARRC      55.866667   15.599387  96.1339464 0.0005037
## D1:NOAR-D4:SMAARRC      62.266667   21.999387 102.5339464 0.0000510
## D2:NOAR-D4:SMAARRC     172.600000  132.332720 212.8672797 0.0000000
## D3:NOAR-D4:SMAARRC      71.866667   31.599387 112.1339464 0.0000012
## D4:NOAR-D4:SMAARRC      68.666667   28.399387 108.9339464 0.0000043
## D2:NOAR-D1:NOAR        110.333333   70.066054 150.6006130 0.0000000
## D3:NOAR-D1:NOAR          9.600000  -30.667280  49.8672797 0.9997232
## D4:NOAR-D1:NOAR          6.400000  -33.867280  46.6672797 0.9999953
## D3:NOAR-D2:NOAR       -100.733333 -141.000613 -60.4660536 0.0000000
## D4:NOAR-D2:NOAR       -103.933333 -144.200613 -63.6660536 0.0000000
## D4:NOAR-D3:NOAR         -3.200000  -43.467280  37.0672797 1.0000000
```

3.4. Usability study

Present results overview with basic statistics. Summarise data structure and basic statistics.

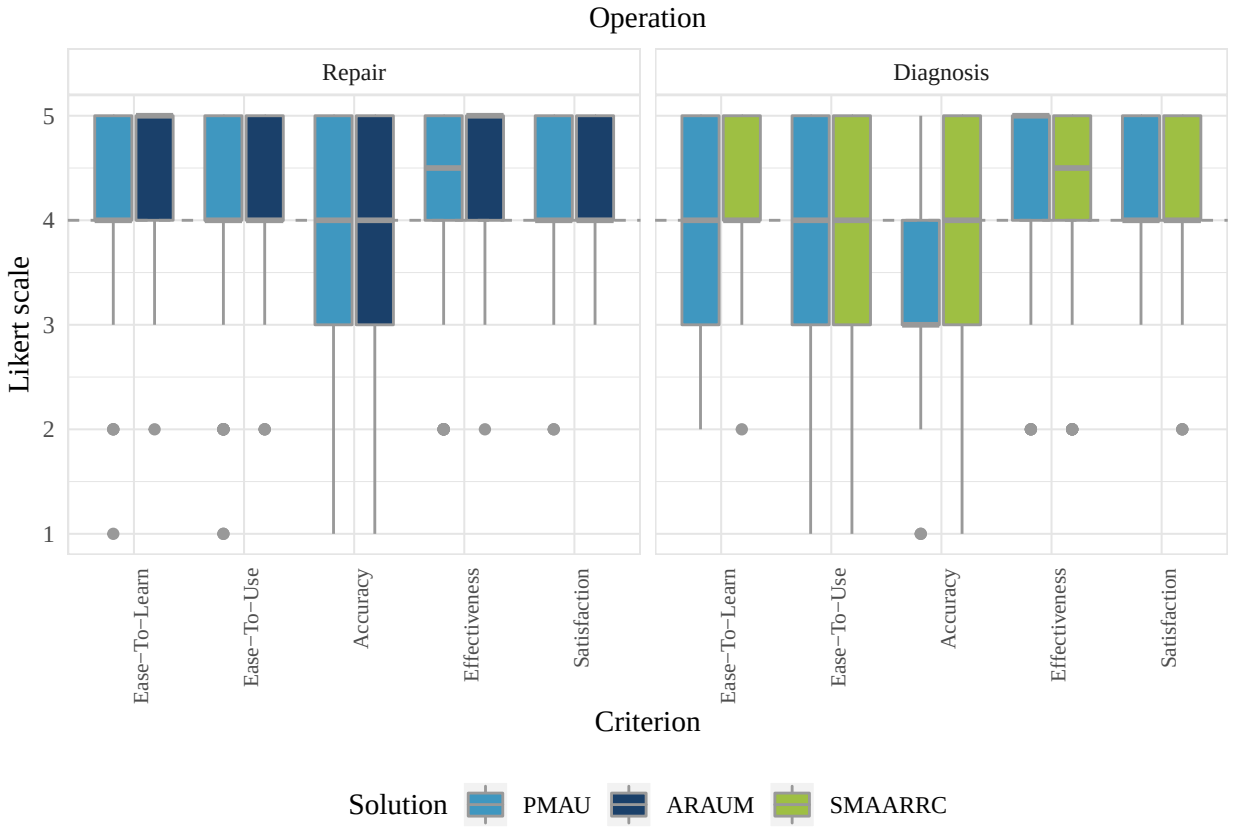
```
## 'data.frame':    1440 obs. of  6 variables:
##  $ Tester   : Factor w/ 30 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Operation: Ord.factor w/ 2 levels "Repair"<"Diagnosis": 1 2 1 2 1 2 1 2 1 2 ...
##  $ Solution : Ord.factor w/ 4 levels "PMAU"<"ARAUM"<..: 2 1 2 1 2 1 2 1 2 1 ...
##  $ Criterion: Ord.factor w/ 5 levels "Ease-To-Learn"<..: 1 1 1 1 1 1 2 2 2 2 ...
##  $ Aspect   : Factor w/ 24 levels "Animations","Buttons",..: 21 21 7 7 14 14 2 2 10 10 ...
##  $ Response : int  5 4 5 5 5 5 4 4 NA 5 ...
```

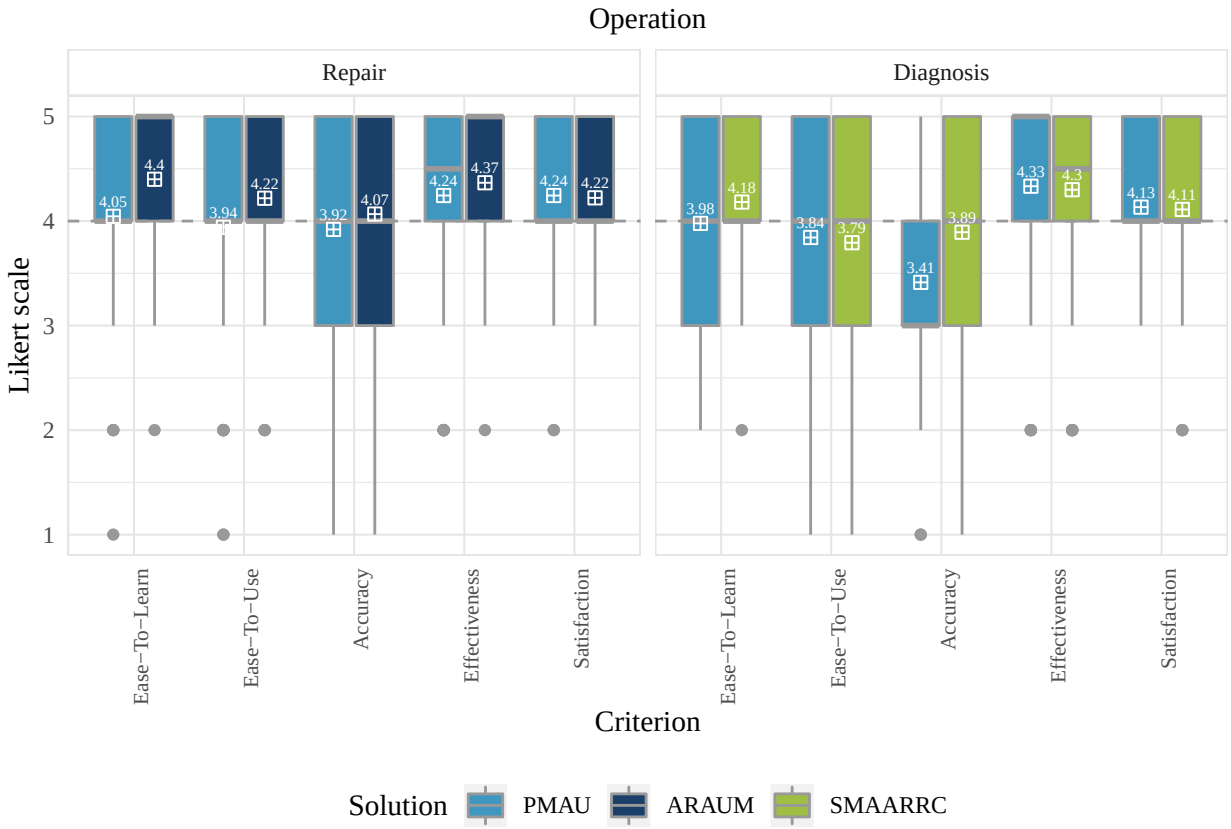
```
##      Tester         Operation      Solution           Criterion
##  1      :  48   Repair   :720   PMAU   :720   Ease-To-Learn:180
##  2      :  48   Diagnosis:720   ARAUM  :360   Ease-To-Use  :420
##  3      :  48                   SMAARRC:360   Accuracy     :300
##  4      :  48                   NOAR   :  0   Effectiveness:360
##  5      :  48                                 Satisfaction :180
##  6      :  48
##  (Other):1152
##                  Aspect        Response
##  Animations         :  60   Min.   :1.000
##  Buttons            :  60   1st Qu.:4.000
##  Confidence-Increase:  60   Median :4.000
##  Content-Suitability:  60   Mean   :4.074
##  Design             :  60   3rd Qu.:5.000
##  Efficiency-Increase:  60   Max.   :5.000
##  (Other)            :1080   NA's   :80
```

Analyse average responses per criterion, solution and operation. Calculate mean and standard deviations per factor group.

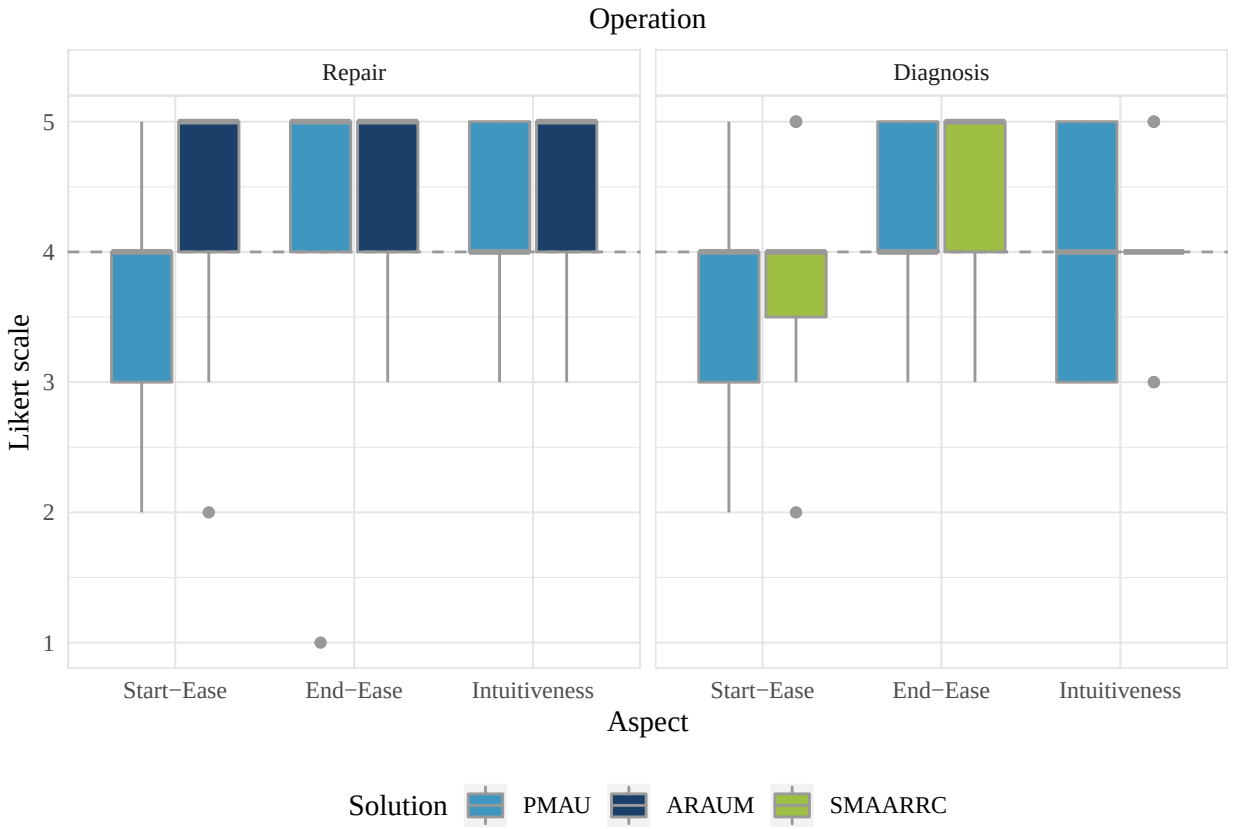
Operation	Criterion	Solution	count	mean	sd
Repair	Ease-To-Learn	PMAU	45	4.045454	1.0105155
Repair	Ease-To-Learn	ARAUM	45	4.400000	0.7804428
Repair	Ease-To-Use	PMAU	105	3.940476	0.9737666
Repair	Ease-To-Use	ARAUM	105	4.218391	0.8411687
Repair	Accuracy	PMAU	75	3.920000	0.9552133
Repair	Accuracy	ARAUM	75	4.066667	0.9772180
Repair	Effectiveness	PMAU	90	4.244444	0.9278598
Repair	Effectiveness	ARAUM	90	4.366667	0.7854005
Repair	Satisfaction	PMAU	45	4.244444	0.7433204
Repair	Satisfaction	ARAUM	45	4.222222	0.7035265
Diagnosis	Ease-To-Learn	PMAU	45	3.977778	0.8915994
Diagnosis	Ease-To-Learn	SMAARRC	45	4.181818	0.7555293
Diagnosis	Ease-To-Use	PMAU	105	3.842697	1.0101629
Diagnosis	Ease-To-Use	SMAARRC	105	3.792683	1.1081940
Diagnosis	Accuracy	PMAU	75	3.413333	1.0012605
Diagnosis	Accuracy	SMAARRC	75	3.893333	1.0851994
Diagnosis	Effectiveness	PMAU	90	4.333333	0.8740966
Diagnosis	Effectiveness	SMAARRC	90	4.300000	0.8669979
Diagnosis	Satisfaction	PMAU	45	4.133333	0.6941312
Diagnosis	Satisfaction	SMAARRC	45	4.111111	0.8040303

Graphically analyse responses averages for each criterion per operation. Plot average responses count per tester as box and whiskers per criterion, solution and operation with conservative average for Likert scale.

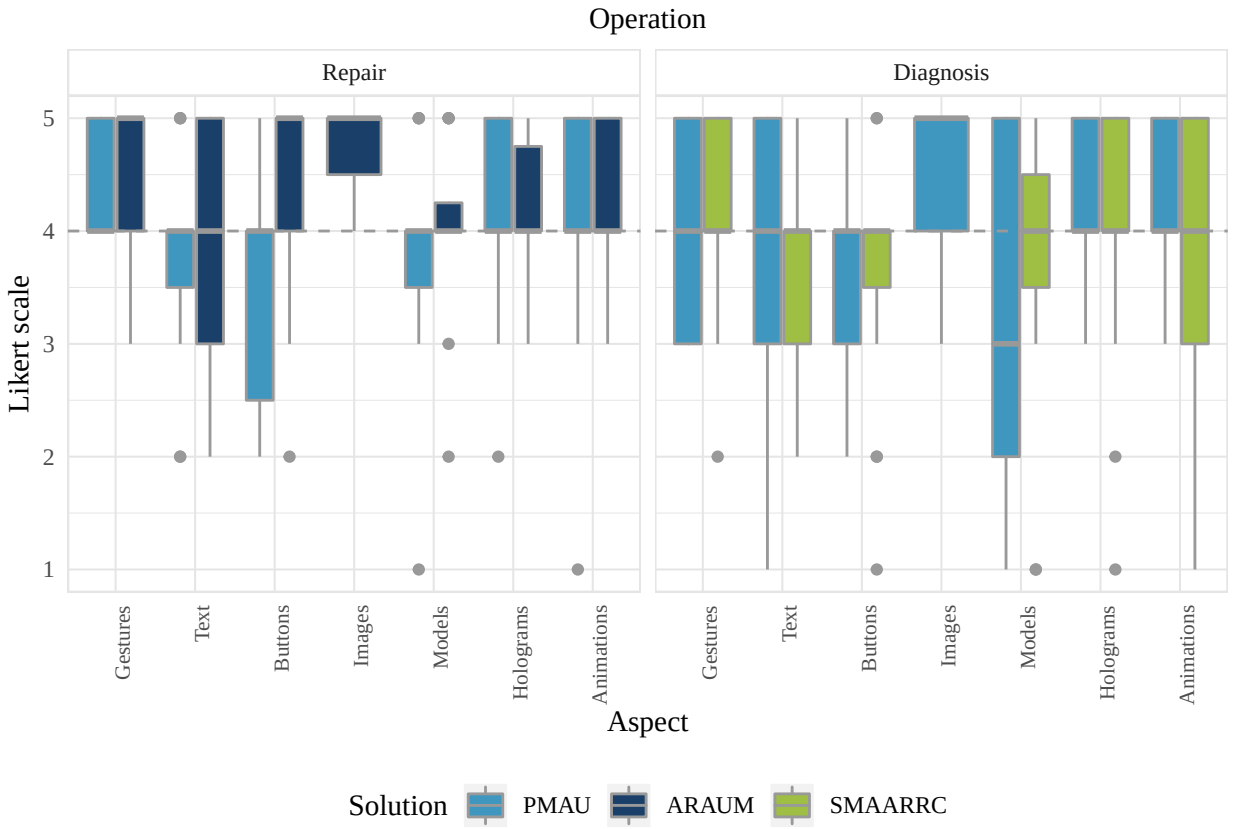




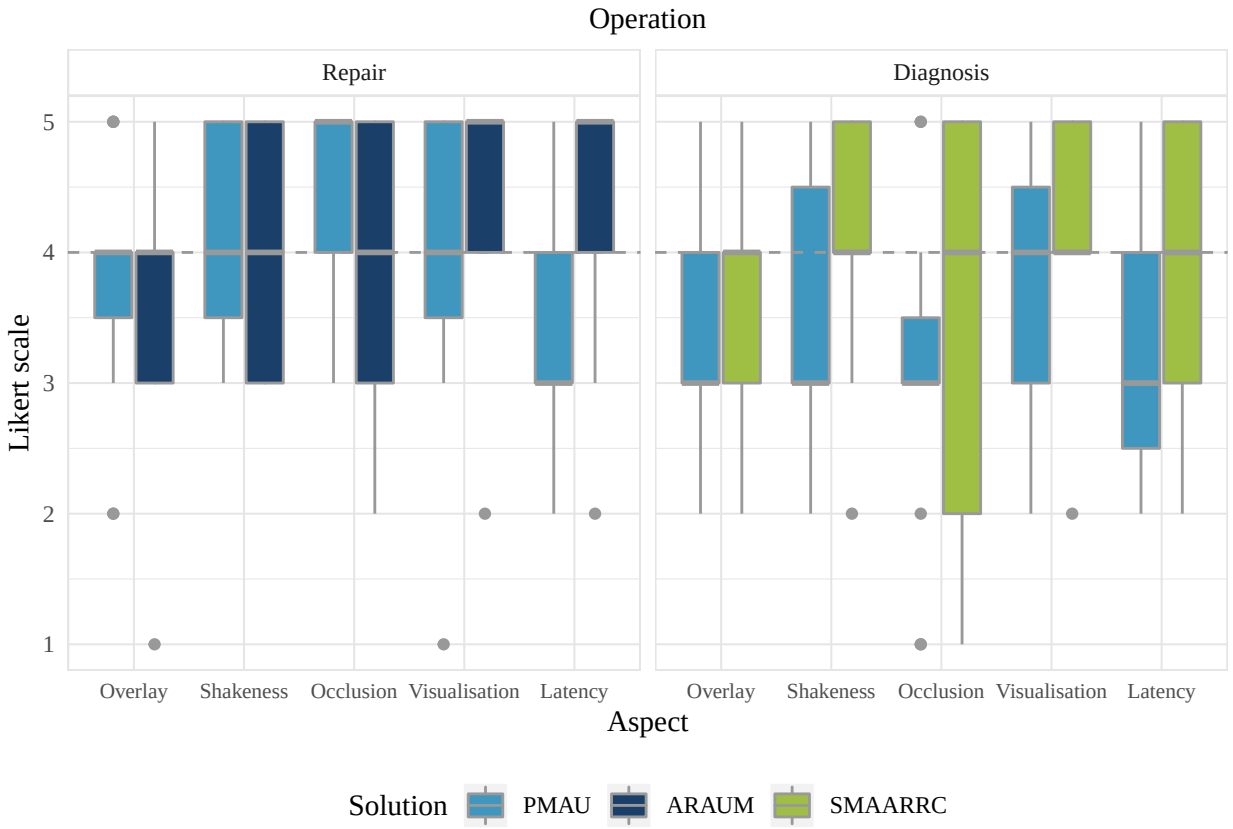
Graphically analyse responses averages for each aspect regarding Ease-To-Learn criterion per operation. Plot average responses count per tester as box and whiskers per aspect, solution and operation with conservative average for Likert scale.



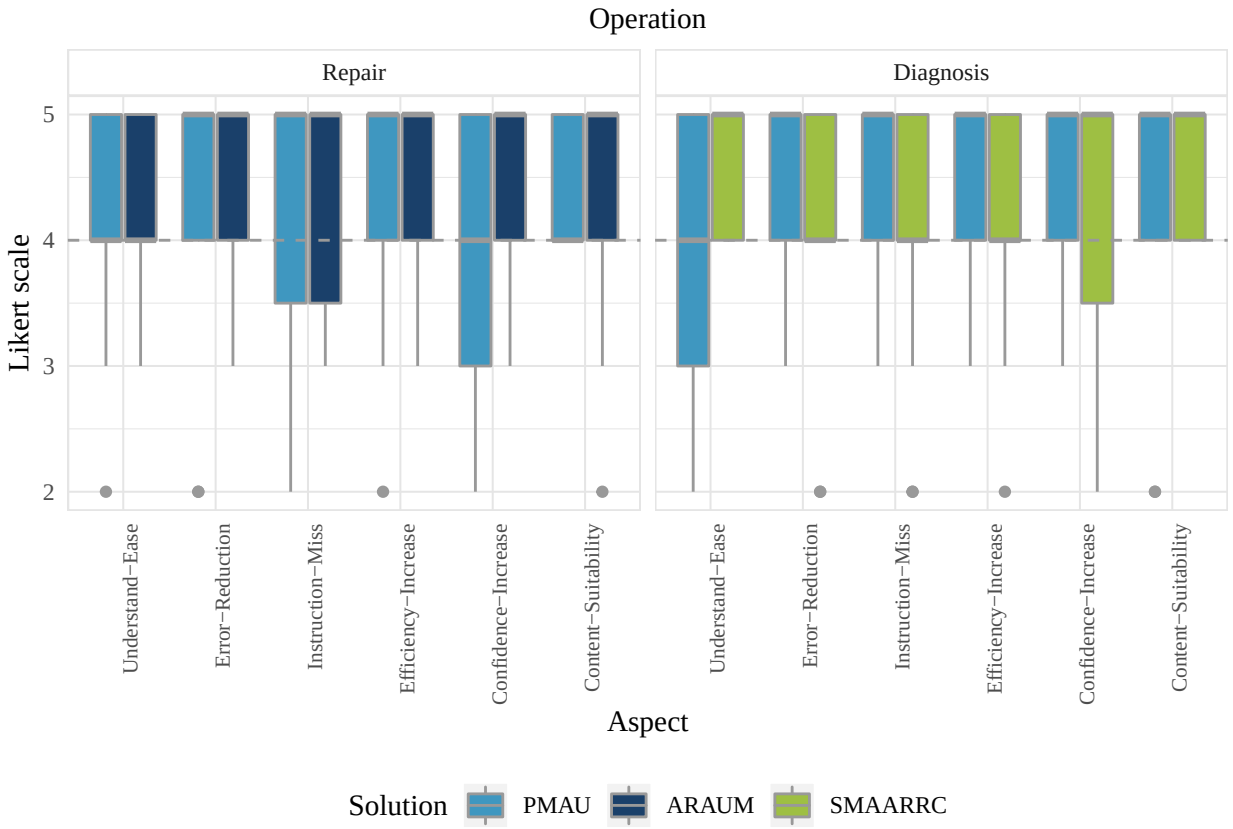
Graphically analyse responses averages for each aspect regarding Ease-To-Use criterion per operation. Plot average responses count per tester as box and whiskers per aspect, solution and operation with conservative average for Likert scale.



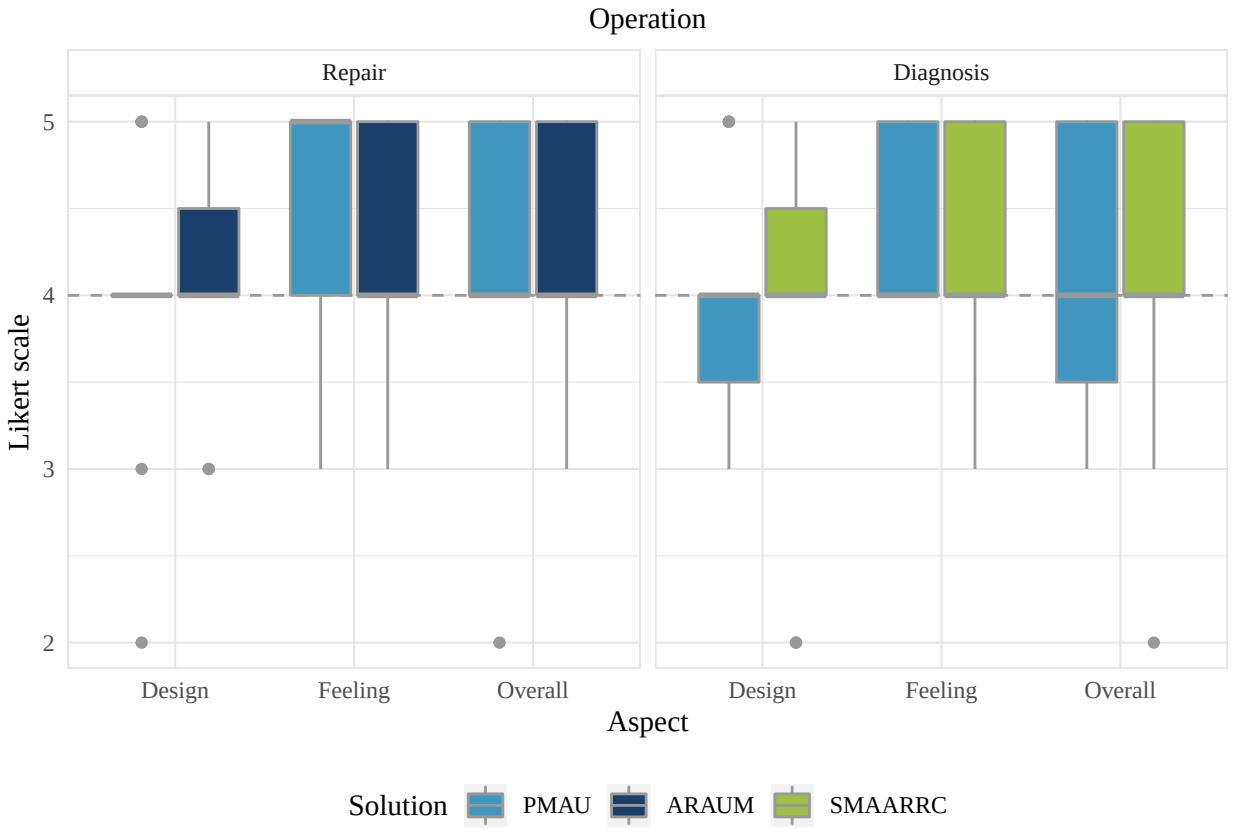
Graphically analyse responses averages for each aspect regarding Accuracy criterion per operation. Plot average responses count per tester as box and whiskers per aspect, solution and operation with conservative average for Likert scale.



Graphically analyse responses averages for each aspect regarding Effectiveness criterion per operation. Plot average responses count per tester as box and whiskers per aspect, solution and operation with conservative average for Likert scale.



Graphically analyse responses averages for each aspect regarding Satisfaction criterion per operation. Plot average responses count per tester as box and whiskers per aspect, solution and operation with conservative average for Likert scale.



Analyse average responses per aspect, criterion, solution and operation. Calculate mean and standard deviations per factor group.

Operation	Criterion	Aspect	Solution	count	mean	sd
Repair	Ease-To-Learn	End-Ease	PMAU	15	4.400000	1.0555973
Repair	Ease-To-Learn	End-Ease	ARAUM	15	4.533333	0.6399405
Repair	Ease-To-Learn	Intuitiveness	PMAU	15	4.214286	0.6992932
Repair	Ease-To-Learn	Intuitiveness	ARAUM	15	4.333333	0.8164966
Repair	Ease-To-Learn	Start-Ease	PMAU	15	3.533333	1.0600988
Repair	Ease-To-Learn	Start-Ease	ARAUM	15	4.333333	0.8997354
Repair	Ease-To-Use	Animations	PMAU	15	4.133333	1.0600988
Repair	Ease-To-Use	Animations	ARAUM	15	4.133333	0.7432234
Repair	Ease-To-Use	Buttons	PMAU	15	3.533333	1.1254629
Repair	Ease-To-Use	Buttons	ARAUM	15	4.466667	0.9154754
Repair	Ease-To-Use	Gestures	PMAU	15	4.400000	0.5070926
Repair	Ease-To-Use	Gestures	ARAUM	15	4.461538	0.7762500
Repair	Ease-To-Use	Holograms	PMAU	15	4.000000	1.0000000
Repair	Ease-To-Use	Holograms	ARAUM	15	4.071429	0.7300459
Repair	Ease-To-Use	Images	PMAU	15	NaN	NaN
Repair	Ease-To-Use	Images	ARAUM	15	4.666667	0.5773503
Repair	Ease-To-Use	Models	PMAU	15	3.800000	1.0141851
Repair	Ease-To-Use	Models	ARAUM	15	4.000000	0.8528029
Repair	Ease-To-Use	Text	PMAU	15	3.800000	0.9411239
Repair	Ease-To-Use	Text	ARAUM	15	4.066667	1.0327956
Repair	Accuracy	Latency	PMAU	15	3.266667	0.7988086
Repair	Accuracy	Latency	ARAUM	15	4.333333	0.9759001
Repair	Accuracy	Occlusion	PMAU	15	4.400000	0.7367884
Repair	Accuracy	Occlusion	ARAUM	15	3.933333	1.0997835
Repair	Accuracy	Overlay	PMAU	15	3.800000	0.9411239
Repair	Accuracy	Overlay	ARAUM	15	3.600000	0.9856108
Repair	Accuracy	Shakeness	PMAU	15	4.200000	0.8618916
Repair	Accuracy	Shakeness	ARAUM	15	4.066667	0.8837151
Repair	Accuracy	Visualisation	PMAU	15	3.933333	1.0997835
Repair	Accuracy	Visualisation	ARAUM	15	4.400000	0.8280787
Repair	Effectiveness	Confidence-Increase	PMAU	15	4.000000	1.1338934
Repair	Effectiveness	Confidence-Increase	ARAUM	15	4.533333	0.7432234
Repair	Effectiveness	Content-Suitability	PMAU	15	4.466667	0.5163978
Repair	Effectiveness	Content-Suitability	ARAUM	15	4.266667	0.9611501
Repair	Effectiveness	Efficiency-Increase	PMAU	15	4.333333	0.9759001
Repair	Effectiveness	Efficiency-Increase	ARAUM	15	4.333333	0.8164966
Repair	Effectiveness	Error-Reduction	PMAU	15	4.266667	1.0327956
Repair	Effectiveness	Error-Reduction	ARAUM	15	4.400000	0.7367884
Repair	Effectiveness	Instruction-Miss	PMAU	15	4.200000	1.0141851
Repair	Effectiveness	Instruction-Miss	ARAUM	15	4.266667	0.8837151
Repair	Effectiveness	Understand-Ease	PMAU	15	4.200000	0.8618916
Repair	Effectiveness	Understand-Ease	ARAUM	15	4.400000	0.6324555
Repair	Satisfaction	Design	PMAU	15	4.000000	0.7559289
Repair	Satisfaction	Design	ARAUM	15	4.133333	0.6399405
Repair	Satisfaction	Feeling	PMAU	15	4.466667	0.6399405
Repair	Satisfaction	Feeling	ARAUM	15	4.333333	0.7237469
Repair	Satisfaction	Overall	PMAU	15	4.266667	0.7988086
Repair	Satisfaction	Overall	ARAUM	15	4.200000	0.7745967
Diagnosis	Ease-To-Learn	End-Ease	PMAU	15	4.200000	0.7745967
Diagnosis	Ease-To-Learn	End-Ease	SMAARRC	15	4.600000	0.6324555
Diagnosis	Ease-To-Learn	Intuitiveness	PMAU	15	4.000000	0.9258201
Diagnosis	Ease-To-Learn	Intuitiveness	SMAARRC	15	4.071429	0.6157279
Diagnosis	Ease-To-Learn	Start-Ease	PMAU	15	3.733333	0.9611501
Diagnosis	Ease-To-Learn	Start-Ease	SMAARRC	15	3.866667	0.8338094

4. Results

4.1. Errors study

Hypothesis

- Errors do not vary with the use of different solutions for each maintenance operation.

Assumptions

- An estimate value for errors per test based on a generic error rate per step.
- Minimum estimate can be set at 1 error per step.
- Medium estimate can be set at 2 errors per step.

Results

- Graph does not show a significant difference on errors per test per solution, but it seems to be a different average per operation. Compared average number of errors per test with conservative estimate, averages are below 15% error rate assuming one error per step in test.
- Compared group means per solution and operation, they range from 0.267 to 0.6. In repair, PMAU average errors are the smallest, while ARAUM is above NOAR. In diagnosis, PMAU average lays equal to NOAR, while SMAARRC is the smallest.
- Tested significance of errors variance for solutions per operation, variances are not significant according to ANOVA results.
- Tested significance of errors variance for operations, variances are significant according to t-test results.
- These results show validity of the following hypotheses:
  - Authoring solutions do not affect errors.
  - Errors are different for each maintenance operation, maybe because operations are different in nature.

4.2. Time study

Hypotheses

- Seconds are reduced with authoring solutions compared to non-AR solutions for each maintenance operation.
- Seconds do not vary significantly between authoring solutions for the same maintenance operation.

Assumptions:

- Differences in maintenance operations do not allow to analyse experiments together. Alternative solutions and steps are not the same, and so the results will differ.

Results:

- Graph does show a considerable difference in completion times per step for each maintenance operation. Besides, it also shows a difference between AR and non-AR solutions, but not between different authoring solutions. A relevant case is D1, it can be seen that in this case the effect of AR solutions is minimum. This case is similar to the findings presented in **[ref]**, where the kind of step had an effect on AR impact.
- Compared groups means per solution and operation, they range from 134 to 231 seconds in repair and from 74 to 134 seconds in diagnosis. These numbers show a difference between repair and diagnosis operations and so, indicate the assumption for separate experiment analyses was senseful. In repair, means show a considerable difference (42%) in completion times between NOAR and AR (PMAU and ARAUM) solutions. In diagnosis, means show a similar difference (43%) in completion times of NOAR and AR (PMAU and SMAARRC), although there is also a smaller difference (~5%) between SMAARRC and PMAU

- Time variance analysed in repair operations show a significant variance for steps and solutions. Thus, indicating validity for the first hypothesis. The interaction between these two effects cannot be considered significant according to anova results. Hence, it can be said that for repair operations, the support AR provides does not depend on the type of step being considered.
- Time variance analysed in diagnosis operations show a significant variance for steps and solutions. Thus indicating validity of the first hypothesis. The interaction between these two effects can be considered significant in this case. This is a similar results to that identified in **[ref]**, where AR support was more effective according to the complexity of the step being conducted.
- Post-hoc comparisons calculated indicate that the second hypothesis can also be considered true. Although anova results show Solution as a significant effect, solution group means differences between authoring solutions are low compared to the difference with non-AR solutions. Moreover, post-hoc comparisons for repair and diagnosis operations show that the mean differences for same-step groups of PMAU and alternative authoring solutions (ARAUM and SMAARRC) are not significantly different. Hence, it can be said that the main effect is driven by the difference between AR and NOAR solutions rather than in-between AR solutions.
- These result indicate validity of the following hypotheses:
  - Seconds are reduced with authoring solutions compared to non-AR solutions for each maintenance operation.
  - Seconds do not vary significantly between authoring solutions for the same maintenance operation.

4.3. Usability study

Hypothesis:

- The proposed authoring solution's usability is similar to that of alternative specific authoring solutions for each maintenance operation.

Assumptions:

Results:

- Counted total number of responses per criterion and tester indicate that the number of survey questions per aspect is 2 with criterion ranging from 3 to 7 aspects per criterion. Hence, it can be said that the survey lenght is quite extensive and so, detailed regarding authoring usability.
- Graph on criterion means per solution and operation show that there are not considerable differences betweem PMAU's and ad-hoc authoring solutions' usability. Most criterions scored above 4 in a Likert Scale out of 5, with higher variabilities in diagnosis scenarios.
- Compared group means per criterion and solution for each operation suggest that authoring solutions achieved similar usability according to testers' opinions. In absolute numbers, group means range from 3.9 to 4.1 in a Likert Scale (1-5) with the exception of PMAU's accuracy in diagnosis, which goes down to 3.4. Percentual differences between PMAU and ad-hoc authoring solutions means range in between -1% and 12%. In repair, ARAUM is considered more usable (5%-8%) regarding all criterions except for Satisfaction. In diagnosis, SMAARRC and PMAU have similar considerations for all criterions but for accuracy, where SMAARRC is considered best by 12%. Overall, these numbers suggest that PMAU's content achieves similar usability than that from ad-hoc authoring solutions because most group means are close to 4 in a Likert scale out of 5. The only exception is PMAU's accuracy in diagnosis operation. A reason for this might be related to an event that occurred recurrently during experimentation and that is connected with HoloLens behaviour: tracking was being lost when testers were asked to get closer to the equipment for taking photographs.
- Total number of responses per aspect (60) provide sufficient data to analyse each criterion's aspect separately. Independent graphs for each criterion showing response averages per aspect for each solution and operation can provide additional insights regarding further improvements on PMAU's usability:
  - Ease-To-Learn results suggest that PMAU's content was slightly more difficult to learn compared to other authoring solutions. ARAUM (tablet-based) had almost no differences between ease-to-use at start and at finish, while SMAARRC's had a slightly smaller difference between start and

finish compared to PMAU. In terms of intuitiveness, only ARAUM's results indicate a better performance.
  - Ease-To-Use results does show interesting differences between authoring solutions in terms of content formats. Tablet-based solutions (ARAUM) showed better responses for text and buttons, while SMAARRC showed the worst results for 3D models. This can be related with the event described above.
  - Accuracy results indicate a worse PMAU's performance in terms of latency. This could be caused due to the real-time PMAU's requirements regarding content generation. For other aspects, responses are quite similar for all three authoring solutions except for occlusion, where SMAARRC recieved a great variability on its responses.
  - Effectiveness results indicate that all three authoring solutions are considered very similar in terms of their abilities to reduce errors, missed instructions and improve efficiency and confidence. One exception is PMAU's variability in ease-to-understand for diagnosis operations. Few testers noted during experiments that ontological naming conventions were sometimes difficult to understand. Thus, it seems important to adapt ontological's wording for improved usability.
  - Satisfaction results were relatively higher for PMAU compared to ad-hoc authoring solutions. A reason for this can be the potential improvements testers figured about PMAU's ontological approach. After experiments, few of them noted the ability of PMAU's approach to **track user's performance through more accurate content monitoring**.
- Overall, usability surveys did not suggest a significant difference between ad-hoc and generic authoring solutions. PMAU was scored relatively lower in accuracy and text understanding, which are areas for further improving its usability. Moreover, PMAU's ability to track user's performance through content monitoring was also perceived as a good solution to further adapt content according to user's expertise. Hence, it can be said that these survey results indicate validity of the following hypothesis: the proposed authoring solution's usability is similar to that of alternative specific authoring solutions for each maintenance operation.

4.4. Discussion

Ideas: Results to be considered in the context of experiments: two different maintenance operations so can be assumed that it would also serve for others. However, real-life experimentation is still required because there are some factors that have not been considered.

5. Conclusions

5.1. Analysis assumptions

5.2. Results conclusions