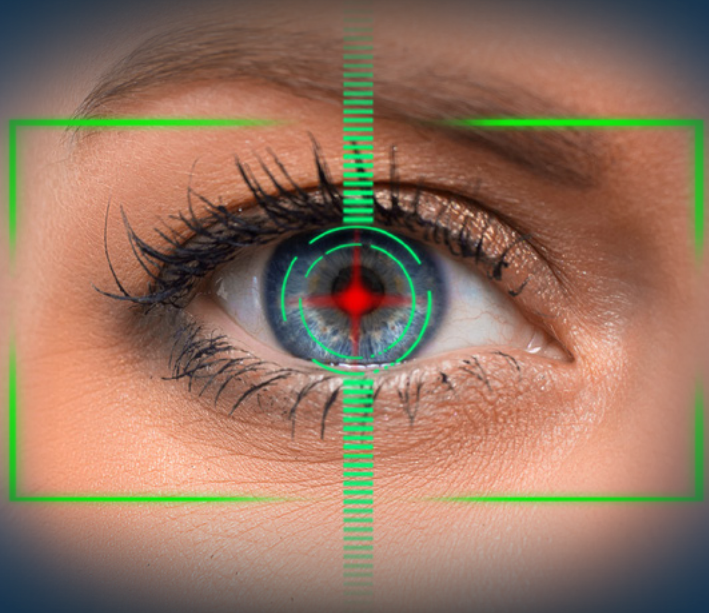
Experimental design and biometric research. Toward innovations

Sylwester Białowąs
Editor



PUEB PRESS
Poznań University of Economics and Business

POLISH NATIONAL AGENCY FOR ACADEMIC EXCHANGE

CEENET SIE

AKADEMICKIE PARTNERSTWA MIĘDZYNARODOWE

Experimental design and biometric research. Toward innovations

Sylwester Białowąs
Editor

PUEB PRESS
Poznań University of Economics and Business

Poznań 2021

NAWA
POLISH NATIONAL AGENCY FOR ACADEMIC EXCHANGE

CENTSIE

AKADEMICKIE PARTNERSTWA MIĘDZYNARODOWE

EDITORIAL BOARD
Barbara Borusiak, Szymon Cyfert, Bazyli Czyżewski,
Aleksandra Gaweł (chairwoman), *Tadeusz Kowalski, Piotr Lis, Krzysztof Malaga,*
Marzena Remlein, Eliza Szybowicz (secretary), *Daria Wieczorek*

REVIEWER
Renáta Benda Prokeinová

COVER DESIGN
Piotr Gołębniak

STATISTICAL EDITOR
Wojciech Roszka

MANAGING EDITOR
Grażyna Jeżewska

PROOFREADER
Katarzyna Smith-Nowak

DTP: eMPi^2
Reginaldo Cammarano

Publication financed by Polish National Agency for Academic Exchange
Project *Central European Network for Sustainable and Innovative Economy*,
no. PPI/APM/2019/1/00047/U/00001

© Copyright by Poznań University of Economics and Business
Poznań 2021

eISBN 978-83-8211-079-1
https://doi.org/10.18559/978-83-8211-079-1

This textbook is available under the Creative Commons 4.0 license—
Attribution-Noncommercial-No Derivative Works

POZNAŃ UNIVERSITY OF ECONOMICS AND BUSINESS PRESS
ul. Powstańców Wielkopolskich 16, 61-895 Poznań, Poland
phone: +48 61 854 31 54, 61 854 31 55
www.wydawnictwo.ue.poznan.pl, e-mail: wydawnictwo@ue.poznan.pl
postal address: al. Niepodległości 10, 61-875 Poznań, Poland

CONTENTS

PREFACE

Over the past years, experiments went in the world of economists from a very rare and not really recognized method into the standard tool for empirical research. This book provides the basic knowledge about using experiments in economics and practical tools for using them. The topic is extended to the more advanced and increasing in popularity area of biometric research.

The book is divided into three parts mirroring experimenting.

The first part provides theoretical background and tips about organizing own research. The chapter is concluded with a guide focused on writing a research report in APA style. This part includes an example of the actual research report.

The next part has two chapters, and both are guided tours allowing to plan and conduct eye-tracking research and electrodermal activity research (EDA).

The last part is devoted to the data analysis. There are three chapters in this part covering the common procedures used in analysis of experiments (all types of experiments, not only biometric). The first part includes tests for one hypothesis: parametric *t*-test and one-way ANOVA and non-parametric siblings: Mann--Whitney U test and Kruskal-Wallis H test. The next part describes test allowing testing more hypotheses: ANOVA without repetition and ANOVA with repetitions. Furthermore, the last chapter deals with dependent samples, which are a popular approach in experiments. This part describes the dependent sample t-test and Wilcoxon test. The effect sizes calculations are included; each test is shown with screenshots from SPSS and some additional screenshots from Excel. This approach allows following the procedure step-by-step. To help easily understand procedures and interpretations, all the examples are basic; they were chosen from areas of sustainability and innovations to match the general idea of the e-books series prepared within the CENETSIE program.

The book contains texts that can be useful in the teaching process. They can be used in graduate programs in economics and business schools, some programs of doctoral schools will benefit from this book as well.

Acknowledgements
The book is the fruit of the cooperation of seven authors from four different universities: the University of Zagreb, Croatia, The D. A. Tsenov Academy of Economics, Bulgaria, and the Poznań University of Economics and Business, Poland.

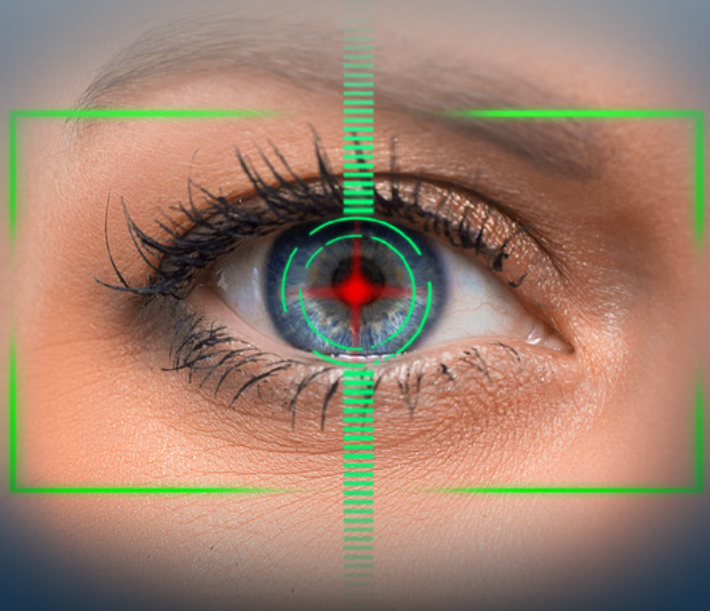
This publication was possible thanks to the international project titled Central European Network for Sustainable and Innovative Economic (CENETSIE), financed by NAWA funds intended to develop international cooperation (PPI/APM/2019/1/00047/U/00001) coordinated by prof. Barbara Borusiak from Poznań University of Economics and Business. The project was implemented in the years 2020-2022.

We, the authors, really hope that this book will help understand the experiments and be helpful when planning and conducting the experiments by the students.

Sylwester Białowąs

PART 1.

PLANNING AN EXPERIMENT



Experimental design and biometric research. Toward innovations, pp. 9-35
https://doi.org/10.18559/978-83-8211-079-1/l1

1.

EXPERIMENTAL DESIGN

Sylwester Białowąs, Atanaska Reshetkova, Adrianna Szyszka
Poznań University of Economics and Business

Atanaska Reshetkova
D.A. Tsenov Academy of Economics

Adrianna Szyszka
Poznań University of Economics and Business

Abstract: Experiment is a research method appropriate to examine causal relationships, also in relation to the current problems of science, including sustainable development.

Conducting experiments can take place in laboratory conditions, but also in natural environments. The main objective of an experiment is always to test what the researcher actually wants and to obtain results that can be generalised to the entire population. In other words, planning experiments requires considering many aspects related to their internal and external validity. The key aspect that needs to be considered in conducting experiments is proper problem defining, as well as the concepts of causality, manipulation or null and alternative hypotheses. It is also worth bearing in mind that in social sciences, when engaging participants in research, caution must be exercised. Depending on whether each participant of the experiment is exposed to all conditions or different people test different ones, the classification of experiments is distinguished into within-subjects and between-subjects design. In this chapter, the most commonly used experimental designs in this division are presented. However, the experimental method offers more complex schemes such as randomised block design or Latin square design. Finally, the obtained findings should be properly presented—in the form of a report following APA standards.

Keywords: experiment, experimental design, randomisation, validity.

1.1. Introduction to the experimental method

1.1.1. The definition of experiment

The experimental model is the best way to check research hypotheses about cause and effect relationships between variables. Experiments allow a researcher to observe and influence a specific phenomenon. Conducting an experiment requires a precise definition of the problem under investigation, as well as analysis of the conditions related to the phenomenon (Stachak, 1997, p. 146). Thus, an experiment should be defined as a conscious, purposeful and planned evocation of specific states and processes. While performing the experiment, the researcher deliberately changes certain factors in the examined situation and, at the same time, controls other factors in order to learn what the effects of the undergoing change are in the course of observation (Sułek, 1979).

An experiment depends on manipulating one or more independent variables and measuring their effect on one or more dependent variables while simultaneously controlling the effect of extraneous variables (Burns, Veeck, & Bush, 2017; Malhotra, Nunan, & Birks, 2017). In other words, experiments help discover causal relationships and ensure that the observed effect concerning the dependent variable is because of the independent variable and not due to other aspects (extraneous variables) (Burns et al., 2017). For example, if checking the effectiveness of fertiliser for plants, in the experiment, two flowers are planted separately and their growth is observed for 60 days, with one of the plants being additionally fertilised.

The three key concepts related to experimental research are:

- independent variable **manipulation** (in the example, one of the flowers will be additionally fertilised);
- **control** of extraneous variables, which may be relevant for the independent variable (in this example, it must be certain that the two flowers grow under similar conditions, i.e. the flowers are planted close together, have the same lighting parameters and get the same amount of water every day);
- variability **measurement** of a dependent variable resulting from the researcher's influence on this variable using independent variables (Brzeziński, 1999, p. 282).

An important distinction is the division of experiments into those performed at a laboratory and in the field. In the case of laboratory experiments, the researcher creates an artificial environment meeting conditions for the tested problem. Field experiments are carried out under real market conditions (Malhotra et al., 2017). Conducting experiments in natural settings creates a more realistic environment but is more expensive and time-consuming in comparison to laboratory experiments (Hair, Bush, & Ortinau, 2003).

1.1.2. Experiments and other methods of scientific research

There are three types of research designs, i.e. exploratory, descriptive and causal. Together, descriptive and causal methods are called conclusive (Malhotra et al., 2017). The main difference between these categories regards their goals—in exploratory studies, the researchers aim to understand the nature of the problem, while in those conclusive—measuring the phenomenon and examining dependencies are of concern (Malhotra et al., 2017). Exploratory research designs are used to clarify and define the problem, obtain additional insights and formulate research objectives. It is especially helpful when little is known about the investigated phenomenon (Burns et al., 2017). Exploratory methods find their application when the problem is difficult to be measured quantitatively (Malhotra et al., 2017).

Descriptive studies provide information about certain aspects of the problem: who, what, where, when and how. This research design allows researchers to describe and measure the phenomena (Burns et al., 2017). This process should be preceded by formulating a hypothesis and defining a problem. This is usually the description of market characteristics or functions that are planned, structured and based on a representative sample (Malhotra et al., 2017).

The last category—causal research designs, enable the measurement of causality in relationships which can be observed when one (or more) variables affect one (or more) variables (Burns et al., 2017). Experiments are the primary method among causal research designs (Malhotra et al., 2017), providing the researcher with the ability to answer the question as to why something occurs and why it may be observed under specific conditions. Examining cause and effect dependencies further allows the researchers to make predictions about various phenomena occurring on the market (Hair et al., 2003). The experiments are considered as research designs that measure the causes and effects of the variables most accurately. Non-experimental studies that are also used for examining cause and effect relationships sometimes do not fulfil all the aspects of causality (Malhotra et al., 2017).

Advantages of the experimental method include:

- enabling the verification of cause-effect relationships relatively easily in comparison to other methods (Moore, McCabe, Alwan, Craig, & Duckworth, 2011);
- helping the researcher to control the experimental conditions and factors that are not significant for the study (Moore et al., 2011);
- easy replication (experiments are repeated more often than other methods), proving the experiment's accuracy (Sułek, 1979);
- making it possible to study the simultaneous influence of more than one factor—separately, the variables may affect a dependent variable in a different way than their interaction (Moore et al., 2011).

Limitations of the experimental method include:

- the researcher potentially not being able to control extraneous variables, particularly in field experiments (Malhotra et al., 2017);
- being time-consuming—especially when the effects of the manipulation are examined in the long run (e.g. effectiveness of an advertising campaign) (Malhotra et al., 2017);
- conducting the experiment on many occasions being relatively expensive (Malhotra et al., 2017);
- potential ethical implications of the conducted experiments (Burns et al., 2017);
- the experiment's results potentially being affected by the artificiality of an experimental situation (Moore et al., 2011).

1.1.3. Research design: type of data

In research design, there are two types of data. The first category refers to the kind that is not obtained by the researcher in the research project or that has been collected for other purposes. This group of data sources refers to surveys and records that are prepared by different companies or organisations. If this data is publicly available, the researcher may use it for research purposes (Burns et al., 2017). It is usually in the form of written documents and is referred to as "desk research". Secondary data is helpful both in defining the objectives of a study and confronting the obtained results. The fact that the data was collected previously by someone else makes it relatively inexpensive and easy to access. On the other hand, since it has been collected for different purposes, the secondary data may cover issues that do not fit perfectly with the research objectives. There is also a risk that this information will be out of date. The secondary data may be obtained from government sources represented by statistical departments. Information is also provided by academic sources and company documents or annual reports. Secondary data come from market research publishers, organisation websites and even private citizens (Bridley, 2013).

Primary data is collected intentionally for a specific purpose. The researcher obtains this kind of data while conducting the research project (Bridley, 2013; Burns et al., 2017). She/he has various possibilities to contact participants and gain the information. This may take place by phone, e-mail, post and/or in person. Among the forms of this kind of data, we can distinguish interviewing and self-completion methods (Bridley, 2013). Experimentation also belongs to this category as a form of gathering primary quantitative data (Malhotra et al., 2017).

1.1.4. Application in economics and management

The experimental method (including field experiments) is commonly used in marketing research, especially in the aspects of communications and advertising (Malhotra et al., 2017). In this area, a popular practice is test marketing which is a type of field experiment. The researchers mostly use test marketing for two purposes—to evaluate the sales potential of a product or service or to test elements of the marketing mix. Furthermore, test markets are used to assess media usage, prices or sales promotions. Although this practice may be expensive for the company, it enables in-advance testing if the product may succeed in the market. There are four main types of test markets—standard, controlled, electronic and stimulated. Standard test markets may provide reliable results because they are conducted in real settings, i.e. using the regular distribution channels of the company. In a controlled model, the experiments are conducted by out-company research firms that test the adjusted distribution channels (Burns et al., 2017). Electronic test markets depend on gathering data from consumers who use an identification card that registers the purchase of goods or services. In simulated test markets, researchers interview selected participants and observe their purchasing behaviours as well as attitude towards the product (Hair et al., 2003).

Experiments are successfully implemented in the area of organisational research. For example, in examining issues related to work efficiency, the profitability of an organisation, the level of task performance or the attitudes and satisfaction of employees (Stachak, 1997). Another field in which companies use experiments is consumer behaviour.

1.2. Key concepts prior to planning an experiment

1.2.1. Causality

Identifying causal relationships is one of the most interesting yet challenging research goals in any scientific field. The concept of causality is rather simple to understand: when one phenomenon is the reason why another manifests, then we have a cause and its effect. How to validate this assumed causal relationship is a question requiring more attention. As in any other research methodology, the experimental method has certain significant criteria that need to be met in order to conclude that a causal relationship really exists.

In any book of statistics and/or research methodology, it is said that correlation does not necessarily mean causation. Two variables can be associated and this

still might not mean that one of them is responsible for the changes in the values (levels) of the other. There may be, for example, a third variable influencing both of them, and thus—making them seem like a cause and its effect, when they are actually both an effect of another cause variable. The pollution of oceans and of air are correlated, but that does not mean that one is a cause of the other. However, the correlation between two variables is the first criterion to pronounce their relationship for causation. In other words, the researcher must be sure that there is an observed association between the cause variable (also called independent variable), and the effect variable (dependent variable).

Another important condition is that the cause must precede the effect. Only this way can we assume that the variation in the independent variable is the cause for variation in the dependent one. For example, suppose we are trying to prove that bad marks at school make admission to college harder. The marks should be received before students send their application to a given college—any other way around contradicts common logic. There are many cases in which it is difficult to decide which came first which makes pointing to the cause and effect variables difficult.

To conclude that there is a causal relationship between any two phenomena, it is necessary to manipulate the impact of the influencing factor, to control all other factors that may influence the test subjects and to compromise the validity of the results. Of course, this is possible only if we conduct an experimental study to conclude causality. The adoption of any other research strategy may lead to the assumption of some correlational degree between variables, which solely, cannot serve as an argument for causality. The choice of experimental design plays a critical role in drawing a conclusion about the causality.

Finally, testing for causality requires the influencing factor to have at least two levels to compare their effects on the response variable.

1.2.2. Independent and dependent variables

As previously stated, in experimental data we can identify two types of variables: dependent and independent. The independent variables are those manipulated by the researcher with the expectation that they will cause some effect on the experimental subjects. If no effect is observed, the reason could be either that there is no causal relationship or that the manipulation of the independent variable was not done properly. For example, the researcher may choose to experiment with only two levels of the variable while there are three or more that can be tested. The dependent variables represent the response and their values are expected to be a result of independent variable manipulation. In social experiments, dependent variables may measure, for example, the actual or intended behaviour

of participants, or different psychological processes. Choosing the right dependent variables that can actually capture the supposed effect is just as important as the right manipulation of the independent variables. Sometimes more than one dependent variable can be used in order to ensure accurate measurement. The manipulation and application of the independent variable on the participants is often called 'treatment'.

The goal of the experiment is to determine the function f that relates the dependent variable y to the independent variables $x_1, x_2, x_3 \ldots, x_k$, i.e., the cause and effect:

$$y = f(x_1, x_2, x_3 \ldots, x_k)$$

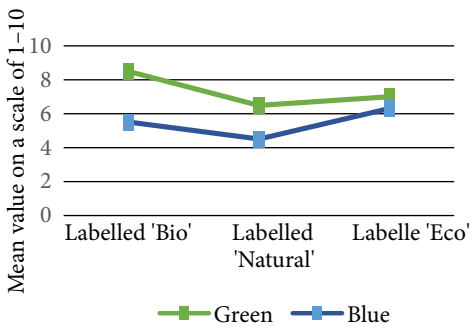
The independent variables are also called factors. There are different types of factors:

- Continuous factors—these are variables that can assume any value in a given interval. Values taken on by continuous factors are therefore represented by continuous numbers.
- Discrete factors—these can assume only a limited number of values. Values taken on by discrete factors can be names or words. Numbers are usually used as codes or labels, and not to denote quantity. For example, the type of labelling used on a product—bio, natural, eco—represents the possible values (levels) that a discrete factor can assume.
- Ordinal factors—these are discrete factors that can be put in a logical order. For example, the ranking of some objects as first, second and third is an ordinal factor. Size defined as small, medium and large is also this type of factor.

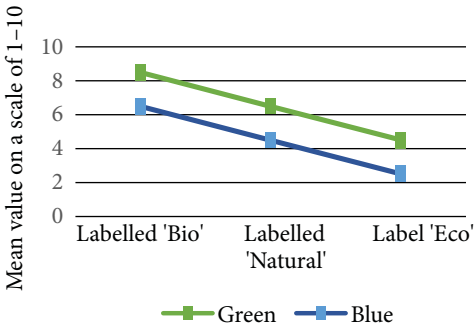
It should be noted that some continuous factors can be transformed into discrete or ordinal ones by creating two or more categories. For example, age is a continuous factor that can be transformed into an ordinal one with three levels: young adults, adults, seniors.

Very often, researchers are interested in testing the effect of more than one independent variable at the same time. In experiments with two or more factors, an interaction between these factors can be observed. If there is an interaction between two independent variables, the effect produced by one of them is different at each level of the other one. For example, let us suppose that we want to compare the effect of different product packaging on consumer perception of its healthiness. We decide to manipulate two independent variables of the packaging—labelling (bio, natural, eco) and colour (green and blue). If the labelling has a different effect on consumers' perceptions when the package is green and when it is blue, then this means that the two factors are in interaction. The mean values regarding the stated perceptions of the product healthiness across different levels of independent variables are presented in Figure 1 (a).

When the package is green and the product is labelled as "bio", it is perceived as healthier than when the labelling is "natural" or "eco". But when the package is blue, the product is perceived as healthiest if it is labelled as "eco". We can conclude that there is an interaction between these two factors. It should be borne in mind that when there is a significant interaction between the independent variables, it is only meaningful to interpret the interaction effects on the dependent variables.



(a) Consumer perception of a product's healthiness



(b) Consumer perception of a product's healthiness

Figure 1. Interaction between independent variables (a). No interaction between independent variables (b)

Source: Own elaboration.

If no interaction exists, then each independent variable effect is interpreted. With regard to the previous example, if there is no interaction between the factors, the results would look like those presented in Figure 1 (b). The product is perceived as healthiest when it is labelled "bio", followed by "natural" and "eco". This does not change when the colour of the packaging changes from green to blue. When the packaging is green, the product is perceived as healthier than when it is blue, regardless of the labelling. These independent effects of each factors are called main effects.

1.2.3. Experimental and control groups

In order to test for a causal relationship between two variables, the one assumed as the cause should be manipulated and the participants exposed to its impact. Then, the response variable should be measured. Following, it needs to be concluded whether a significant effect is observed. There are different ways to organise the experiment depending on its goals and the studied phenomena. Sometimes, all

participants are subjected to the same levels of the independent variables and other times—they are divided into groups, each group subjected to a different level of the independent variable. The specific way of dividing participants into groups and applying treatments is discussed later in this chapter. However, participants in every experiment can be distributed either in an experimental or in a control group. Every experiment has at least one experimental group which is exposed to one or more factor levels. A true experiment includes a control group as well. This is a group of participants not subject to the impact of any factor, thus, providing a baseline for comparison of the dependent variables' mean values.

1.2.4. Selecting research participants

While conducting research, the main purpose is to draw conclusions about the population that is the entire group under the study. However, markets sometimes consist of millions of individuals and the researcher is not able to carry out the experiment on the entire group. Conducting the experiment on the whole population would be time-consuming, expensive and also ineffective. Fortunately, in order to achieve research objectives, the researcher may use a sample (Burns et al., 2017). A sample is a subgroup of the population that represents the entire group and is selected for participation in the experiment. A representation should be suitable since the sample is designed to accurately reflect the characteristics of the group (Burns et al., 2017).

Sample sizes differ across different studies. In order to choose the number of elements to participate in a study, the researcher should take several aspects into account. It is crucial to consider the significance of the study, the number of variables, the sample sizes used in similar experiments and the available resources to conduct the study (Malhotra et al., 2017).

The way of assigning participants to samples induces the division into probability and non-probability sampling. In probability sampling, each unit has a specific chance of being assigned to the sample—the general probability is known. On the other hand, in the non-probability designs, the selection of a sample depends on subjective criteria. In this case, the researcher may correctly assign units to the sample based on his/her expertise, but, at the same time, there is no way to ensure that this selection will be free from bias (Mazzochi, 2008). Therefore, the main difference among both schemes depends on the intervention of the researcher. In probability sampling, the selection of participants is determined by the applied method (Burns et al., 2017). In this chapter, focus will be mostly on probability sampling methods. In this group, the following methods of selection for the sample can be distinguished: simple random, systematic, cluster and

stratified sampling types. In the most basic of these methods—simple random sampling—participants have the same chances of being selected (the selection hinges on luck and probability). A similar procedure is systematic sampling, for which the participants are listed by the researcher who randomly selects only the first assigning number for the first unit in the sample. The rest of the participants is extracted consecutively following the selected starting point. In cluster sampling, the population is divided into complementary groups—each of them should reflect the population. The random selection applies to the clusters. This type of sampling is often an initial step in a more advanced procedure and is useful in relation to electronic databases (e.g. people whose name starts with the letter A, B, C, etc.) or geographical areas (cities, neighbourhoods). The method depending on dividing the population into groups is also stratified sampling. However, in the case of this sampling procedure, the groups are distinguished on the basis of the common characteristic so that the units are similar inside groups and heterogeneous among different strata. It is especially helpful when the distribution of the population is not normal.

1.3. Planning an experiment

1.3.1. Defining the problem and research questions

The key part of the research process is to properly define the problem. It is important because defining the problem influences the research questions, hypotheses and research procedure. If this part is not carried out correctly, there is a risk that we will not get answers to the issues that we want to examine.

In economic practice, a problem is often defined after failure to achieve a goal or after an opportunity has been identified. After that, managers aim to understand the background of the problem, define what decisions should be made and learn about additional sources of information to fully understand it (Burns et al., 2017). This may constitute a basis for scientific exploration. Therefore, the first step in the research process is initial observation and identification of an issue that needs explaining. After finding a lack of knowledge and solutions in some areas, exploratory research should be conducted. It will help the researcher to better structure the problem and clarify the scope of the investigation. The common practice is to investigate previous studies regarding the topic. This step mainly involves conducting a review of literature—before further exploring the problem it is essential to examine scientific publications, books and articles relevant to the issue. On this basis, specific research questions should be identified, and then—the resulting hypotheses (Zikmund, Babin, Carr, & Griffin, 2010).

1.3.2. Null and alternative hypotheses as well as significance

By conducting an experiment, the researcher tries to verify whether the empirical evidence is consistent with the assumed hypothesis. In statistics, it is impossible to show that any statement is undeniably true. However, there are ways to show that some dependencies are not true. In this part of the chapter, the main assumptions of statistical hypothesis testing are presented. Here, the null hypotheses can be distinguished according to whether there is no difference between the groups that are compared in the study. The null hypothesis allows to suggest that no effect is observed but the researcher usually wants to demonstrate that there are dependencies. Thus, there is an alternative hypothesis which predicts that there is a significant difference between the groups under study. This hypotheses is mainly the one that the researcher aims to support (Jackson, 2008). While conducting the experimental procedure, we want to reject the null hypothesis, which would indicate that the findings are consistent with the alternative hypothesis.

The concept of statistical significance is crucial in testing statistical hypotheses. If some difference is statistically significant, this means that it does not happen by chance. In social sciences, the usually chosen level of statistical significance (alpha level) is 5%. It allows the researcher to reject the null hypothesis and indicates that the probability that the tested dependence is due to chance is 5 in 100 (Jackson, 2008).

1.3.3. Data presentation and report structure (APA standards)

After planning, conducting the experiment and data analysis, the results should be properly presented. In this section, the principles of presenting and reporting results will be discussed. Creating the report follows the standards of scientific papers. The research should be fully clear for readers, the conclusions thoroughly explained and presented in a way that allows them to be compared to other studies. This is why the comprehensive standards of reporting are indispensable. The main rule is that all information relevant to the experiment should be included in the report.

The structure of a typical report follows the structure of scientific articles and is presented below:

- Introduction (literature review, main hypotheses)
- Method (design, participants, procedure)
- Results
- Discussion (interpretation, limitations)

The document should also follow the formal structure including: title, abstract, keywords, references and appendix.

Introduction

The first part of a report is the 'Introduction', in which the importance of the problem under study is shown. In this part, the 'Literature review' should also be presented—it is suggested to define the scope of the problem, its theoretical and practical aspects and to indicate what was the subject of research earlier and what remains unexplained. The main hypothesis should be formulated on the basis of the analysed theories. Thus, the introduction involves the description of the study goals as well.

Method

In the next part of the report, the implemented method(s) should be described. The 'Method' section should contain a description of the study participants, including information about their demographic characteristics (e.g. age, nationality, level of education), as well as aspects relevant to the study. Here, the procedures for selecting participants should be presented—the sampling method, time and place of collecting the data, agreements with participants and ethical and safety considerations. In the report, the number of participants taking part in the experiment, the number of participants in experimental and control groups as well as the number of participants that did not complete the experiment should be shown. The 'Method' section involves the inclusion and exclusion criteria for participants. Then, there should be a description of the sample—the number of participants in the study and the planned sample size. If such procedures were used, the methodological part of the report should include information on masking the purpose of the study, training to which collectors were subjected or additional methods. In this section, the research design (whether the between-subjects or within-subjects procedure was applied), the conditions of the study (natural or manipulated) and the assignment to different conditions (if applicable) are described. If the experiment includes manipulations/interventions, it should be precisely described what they consisted of and how they were applied—settings, the duration of exposure and the number of manipulations.

Results

The next section of the report focuses on the 'Results' section of the experiment. An accurate and impartial presentation of the results is the crucial part of the report. All the important results of the study should be presented

with attention to detail and as clearly as possible. In the report, data that are not consistent with the assumed hypotheses should not be omitted—the insignificant dependencies and small effect sizes should be mentioned as well. Raw data and additional materials may be included in the 'Appendix'. When reporting the results, it is recommended to reflect the sequence of the hypotheses presented earlier. When it comes to statistical tests, reporting involves a sufficient set of statistics that are indispensable to understand the outcome. The description should include the value of the test statistic, the degrees of freedom, the *p* value and the magnitude of the effect. The measures of effect size may also be added to this section.

Discussion

The next part of the report regards the 'Discussion' section. The next step, after presenting the results, is to interpret them and draw conclusions from the conducted experiment. It is important to keep this section consistent with the previous one regarding the results. In this section of the report, it should be indicated whether the findings support or do not support the hypotheses. If contradictory or unclear results are obtained, possible causes need to be indicated. Moreover, in the report, the results obtained in relation to the studies of other researchers are presented and the observed differences and similarities are explained. In general, the main implications of the study should be emphasized. In this section, the limitations and strengths of the study are given.

Example

Perception time in forming attitudes towards art

Abstract: In the study, it is examined whether an extremely short exposure to stimuli enables the formulation of aesthetic judgments. In order to determine the time of aesthetic experience formation, an experiment has been conducted in which 12 paintings were displayed during 40 ms. In the previous study, 40 ms was assessed as the minimum exposure duration to process the visual stimuli. The initial judgments were confronted with the judgments formed after longer exposure (10 s). By comparing long- and short-term exposure, it is possible to establish consistency of the observed judgments. The database comprises pairs of works of art by the same artists with a similar composition and auctioned at similar prices, which makes it possible to assess the consistency of judgments with regard to a particular

style. The experiment was conducted on a sample of 30 participants. The main findings allow to indicate that 40 ms is a sufficient time to formulate aesthetic judgment.

Keywords: art perception, formulating aesthetic judgments.

Introduction

When thinking of an aesthetic judgment, it must be considered how well a work of art expresses and influences others with feelings and emotion. The processes underlying the aesthetic experience have been described from both perceptual/cognitive and motivational viewpoints.

In previous research, it has been confirmed that ultra-short exposures (below 1 s) may be sufficient to formulate aesthetic judgements and attitudes. Cupchik and Berlyne (1979) assessed whether people are able to distinguish collative properties with presentation times of 50 ms. They have confirmed that this time allowed the participants to obtain relevant visual information. Locher, Krupinski, Mello-Thoms and Nodine (2007) noted that the time needed to form a significant holistic impression of the painting is about 100 ms.

The most extreme time range was tested in the study by Augustin, Leder, Hutzler and Carbon (2008). They found that 10-ms exposure may be enough to find traces of visual processing effects. In the same study, they confirmed such a significant effect after the presentation of 50 ms.

Main hypotheses

The previous study allows us to state that within the range of 50 to 100 ms, people are able to process visual stimuli and formulate judgment. We aimed to test if the shorter presentation time could be sufficient for similar effects to be observed.

The main hypothesis allows to indicate that a presentation time of 40 ms is sufficient to formulate aesthetic judgments.

Method

In the study the within-subjects, one-group pretest–posttest design was used. There was one independent variable (exposure time) with two levels (40 ms, 10 s). The dependent variable was the aesthetic pleasure measured as a self-reported assessment on the interval scale of 0 (not at all) to 10 (extremely pleasing).

Participants

The recruited participants were students from PUEB. The sample included 52 participants who were not selected randomly. There were 35 women and 17 men between the age of 18 and 31.

Procedure (including technical aspects)

The study took place under manipulated conditions—in the laboratory at the university. We have displayed the stimuli on a 75-inch screen in constant and dimmed light conditions. In the first stage of the experiment, each participant was shown 9 images for 40 ms each. After every stimulus, the participants evaluated their experience by answering the question as to whether the image was pleasing. Each image was preceded by one second of a grey screen with a cross sign in a circle (attention focusing point assuring same visual range for each picture). The second stage was a series of tasks not related to the experiment which was intended to clear the short-term memory of the previously seen stimuli. The third stage was conducted in the same manner as the first one, but the exposure time for each picture was 10 s.

Results

In order to test aesthetic judgment, it was decided to test if there was a difference between scores obtained for short- and long-time exposures. Due to the lack of normal distribution of differences, we decided to apply the Wilcoxon signed-rank test performed 12 times for each picture separately. In 11 cases, it was found that there were no differences between scores (see appendix)—p value was higher than the assumed alpha level ($p > .05$).

In one example, it was found that the evaluation of experience (whether the image was pleasing) changed significantly $Z(52) = -2.54$, $p = .011$. The evaluation that the image was pleasing for the 40-ms exposure time was higher ($Mdn = 5$) compared to the 10-s exposure ($Mdn = 4$). However, the effect size was rather small ($r = .25$).

Discussion

The findings are consistent with the assumed hypotheses that the visual exposure of 40 ms can be sufficient to formulate aesthetic judgment. This means that the obtained results are consistent with the previous findings.

In our study, it was shown that aesthetic judgment may be formed even in a shorter time (40 ms) than expected by other researchers.

Limitations

It may be considered whether the second stage sufficiently separated both experiences of processing visual stimuli. For future research, random sampling could also be considered. In further research, an even shorter time for the initial exposure could be applied—our results do not ensure that 40 ms is the limit of processing visual information enabling the formulation of aesthetic judgment.

References and appendix

Here, tables with detailed results should be presented. Due to the space limitations in this sample report, they have not been included.

References

Augustin, D. M., Leder, H., Hutzler, F., & Carbon, C.-C. (2008). Style follows content: on the microgenesis of art perception. *ActaPsychologica*, *128*(1), 127–138. https://doi.org/10.1016/j.actpsy.2007.11.006

Cupchik, G. C., & Berlyne, D. E. (1979). The perception of collative properties in visual stimuli. *Scandinavian Journal of Psychology*, *20*(2), 93–104. https://doi.org/10.1111/j.1467-9450.1979.tb00688.x

Locher, P., Krupinski, E. A., Mello-Thoms, C., & Nodine, C. F. (2007). Visual interest in pictorial art during an aesthetic experience. *Spatial Vision*, *21*(1–2), 55–77. https://doi.org/10.1163/156856807782753868

1.4. Types of experimental research design

In this chapter, we will focus mainly on true experimental designs. However, we present other types of experiments as well to properly adjust the design to the goal of the study.

1.4.1. Within-subjects and between-subjects experimental designs

Conducting an experiment involves deciding what specific experimental design is to be used. Only after this decision is made, we can determine other important elements of the experimental procedure. Experimental design refers to the way of organising the tested subjects and intervening factors so as to minimise the uncontrolled variation in the effect variable. Of course, the choice of experimental design depends mainly on the purpose of the experiment and the nature of the studied phenomena, but also requires balance between the ability to correctly detect an existing causal effect and the precision with which this effect can be measured (Bellemare, Bissonnette, & Kröger, 2014). However, this is not the only consideration that makes the choice of an experimental design critical for the success of the study: the results can vary considerably depending on the chosen design. In social sciences, there are two major types of experimental designs: within-subjects and between-subjects design.

In the within-subjects experimental design, each participant (or subject) is exposed to all factors levels. In other words, all levels of the independent variable are administered in a consecutive manner on the same group of participants. Because we do not need to assign subjects to different groups, this experimental design requires fewer participants. From a practical point of view, the design is preferred when participants have to fulfil specific conditions to be recruited for the experiment and generally, it is hard to find enough people who are willing to participate.

In the between-subjects experimental design, the participants are divided into separate groups and each group is subjected to only one factor level. To describe the impact of the factor, the differences in the mean values of the effect variable are observed.

The main types of within-subjects and between-subjects designs are presented in Table 1. The experimental groups are indicated by an 'E', a factor is indicated with an 'X' (but not its levels), observations are indicated by an 'O', and control groups are indicated with a 'C'. In the between-subjects design, the indication 'A' represents the sample that is divided into equivalent groups prior to experimental treatment.[1]

[1] The sequence of letters in each row represents the order in which particular actions are taken. For example, in the first design (pretest / posttest), there is only one experimental group –E_1. The dependent variable is measured before the group is exposed to the treatment (O_1). Then, treatment X takes place, and the dependent variable is measured once more (O_2). Parallel rows represent independent, parallel testing of another experimental group.

Table 1. Types of within-subjects and between-subjects designs

Experimental design			Type
	Pretest / Posttest	(1)	$E_1\ O_1\ X\ O_2$
Within-subjects	Pretest / Posttest with control group	(2)	$E_1\ O_1\ X\ O_2$ $C_1\ O_3\ O_4$
	Pretest / Posttest: four-group design	(3)	$E_1\ O_1\ X\ O_2$ $C_1\ O_3\ O_4$ $E_2\ X\ O_5$ $C_2\ O_6$
Between-subjects	Posttest with control group	(4)	$A\ X\ O_1$ $A\ O_2$

Source: (Campbell, 1957; Dimitrov & Rumrill, 2003).

The simplest type of within-subjects design is the pretest / posttest with a control group. The experimental group is observed at least twice: prior to and after testing. The number of observations increases with the number of factor levels that are to be tested. Then the mean values of the dependent variable at each observation are compared for significant differences. Because different external factors that are outside the researcher's control can cause the dependent variable's mean to change between the observations (such as history or maturation), a control group is also included in the experiment. Measuring the dependent variable in the control group in parallel to the experimental group, but without applying any treatment, can help determine whether the observed effect can be attributed to the applied treatment or other external factors are contributing.

The pretest / posttest: four-group design, also known as the Solomon four group design, is somehow an extension of the previous design. It includes two more groups—one experimental and one control group, tested in parallel. Only the post-test is performed in the second experimental group. By omitting the pre-test, the researcher aims to avoid some threats to internal validity—performing the same testing twice can lead to the occurrence of the testing effect. The same logic applies to the second control group, which is only tested once.

The between-subjects design aims to overcome the weaknesses of the described designs by dividing the participants into different groups that are exposed to only one treatment. There is no pretest in this type of experimental design, thus, there is no threat to the internal validity because of the testing effect. Each experimental group is exposed to different conditions and this is why participants' behaviour cannot be influenced by more than one combination of factor levels.

Considerations when choosing an experimental design

True experimental designs are associated with different degrees of external factor control, leading to the occurrence of systematic error, and with a different statisti-

cal power of the tests used. The researcher must decide whether to accept a certain decrease of internal validity at the expense of increasing statistical power, or vice versa; what possible measures can be taken to address the limitations of the preferred experimental design. Comparing the within-subjects and between-subjects experimental designs can be done within two aspects: the potential to provide internal validity and statistical power.

Statistical power of tests

The interference theory poses that the null hypothesis expresses the lack of difference or effect in the observed means (Sawyer & Ball, 1981, p. 275). In most cases, the researcher seeks to reject the null hypothesis in order to accept that there is a significant effect of tested factor levels. Statistical power expresses the probability that the applied statistical test will lead to correct rejection of the null hypothesis[2], therefore, the higher the power of the test, the greater the probability that the conclusion made about the existence of a causal relationship is correct. Statistical power is a function of the test's significance level, the sample size and effect size, thus increasing the sample will increase the statistical power when fixing the other two components (Chase & Chase, 1976, p. 234).

Determining the desired statistical power level before conducting the experiment is important both for the correct definition of the required sample size and for assessing the appropriateness of the study as a whole.[3] However, choosing an adequate power level can be difficult when the estimated size of the effect is unknown. The empirical level of significance should not be used as a measure of the effect size, since both the statistical significance for a particular level of α and the size of the effect are a function of the sample size: even small effects will almost certainly be significant in large samples, while large effects may not be considered significant if the sample is small (Sawyer & Ball, 1981, p. 281). Overcoming the problem of the lack of a preliminary idea of the effect size can be done by conducting a pilot study.

The link between statistical power and the significance level of the test presupposes its relatively lower values when applying conservative tests.[4] Because the between-subjects design is generally more conservative (Charness, Gneezy, & Imas, 2012, p. 2), it is characterised by lower statistical power. Its conservatism stems from the need to apply post hoc contrast tests, some of which are particularly conservative when it comes to comparing more than three pairs of groups (Privitera, 2015,

[2] Statistical power is equal to 1 – β, where β is the probability of failure to reject the null hypothesis when it should be rejected.

[3] For example, a limited budget may force the researcher to change the research design in order to achieve a satisfactory level of statistical power when a relatively small effect is of interest and it is necessary to experiment with a larger sample (which will be more costly).

[4] A statistical test is conservative when the α level is reduced and as a result, the level of β increases.

p. 374), although the researcher may choose to apply a more liberal test. Dividing participants into separate groups, which should be treated with different factor levels, results in a smaller number of subjects in each group, compared to the within-subject design where all participants are in one group. Repeated measurements in the within-subjects design provide more observations from one subject. This allows the researcher to work with a larger sample size and therefore, the statistical power of the tests is higher. Achieving an acceptable level of statistical power in between-subject experiments may require up to four times more subjects than in within-subject design experiments when the number of experimental sessions is small (Bellemare, Bissonnette, & Kröger, 2014, p. 3). This problem becomes more serious with the inclusion of additional factors or factor levels because if the researcher is unable to recruit more participants, this can lead to forming more groups of even smaller size. In such cases, only the recruitment of additional participants could increase the statistical power of the between-subjects experiment.

1.4.2. Different types of experimental designs

The choice of an appropriate experimental design is essential to precisely explore the aim of research. The previous part of this chapter focuses mainly on true experimental designs. However, different categories of experimental designs can be distinguished as well.

In true experimental designs, the researcher employs randomisation in order to assign participants to groups. On the other hand, the main characteristic of pre-experimental design is the lack of randomisation. Here, the most simple design is a one-shot case study which involves only a single measurement after the treatment applied to one group (Malhotra et al., 2017). Sometimes, while conducting experiments, the researcher is unable to control when the procedure is applied and how the participants are assigned to groups. Such a category of experimental designs is called quasi-experimental. Measurements are made at various time intervals—some of them are taken before the treatment and some after. This design can be helpful for practical reasons, particularly when the researcher wants to observe the effects over a long time period (Field & Hole, 2013). There can be a simple time series with one group and a multiple time series with another group that plays the role of a control group.

Occasionally, the researcher seeks to examine the influence of more than one independent variable. The research design may also intend to control the nuisance factors. A more advanced experimental design that enables the researcher to exert control over an extraneous variable is the randomised block design. In this design, the blocking technique is applied, which depends on dividing the participants into the groups on the basis of a similar variable level. This type of experiment is used when the external factor that may influence the performance has been

identified and may be controlled. For instance, a company introduces a new line of environmentally-friendly cosmetic products. The company aims to popularise the product by organising public campaigns. After creating three public campaigns (E, F, G) that differ from each other with regard to the content of information about the product, the company aims to examine their effectiveness. At the same time, the managers assume that the reception of the campaign may vary depending on the usage of similar products before the experiment. Therefore, the information about using a similar product in the past is considered as a blocking variable. Participants are classified into groups based on the assumption that they have often, rarely or never before used other environmentally-friendly cosmetic products in the past. The random assignment occurs at two stages—in selecting participants for the experiment and in assigning them to the types of public campaign (treatment groups). This design is presented in Table 2.

Table 2. Random block design—example

Number of block group	Using environmentally-friendly cosmetic products before	Type of public campaign		
		High amount of information (H)	Medium amount of information (M)	Low amount of information (L)
1	Often	H	M	L
2	Rarely	H	M	L
3	Never	H	M	L

Source: Own elaboration.

A different kind of experimental design in which the blocking technique also finds its application is the Latin-square design. This kind of practice allows the researcher to control two external factors. Again, blocking intends to reduce the additional source of variability. The main rule is that for both variables, the numbers of levels are the same—the scheme of experimental design that is presented in the table has the same number of rows and columns (Montgomery, 2001).

1.5. Conducting experiments

1.5.1. Internal and external validity

The results of a properly planned and conducted experiment should be valid. In general, the concept of validity is about the "extent to which the conclusions drawn from the experiment are true" (Hair et al., 2003). Specifically, this refers to two aspects—the certainty that the effect on the dependent variable is due to the independent variable and that the results may be applied to a larger population of

interest in a real-world context (Burns et al., 2017; Malhotra et al., 2017). Two forms of validity can be distinguished—that is, internal and external validity. Internal validity concerns the accuracy of the experiment. It is the "extent to which the results of the experiment are attributed to the manipulation of the independent variable" (Jackson, 2008; Malhotra et al., 2017). In other words, the researcher needs to be sure that in the study, the causal relationships that may be explained by the experimental treatment (and not by other reasons) are accurately examined (Hair et al., 2003). Thus, if high internal validity of the experiment is to be ensured, there should not be any confounds, i.e. extraneous variables or flaws of the experiment that are not controlled by the researcher (Jackson, 2008). What this means is that controlling variables other than the treatment is an indispensable condition for internal validity (Malhotra et al., 2017).

External validity, on the other hand, indicates whether the observed relationships between independent and dependent variables can be generalised. This means whether the obtained results can be projected onto conditions beyond the experimental situation and if they are true for the entire population to which the study applies (Hair et al., 2003; Malhotra et al., 2017). While planning the experiment, the researcher needs to be cautious and include all aspects that may be relevant in real-world settings in the experiment. The truth is, considering both internal and external validity in designing a study is a challenge for the researcher. In order to ensure a satisfactory level of these two forms of validity, the need to compromise may sometimes occur. The laboratory experiment can be conducted in the case of wanting to control the extraneous variables which increase the internal validity of the study. However, the laboratory conditions differ from the real ones, which may reduce the external validity (Malhotra et al., 2017).

While planning and conducting experiments, the researcher may encounter several threats, both to internal and external validity. Enlisting those threats and errors may help in avoiding some of them.

1.5.2. Experimental errors (threats to validity)

The following threats regarding the internal validity can be enumerated:

History effect: The history effect occurs when the specific event that is outside the experimental situation may violate the results. What should be noted is that this does not mean that the event occurred in the past, before the experiment. Contrarily, it applies to the factors taking place during the time of the study (Jackson, 2008; Malhotra et al., 2017).

Maturation effect: Another aspect that may limit experiment results is maturation. In the case of the experiments that last over a period of time, the participants may naturally grow and develop (e.g. became older, more tired or interested). These

changes are not caused by a specific event but involve the participants and occur with the passage of time. If people change during the period of the study, the effect on the dependent variable may be caused by this instead of the independent variable. How can we deal with the maturation effect? Implementing a control group in the study may help (Jackson, 2008; Malhotra et al., 2017).

Testing effect: Testing effect refers to the process of experimentation. In some studies, pre- and posttest measures are included, and some are conducted on a daily, weekly or monthly basis. These repeated measures may affect the experiment results in two ways—when the prior measure has impact the later one (the main testing effect) or when the prior measure changes the participant's reaction towards the independent variable (the interactive testing effect). Being tested numerous times itself may well influence the dependent variable, decreasing internal validity. Here, such effects as the practice effect—when participants take some tests several times and learn how to perform it better—can also be mentioned. Also, the fatigue effect may occurs when participants become tired of repeating the same procedure and then get lower scores (Jackson, 2008; Malhotra et al., 2017).

Regression: This effect means that in the course of the study, the extreme scores tend to move closer to the mean. It happens when participants are initially chosen on the basis of their extreme scores and then, with several more tests, their scores regress to the average values. It may also distort the internal validity because the observed change may be caused by the changes in scores instead of the independent variable.

Instrumentation: Another threat to internal validity is directly related to the measuring instrument, observation techniques and measurement processes (Hair et al., 2003). The measurement sometimes takes place through observation and the observers may become tired, bored or may lose their focus during the experiment. This also happens because of the lower accuracy of scorers or due to changes in administration procedures (Hair et al., 2003; Jackson, 2008). This threat occurs especially when there is a pretest and posttest study (Malhotra et al., 2017).

Selection bias: The threat to internal validity that depends on inappropriate selection or assignment of the participants to treatment groups is called selection bias (Hair et al., 2003). In this case, the changes in the dependent variable would be impossible to compare because the group may differ initially. This happens when the researchers select participants on the basis of their subjective judgement or when they let the participants assign themselves to groups on their own (Malhotra et al., 2017).

Mortality (attrition): In the case of experiments with experimental and control groups, there is a risk that along with the course of the study, the number of participants will change. As a result, this inequality between both groups may lead to distortion of internal validity. The differences in the group sizes occur due to many reasons—sometimes people just refuse to take part in the experiment. Thus,

it may not be known whether those who participate in the study would react to the treatment in the similar way to those who resigned from participation (Jackson, 2008; Malhotra et al., 2017).

Diffusion of treatment: Another effect that is particularly risky in maintaining internal validity of the experiment is diffusion of treatment. This matches the relationships between the participants of the study who may react differently to the treatment because of the information which they exchange. It may sometimes occur when students are taking part in the experiment. They may know each other and discuss the study during its course or talk about it with the students who have not yet participated in the study. Sharing information about the experiment may influence the reaction to treatment. The researcher should ask the participants not to communicate during the study. The threat may be limited by conducting each part of the experiment in the shortest time possible (Jackson, 2008).

Experimenter effect: The results of the experiment may well be violated by the experimenter. The experimenter is responsible for designing the study and puts a lot of time and effort into this process. Occasionally, the researcher may unintentionally encourage participants to react in a way desired for the purpose of the study. This may be done by body language or mimics. The possible solution could be using the method of blinding, in which the researcher interacting with the participants does not know the details of the treatment (Jackson, 2008).

There are several threats also connected with external validity. The risk of limiting the possibility of result generalisation may be violated by involving mostly student participants in the study. This is a common practice due to accessibility, low cost and time. However, the researcher should be careful with including students mostly in the experiment as they sometimes may not be representative of the target population (Zikmund et al., 2010). The selection may also be crucial in different aspects. For example, if the study demands a lot of time, the participants who are involved are only those who have the motivation and time to take part in it, which also, may not be a good reference to the whole population. Apart of the inaccurate selection of participants who cannot be applied to the whole population, another issue with generalisation refers to the setting of the experiment. This involves conducing laboratory experiments in which there is a risk that the participants' reactions to the treatment may differ compared to those present in the natural environment. Some other aspects that may affect the external validity are timing of the study or exposure to pre-measurements, which can change the participant's reaction to the treatment.

Limitations of validity in between-subject and within-subject designs

It can be pointed out that in experimental studies, the external validity is unattainable for the whole population, but for its subgroups, it is formed on the basis of the

characteristics of the participating subjects. While controlling the adverse effects of external factors ensures internal validity, it could be an obstacle in achieving external validity for laboratory experiments, as conditions may deviate too much from the actual environment of the studied phenomena.

In within-subjects experiments, at least two observations are performed—one before and one after the treatment. Due to the fact that the group should be subjected to all levels of treatment, increasing the number of factors and/or their levels leads to multiple testing. This allows the comparison of each subject to him/herself from before and after the treatment, but it opens space to some unwanted variation caused by re-testing. In the time period between two observations, various side events may occur, or systematic effects may appear (such as fatigue, distraction, anxiety, boredom). Each subsequent test may result in increased experience of the subjects with the experimental procedure. However, some balancing techniques can be used to deal with these effects. Adding an untreated control group to the design allows generalisation of the results to any other equivalent and pretested group (Campbell, 1957, p. 302). It is assumed that the impact of external factors on the experimental and control groups is relatively the same.

Between-subjects experiments do not suffer from the shortcomings described above, as the observed values of the effect variable in at least two experimental groups are compared, and each group is tested only once. It is assumed that any differences in the group means of the dependent variable would be due to the different level of treatment in each group. However, when conducting a between-subjects experiment, reasonable doubts may arise as to whether the observed differences are caused by different personal characteristics of the subjects in the groups. Application of a randomisation technique and ensuring equivalence of the groups is mandatory to eliminate the possibility that differences in participants' characteristics are the reason for the observed difference.

In this section, the aspects that may help the experimenter to control for extraneous variables and enhance validity of experiments are presented. Among the crucial aspects, the role of randomisation should be distinguished. Randomisation is the procedure of assigning participants to groups randomly, which helps ensure that the groups are equal and comparable (Hair et al., 2003). Participants chosen for the study should also be randomly selected for the experimental and control groups. Thanks to this technique, the researcher may assume that the confounding factors will be displayed in the whole group equally. For this to be ensured, the sample should be reasonably large. Sometimes, for the purposes of the study, the researcher may need to select participants with certain characteristics. This procedure—known as matching—is then a step prior to group assignment (Malhotra et al., 2017). The advisable practice, if possible, is also including a control group in the experimental design, which may help deal with the maturation, history, testing or instrumentation effects (Jackson, 2008).

1.5.3. Ethics in experimentation

When conducting experiments involving people, the researchers need to take care of the participants' well-being and morals during the study. In many studies, especially those demanding active participation, informed consent is indispensable. The participants need to know in what activity they will be involved. The participants should be assured that the confidential information obtained during the experiment is not to be disclosed. What should also be worth noting is that the question of confidentiality is crucial, not only for the participants, but also for the companies that are commissioning or sponsoring the research.

The role of researcher is to precisely explain all procedures to the participants. They should be informed about the extent to which they will be involved. Every participant has the right to leave or resign from the experiment at any time.

For some participants, taking part in an experiment may be a totally new situation, thus, they feel stressed. A helpful practice in dealing with this is called debriefing—this consists in informing the participants about the main objectives of the experiment and its hypotheses. This also creates the chance for participants to ask any questions they may have.

References

Bellemare, C., Bissonnette, L., & Kröger, S. (2014). *Statistical power of within and between-subjects designs in economic experiments* (The IZA Discussion Paper series).

Bridley, N. (2013). *Marketing research*. Oxford University Press.

Brzeziński, J. (1999). *Metodologia badań psychologicznych*. Wydawnictwo Naukowe PWN.

Burns, A. C., Veeck, A., & Bush, R. F. (2017). *Marketing research. Global edition*. Pearson Education Limited.

Campbell, D. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*(4), 297-312.

Charness, G., Gneezy, U., & Imas, A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organisation*, *81*, 1-8.

Chase, L., & Chase, R. (1976). A statistical power analysis of applied psychological research. *Journal of Applied Psychology*, *61*(2), 234-237. https://doi.org/10.1037/0021-9010.61.2.234

Dimitrov, D., & Rumrill, P. (2003). Pretest-posttest designs and measurement of change. *Work*, *20*(2), 159-165.

Field, A., & Hole, G. (2013). *How to design and report experiments*. Sage Publications Ltd.

Hair, J. F. J., Bush, R. P., & Ortinau, D. J. (2003). *Marketing research: Within a changing information environment*. McGraw-Hill/Irwin.

Jackson, S. L. (2008). *Research methods and statistics: A critical thinking approach* (3rd ed.). Wadsworth Cengage Learning.

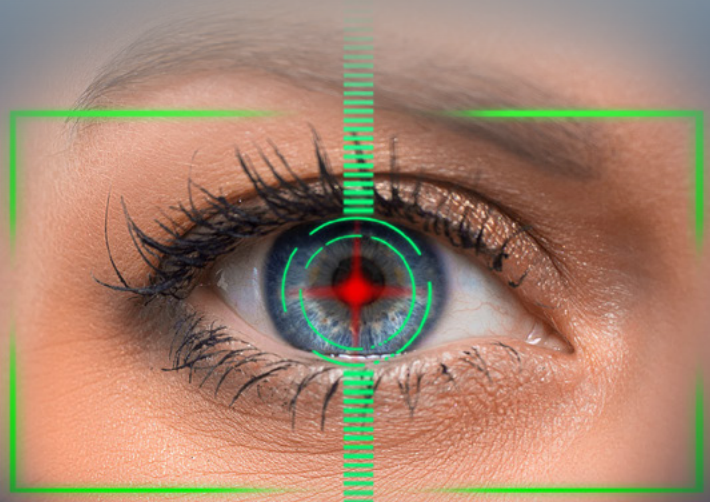
Malhotra, N. K., Nunan, D., & Birks, D. F. (2017). *Marketing research: An applied approach* (5th ed.). Pearson Education Limited.

Mazzochi, M. (2008). *Statistics for marketing and consumer research*. Sage Publications Ltd.

Montgomery, D. C. (2001). *Design and analysis of experiments* (5th ed.). John Wiley & Sons, Inc.
Moore, D. S., McCabe, G. P., Alwan, L. C., Craig, B. A., & Duckworth, W. M. (2011). *The practice of statistics for business and economics* (3rd ed.). W. H. Freeman and Company.
Privitera, G. (2015). *Statistics for the behavioral sciences* (2nd ed.). Sage Publications, Inc.
Sawyer, A., & Ball, D. (1981). Statistical power and effect size in marketing research. *Journal of Marketing Research*, *18*(3), 275-290.
Stachak, S. (1997). *Wstęp do metodologii nauk ekonomicznych*. Książka i Wiedza.
Sułek, A. (1979). *Eksperyment w badaniach społecznych*. Wydawnictwo Naukowe PWN.
Zikmund, W. G., Babin, B. J., Carr, J. C., & Griffin, M. (2010). *Business research methods* (8th ed.). South-Western Cengage Learning.

PART 2.

CONDUCTING BIOMETRIC RESEARCH



Experimental design and biometric research. Toward innovations, pp. 39-60
https://doi.org/10.18559/978-83-8211-079-1/ll1

1.

EYE-TRACKING RESEARCH

Sylwester Białowąs, **Adrianna Szyszka**
Poznań University of Economics and Business

Adrianna Szyszka
Poznań University of Economics and Business

Abstract: Eye movements provide information on subconscious reactions in response to stimuli and are a reflection of attention and focus. With regard to visual activity, four types of eye movements—fixations, saccades, smooth pursuits and blinks—can be distinguished. Fixations—the number and distribution, total fixation time or average fixation duration are among the most common measures. The capabilities of this research method also allow the determination of scanpaths that track gaze on the image as well as heat- and focus maps, which visually represent points of gaze focus. A key concept in eye-tracking that allows for more in-depth analysis is areas of interest (AOI)—measures can then be taken for selected parts of the visual stimulus. On the other hand, the area of gaze outside the scope of analysis is called white space. The software allows for comparisons of static and non-static stimuli and provides a choice of template, dataset, metrics or data format.

In conducting eye-tracking research, proper calibration is crucial, which means that the participant's gaze should be adjusted to the internal model of the eye-tracking software. In addition, attention should be paid to such aspects as time and spatial control. The exposure time for each participant should be identical. The testing space should be well-lit and at a comfortable temperature.

Keywords: areas of interest, calibration, eye-tracking, visual attention.

1.1. Eye-tracking—what it is and how it works

Neuromarketing methods may help obtain a deeper understanding of consumer cognitive processes such as attention and perception. This, of course, enables marketing activities to be addressed in the most efficient manner possible. The subconscious responses to stimuli can be measured using neuromarketing techniques, providing insight into decision-making processes, customer preferences and motivations. One of the most widely used methods of this type is eye-tracking. In recent years, there has been a notable rise in the popularity of using this technique because it offers useful knowledge on how stimuli are processed visually. The need to learn more about the relationship between the brain and the visual system prompted the need to monitor eye movements (Białowąs & Szyszka, 2019; Schall & Bergstrom, 2014).

The eye-tracking system, which tracks the movement of the subject's eyeballs, allows for a thorough examination of the subject's vision direction, as well as the path of attention. As a result, it is possible to isolate the focus areas of the participant's vision, providing an overview of what the subject finds interesting or what has drawn attention. Thanks to this type of information, the researcher may examine how an individual perceives the viewed content (Białowąs & Szyszka, 2019; Duchowski, 2007).

The subject of scientific inquiry has long been how the brain responds to stimuli. Eye-tracking allows to learn more about how the human visual system functions and how the mind works while being exposed to visual content (Schall & Bergstrom, 2014). According to some theories, both attention and eye movements are mediated by the same neural pathways. This means that shifts in attention rely on the stimulation of brain structures involved in eye movement (Hoffman & Subramaniam, 1995).

Eye-tracking is a set of research techniques and methods used to measure, analyse and interpret data on:

- the position and movement of eyeballs (Rojna, 2003);
- where the subject's eyesight falls at a given moment;
- how long the eyesight focuses on a particular point;
- what path it follows (Schall & Bergstrom, 2014);
- pupil size (Bojko, 2013).

Louis E. Javal recorded eye movements using an apparatus mounted on the patient's eye surface in one of the first experiments regarding this field of the 19th century (Wawer & Pakuła, 2012). Eye-tracking has been used in a variety of areas of research, including psychology, medicine, ergonomics and marketing research as well (Białowąs & Szyszka, 2019; Wąsikowska, 2016).

The visual activity consists of four event types:

- fixations—brief pauses in the movement of the eye when the retina stabilizes at a particular point in the field of vision. It means that fixations occur when

the gaze is maintained on a single location. Fixations (visual intake) range in length from 150 to 600 ms (account for 90% of the looking time). They involve the tiniest eye movements such as tremor, drift or microsaccades. As part of visual activity, fixation is a measure of location which reflects the position of the eyes captured in a given time. However, even though the fixation is registered, it does not imply that the subject processed the picture (Duchowski, 2007; Schall & Bergstrom, 2014);

- saccades—rapid eye movements that occur between fixations when the sight shifts from one location to another. Saccades are thought to be an effect of an intention to voluntary change in attention. The duration of saccades is from 10 to 100 ms. Saccades occur when an individual searches different parts of the visual field in a sequential manner. They are not in the main focus of attention research because the visual information is not processed;
- smooth pursuits—movements that allow to track moving objects (Duchowski, 2007);
- blinks.

1.2. What can be examined using eye-tracking

For fixations and saccades, a variety of indicators can be measured.

Fixation measurement indicators include:

- the number and distribution of fixations (which could represent the individual's engagement with the stimuli);
- total fixation time in a specific area and the fixation time per unit area of the visual object (Bylinskii & Borkin, 2015);
- first fixation duration and time to the first fixation (which help determine how long it takes the consumer to recognise a specific element);
- average fixation duration (calculated as the total time divided by the number of fixations);
- revisits—they are observed when the gaze returns to the location where the fixation previously occurred (Garczarek-Bąk & Disterheft, 2018; Tullis & Albert, 2013);
- diversity of fixations—the number of points for which the fixation was observed;
- inter-element fixations—the number of instances when fixations are attributed to various elements (Bylinskii & Borkin, 2015);
- dwell time—the total time of all fixations and saccades (Garczarek-Bąk & Disterheft, 2018).

Saccade measurement indicators include:

- number of saccades;
- saccade duration.

Furthermore, there are measurements of both fixation and saccades as well as some combinations of these eye movements, such as scanpath and heat map.

The order of guiding sight for each space is reflected in the scanpath. Also, it aids the identification of places where the focus is diverted away from the message's vital material during the study. Circles reflect specific points that the subject looks at, with numbers showing the order of perception and lines representing the movement of sight from one point to the next. The following are some of the most commonly used measurements based on the scanpath:

- scanpath length;
- spatial density;
- transition matrix;
- scanpath regularity;
- scanpath direction (Borys & Plechawska-Wójcik, 2017).

The heat map helps participants see which areas earned the most attention and which were missed. Warm colours are used to indicate areas of longer concentration, while cool-toned colours are used to indicate areas of shorter concentration. The items that the subject does not look at, on the other hand, are not coloured. The heat map may also be shown as inverted, displaying the areas of the presented content where the subject focused his or her gaze. Areas of interest, which provide details about the degree to which a given image drew the subject's attention, are another way to view the effects of measuring eye movements (Garczarek-Bąk, 2016; Wąsikowska, 2016).

1.3. How eye-tracking research is prepared

Eye-tracker

An eye-tracker is a device that allows the researcher to get a precise representation and interpretation of how the eyes move. The corneal reflection method is used by most modern eye-trackers to track the location and movements of the eye. This method is based on the use of infrared light sources guided into the eye, accompanied by high-resolution camera reflection. The camera captures an image that can be used. An eye-tracker is a tool that enables obtaining an accurate representation and understanding of eye motion. Most modern eye-trackers follow the position and movements of the eye using the corneal reflection method. The technique is based on the use of light sources (infrared) directed into the eye, followed by a reflection from a camera with high resolution. The image is used to locate the source of light reflection on the cornea, allowing the direction of the subject's sight to be located (Garczarek-Bąk, 2016; Schall & Bergstrom, 2014). In the analysis of visual activity, while using the eye tracker, three main attributes can be distinguished, i.e. the location, duration and movement.

Equipment

1. Smart Recorder—a SmartPhone (based on Samsung Galaxy Note 4) with iViewETG 2.1 Mobile software that is used to create and run experiments.
2. Eye-Tracking Glasses—a mobile eye-tracking device which captures a participant's eye movements. It uses two small cameras on the bottom rim of the glasses and infra-red filtering lenses. The device enables registration of the eye's movements from different distances as well as outdoor (*BeGaze Manual. Version 3.7*, 2017).

   To start the gear, connect part 1 and 2, then power on the recorder.

Figure 1. Eye-tracking equipment

Source: Own elaboration.

Participants

The number of participants included in the study depends on the method of data analysis. For the heat map, the desired number of participants is 39. In qualitative research, it is enough to have at least 6 participants (Pernice & Nielsen, 2009).

The sample should be large enough to maximise the statistical power of analysis and not include too many participants due to study tractability (Duchowski, 2017).

When recruiting people for an eye-tracking study, general information needs to be gathered about the eyes and sight. We should know whether the participant is wearing contact lenses or eyeglasses and/or has any issues with the eyes (e.g. cataracts). The sample should include the participants of the target audience. For

instance, selecting only students for the study may not enable generalisation of the results (Duchowski, 2017).

Before proceeding with the study, participants should be warned that it will take place using technology that tracks eye movements—but the researcher must be careful not to reveal too many details about the procedure, as this may have a negative effect on the obtained results (Pernice & Nielsen, 2009).

1.4. Visual activity testing rules

Conducting the experiment

When carrying out a test using an eye tracker, it is worth remembering a few fundamental principles that will ensure reliability of the obtained results. It must be borne in mind that the exposure time needs to be controlled for each participant—it should be made equal for all the subjects. Furthermore, when controlling the exposure duration is not possible, the solution may be to express the dwell time in percentages instead of absolute values—depending on its duration, other eye movements and other amounts of time spent on watching each element are observed. Time control should only take place when the participant is involved in the study—the time that the respondent spends reporting his/her experience should not be recorded. During the test, the subjects' eye movements should be monitored in real-time, and they should be observed for correct posture. The use of a trigger—the point on which the participants focus their attention at the beginning of the experiment—should be considered. This allows control of the place from which all subjects begin the experiment (Tullis & Albert, 2013).

Space

It is worth remembering that the results obtained through registering eye movements depend on the environment in which the test is performed. When planning an experiment related to tracking the subjects' eyesight, it is worth considering the context in which it is conducted. For example, instead of using an eye tracker in a store space, for reasons of cost and flexibility, researchers decide to use projectors to create a virtual environment. It is worth bearing in mind that the most realistic environment is a real, physical store—the results of the test may be different depending on whether the eye movement is measured in a natural or artificial environment. In the study by Tonkin, Ouzts and Duchowski (2011), it was proved that visual search is faster in a physical environment compared to virtual image—although the perceived difference may not be significant. In turn, if the test is carried out in laboratory conditions, proper lighting should be ensured in the room in which it takes place. It is not recommended to conduct the test in very bright rooms—too much light may affect the device for recording eye movements (Pernice & Nielsen, 2009).

1.5. Before the experiment (proper usage of the equipment, calibration, recording)

Proper usage of the equipment

The glasses should be properly set by adjusting the strip. The position of the glasses should be stable, and the participant is not allowed to change the position of the glasses during the experiment. After turning on the device, the range of the participant's view and the dot showing where the participant is looking at can be seen.

The proper positioning of the glasses is indicated by a green dot on the screen of the recorder (1). If the colour of the dot is not green (yellow or red), the position of the glasses has to be adjusted.

Figure 2. Positioning of the glasses

Source: Own elaboration.

After turning the device on, in the panel on the right, click on the 'NEW EXPERIMENT' button.

After that, you will be asked to name your experiment.

In the next step, a new participant can be added to the experiment. It should be ensured that the participant is added to the experiment. New experiments for new participants of the existing experiment are not to be created. Each new participant should be recorded separately (added as a new participant).

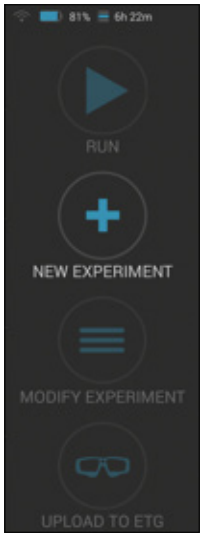


Figure 3. Creating a new experiment on the device

Source: Own elaboration.

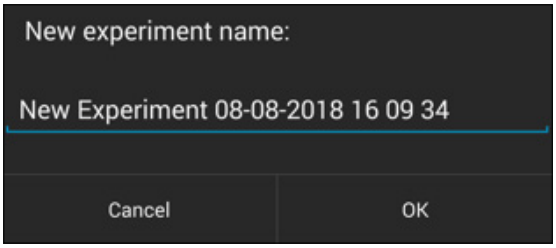


Figure 4. Naming the experiment

Source: Own elaboration.

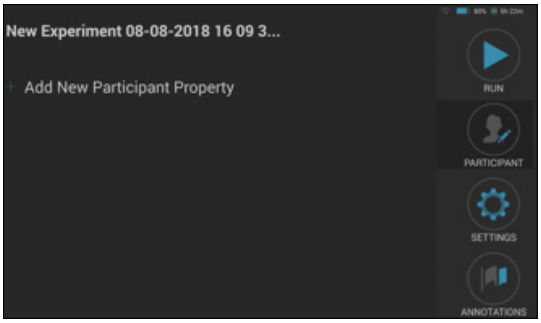


Figure 5. Adding a new participant—part 1

Source: Own elaboration.

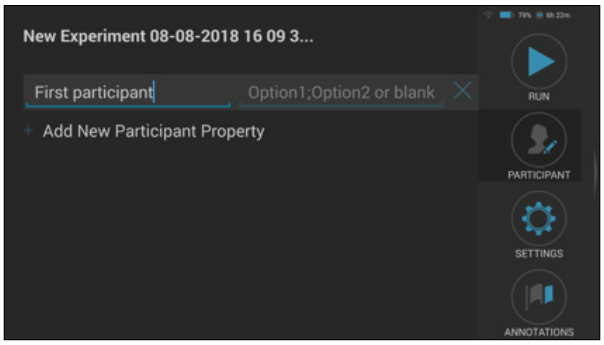


Figure 6. Adding a new participant—part 2

Source: Own elaboration.

Calibration

Calibration enables adjustment of the participant's gaze to the internal model of the eye-tracking software. It is a crucial step in conducting eye-tracking analysis because it helps in precisely tracking the movement of participant's eyes during the experiment (*BeGaze Manual. Version 3.7*, 2017).

In order to calibrate, the CALIBRATE icon on the right panel is to be selected. Before the calibration, the calibration type needs to be chosen (for 1 or 3 points). In this case, calibration will be presented with one point (landmark) that is marked as X.

Calibration should be arranged in the environment similar to real experimental conditions (position of the participant and distance from the object). It must be noted that the calibration should not be conducted with the visible scene of the planned experiment that could bias the experiment results. One or three landmarks (area that we can easily assess the gaze point) are required.

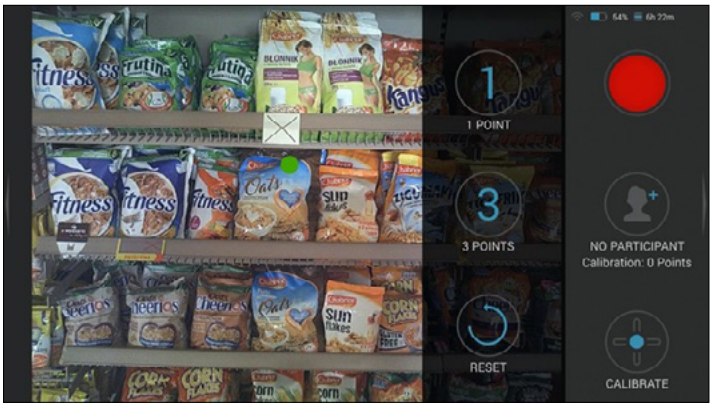


Figure 7. Calibration—step 1

Source: Own elaboration.

On the right panel, the instructions for calibration can be seen. The participant should look at the landmark (X). While the participant confirms gazing at the landmark, even if the dot is not exactly in the place of the landmark, the researcher should tap the screen of the recorder, freezing the image.

Figure 8. Calibration—step 2

Source: Own elaboration.

If the green dot is not exactly on the landmark, the researcher should move the '+' cursor to the landmark, using the touch-screen of the recorder.

Figure 9. Calibration—step 3

Source: Own elaboration.

After positioning the '+' cursor on the landmark, the researcher should click the 'ACCEPT' button.

Figure 10. Calibration—step 4

Source: Own elaboration.

After calibration, it needs to be checked if the position of the dot shows exactly the point at which the participant is looking.

Recording

In order to prevent losing the proper settings, immediately after calibration, the experiment should be conducted. The glasses may not be touched, moved or repositioned.

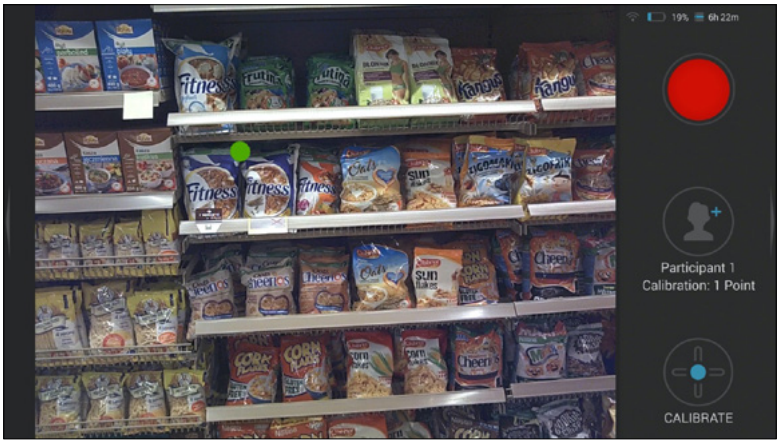


Figure 11. Recording

Source: Own elaboration.

To start the experiment, you should click the 'RED BUTTON' on the right panel (circle-shaped button turns into squared-shaped button, which confirms recording).

It must be remembered to record the whole exposition to the stimulus (full time of the experiment). Recording can be started a few seconds before beginning the experiment.

To end the recording, please click on the same 'RED BUTTON' (squared-shaped button turns into a circle-shaped button, which confirms the end of recording).

Data transfer

After recording chosen participants, the data can be exported to the computer with the BeGaze software. Connect the device to the USB port of the computer. It will appear as a mobile device. You will find the experiment folder in: Card-SMI-A.

Copy the folder of your experiment and save it to the hard drive.

Creating a new experiment in the software

1. Open BeGaze software.
2. Path: File – New experiment from folder – Choose saved folder.

1.6. Data preparation (adding reference image, adjusting gaze points, adding areas of interests, dividing videos, groups)

Preparing experiment analysis

The whole analysis will be conducted on the reference view showing the full visible range of the experiment and allows to set the position of all the fixations. The reference view may be the screenshot from the recorded experiment or a separate image (as in the following example).

1. Adding reference view and selecting fixations for the chosen stimulus. (Path: Change mode – Semantic gaze maping – Confirming it as the default option).
2. Open 'Semantic Gaze Mapping' by clicking on the icon indicated by a red arrow. In order to add a reference view from the folder, click on the icon shown by a green arrow.

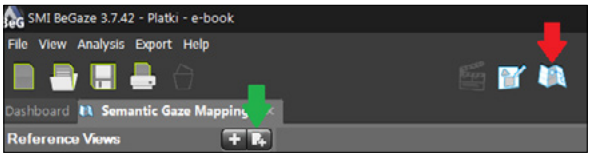


Figure 12. Semantic Gaze Mapping

Source: Own elaboration.

The reference view will be displayed on the left panel. On the right panel, there is a recording of the chosen participant with all the fixations. The allocation of fixations should be conducted for each participant separately. In order to choose the participant, the 'CHANGE STIMULUS' button must be clicked.

The exact length of the experiment can be adjusted by right clicking on the film stripe and setting the starting and ending position of the chosen stimulus.

The first fixation is visible in the right window (displayed as a circle). Please, find and click corresponding position on the reference view. Then, the next fixation will appear in the right window. Please, allocate the fixation to the reference view and repeat the procedure until the final fixation. The allocations are automatically saved and the next participant can be chosen.

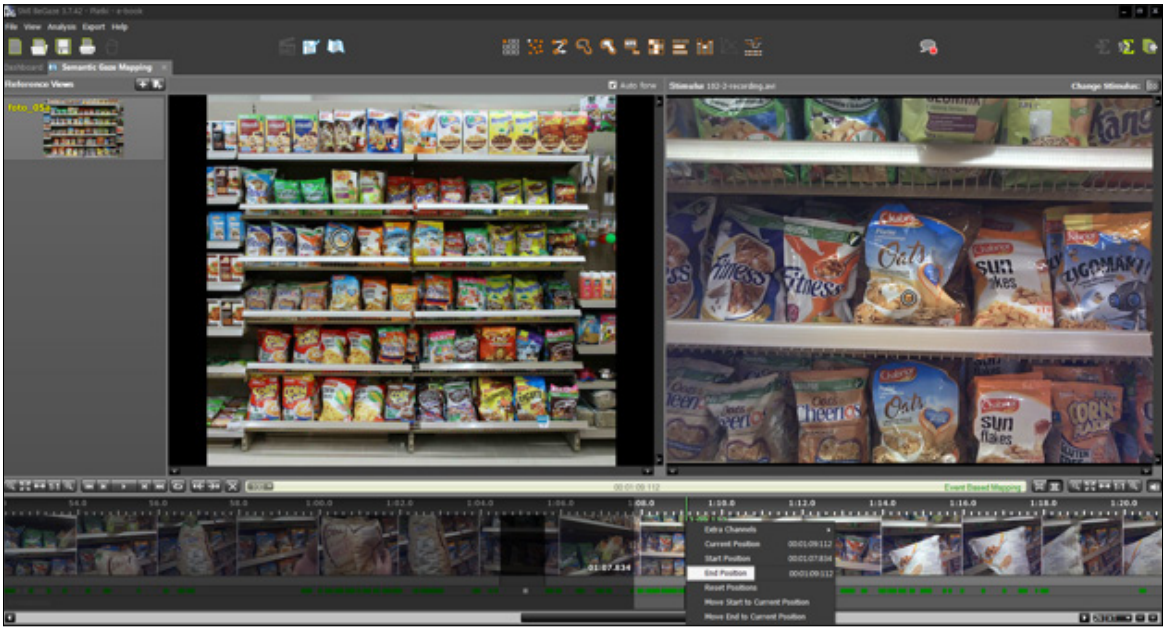


Figure 13. Detecting the fixations

Source: Own elaboration.

Creating AOI

The main analyses are conducted calculating the events within the areas of interest (AOI). Any number of AOI can be set, and results for the chosen areas may be obtained. To set the AOI, the area of our interest can be drawn covering the selected object (e.g. one product, group of products, face, logo, part of logo).

In order to define the AOI for the selected object(s), please click on the 'AOI Editor' indicated by a red arrow.

In the following example, the object is the upper shelf. Please note that on the following screen, AOI has been defined in the rectangle shape. In the AOI toolbar, there are other possible shapes such as those ellipsoidal or polygonal. We can create more AOIs, e.g. the lowest shelf, group of products or even a single product.

If the AOI needs to be deleted, please click on the 'X' in the toolbar (*BeGaze Manual. Version 3.7*, 2017).

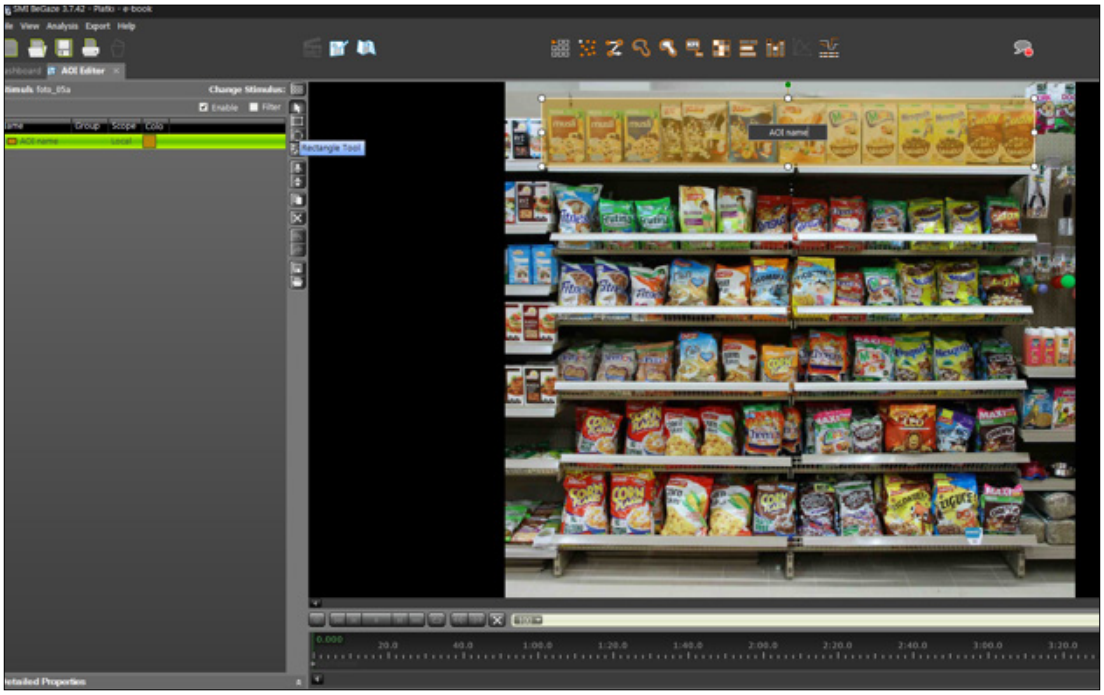


Figure 14. Creating AOI

Source: Own elaboration.

1.7. Analysis using default charts

Bee Swarm shows gaze positions on the reference image (as circles) for selected participant(s) in a given moment. For example, 10 participants have been chosen and their gaze position at the moment of 1:12:771 was checked. Four circles, colour-corresponding to the chosen participants, can be observed. The other six had no gaze positions recorded at that moment.

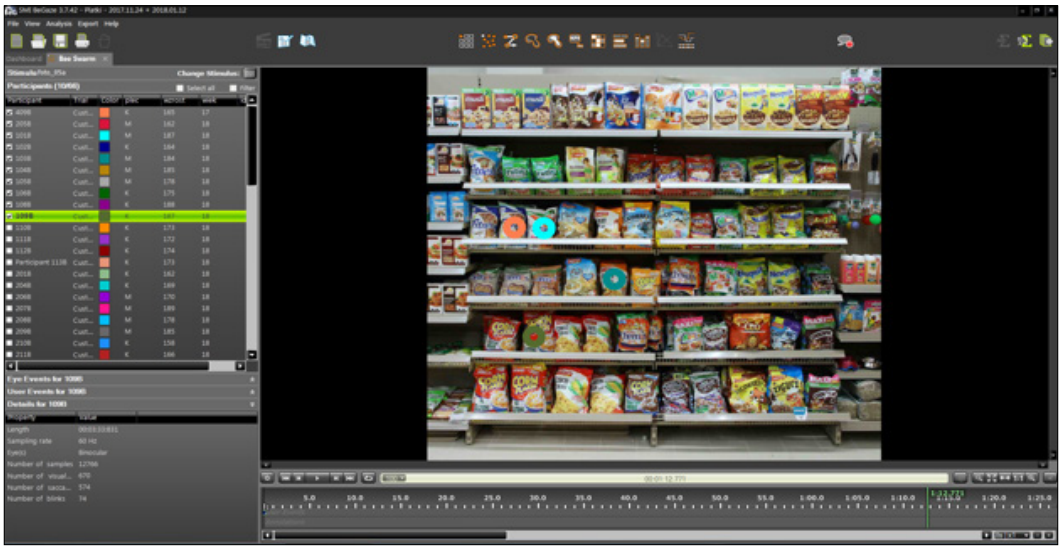


Figure 15. Bee Swarm

Source: Own elaboration.

Scan path shows gaze tracking on the reference image (circles connected by lines) for the selected participant(s). In this example, the scan path for one participant can be seen (matching the colours is the same as in Bee Swarm).

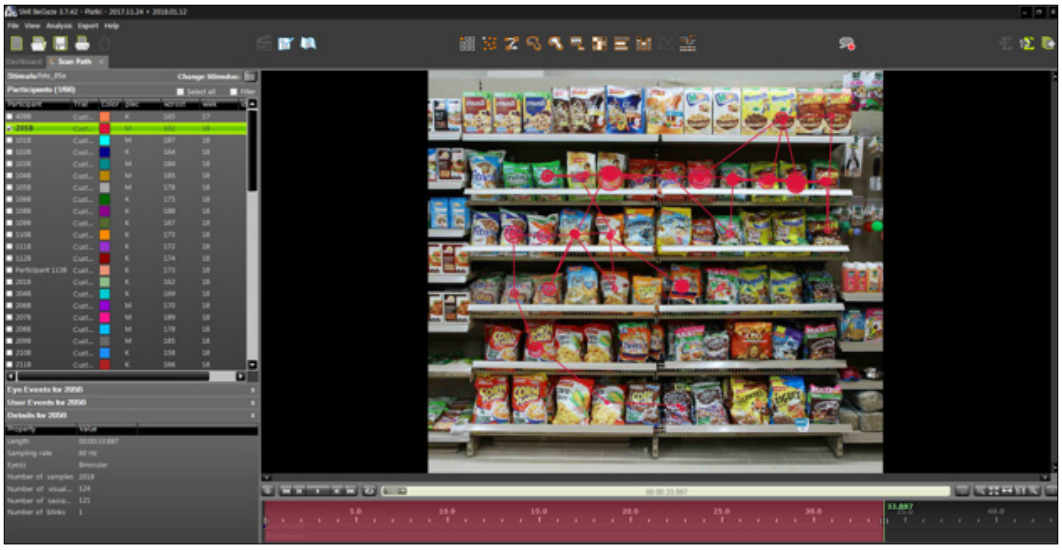


Figure 16. Scan path

Source: Own elaboration.

Heat map allows us to visualise the attention level (number of fixations) of chosen participant(s) by using corresponding colours. From green (lower attention), through yellow (medium attention) to red (higher attention).

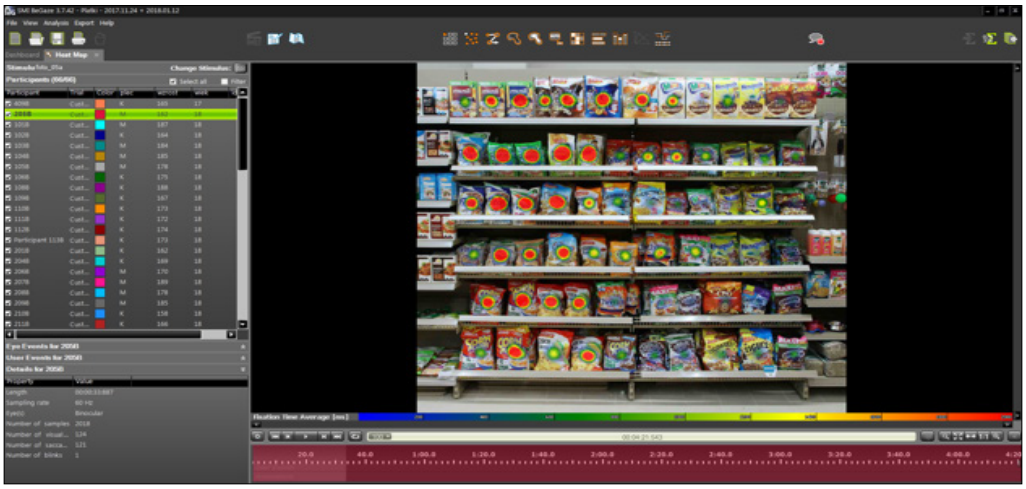


Figure 17. Heat map

Source: Own elaboration.

The focus map is somehow an inversed heat map. It allows to visualise the level of attention by showing the places receiving more fixations.

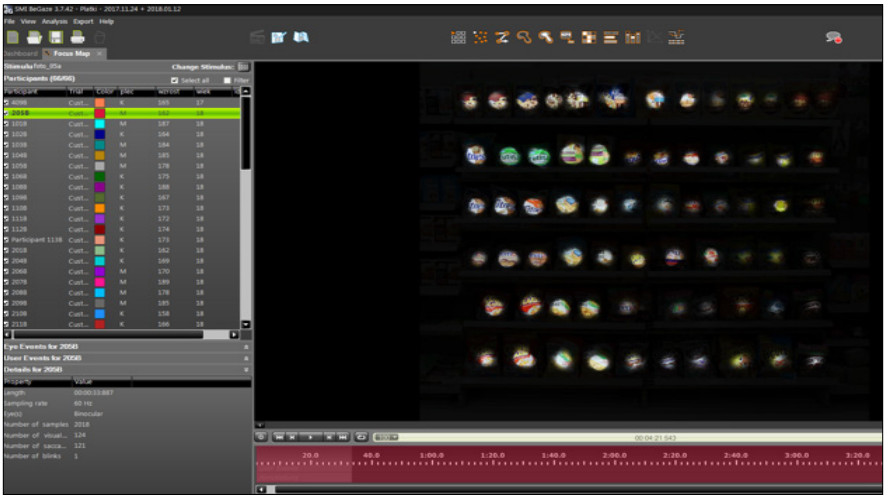


Figure 18. Focus map

Source: Own elaboration.

Key Performance Indicators display the set of the indicators for each AOI of the chosen participant(s). The area that is not covered by the AOI is called White Space, and all the events outside the AOI are summarised in White Space.

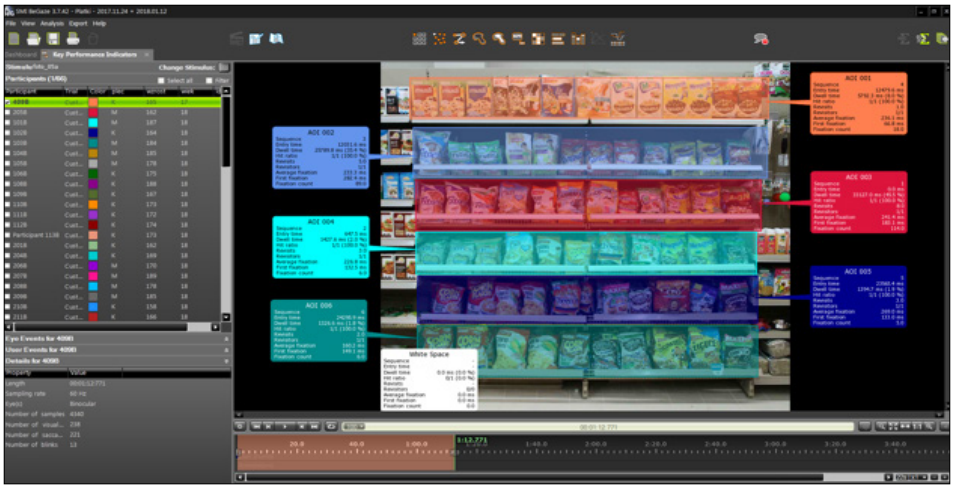


Figure 19. Key Performance Indicators

Source: Own elaboration.

Gridded AOI are default ones proposed by the software as regular squares in the reference image. The Gridded AOI gaze patterns and parameters are visualised by altering the colour of a square based on the level of received attention.

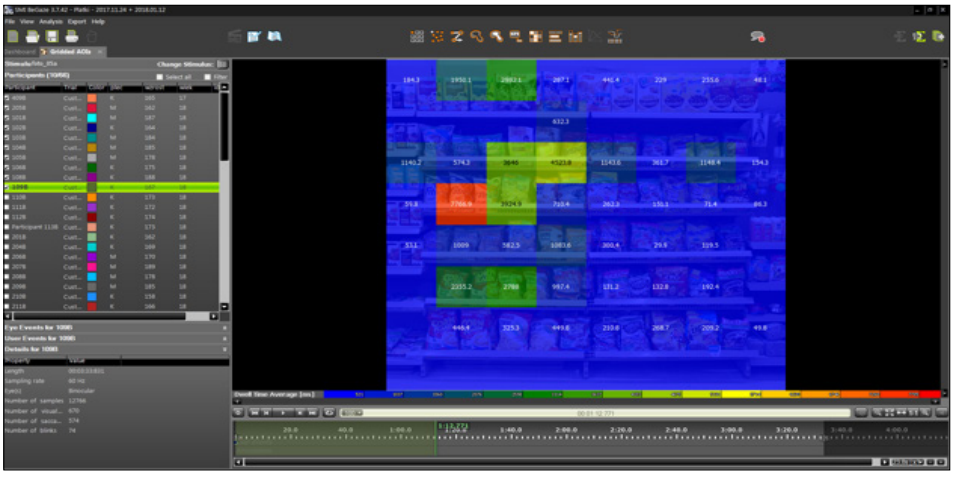


Figure 20. Gridded AOI

Source: Own elaboration.

The AOI Sequence Chart shows the temporal order in which AOI were hit by chosen participant(s). In this example, participant 409B focused gaze for the first 12 000 ms on the red AOI, then onto the blue AOI, and shifted gaze to the orange AOI (for about 5 000 ms), etc.

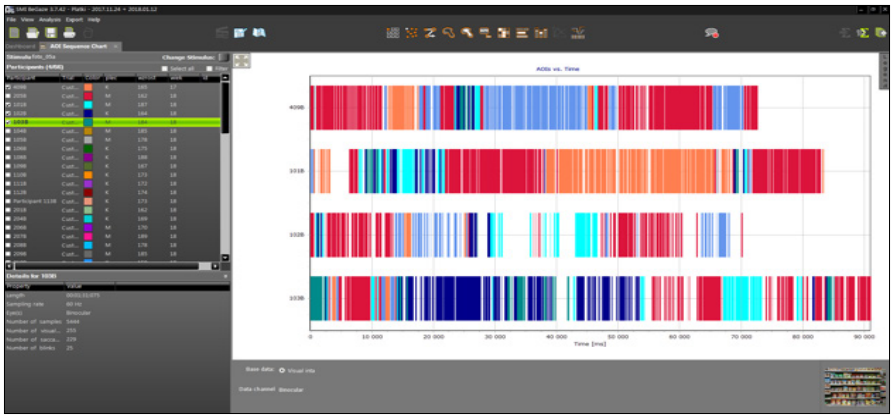


Figure 21. AOI Sequence Chart

Source: Own elaboration.

The Binning Chart shows percentages of AOI dwell time in every time unit.

A value of 100% means that for the whole time of the time bin, for all selected trials, one more AOI was always hit. The time unit of the bins can be adjusted using the 'Bins integration time [ms]' option. In this example, participant 409B in the first second focused gaze for 14% of the time on the blue AOI, for 65% on the red AOI and for 21%, beyond the drawn AOI (White Space).

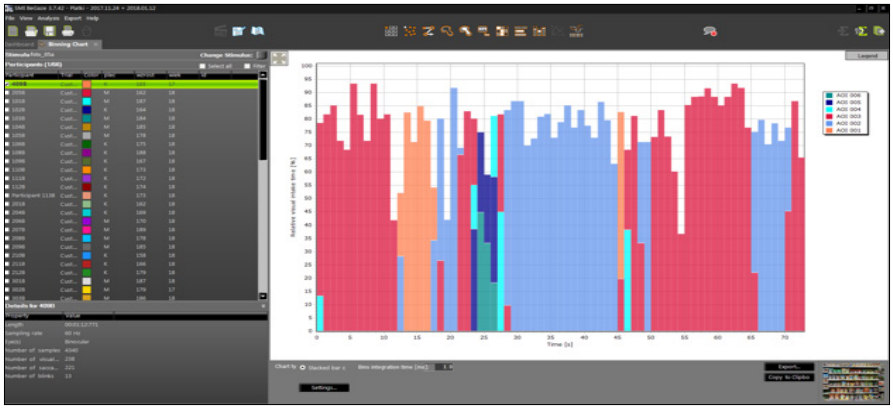


Figure 22. Binning Chart

Source: Own elaboration.

The Line Graph shows a variety of indicators.

In the Line Graph main view, the following gaze data are visualised over the timeline:

- Gaze parameters: the Y-axis on the left displays the gaze position in the stimulus (x- and y-direction) as well as angular velocity and acceleration of the eye.
- Pupil diameter: the Y-axis on the right displays the pupil diameter.
- Time [ms]: the X-axis at the bottom displays fixations, saccades, blinking and user events.

The exact measurements for a chosen time (shown as a red line on time axis) are displayed in the table below. In this example, the diameter increased approximately 28 000 ms, which may indicate higher attention of the participant.

In the presented instance, the diameter of the right pupil with the corresponding events were explored.

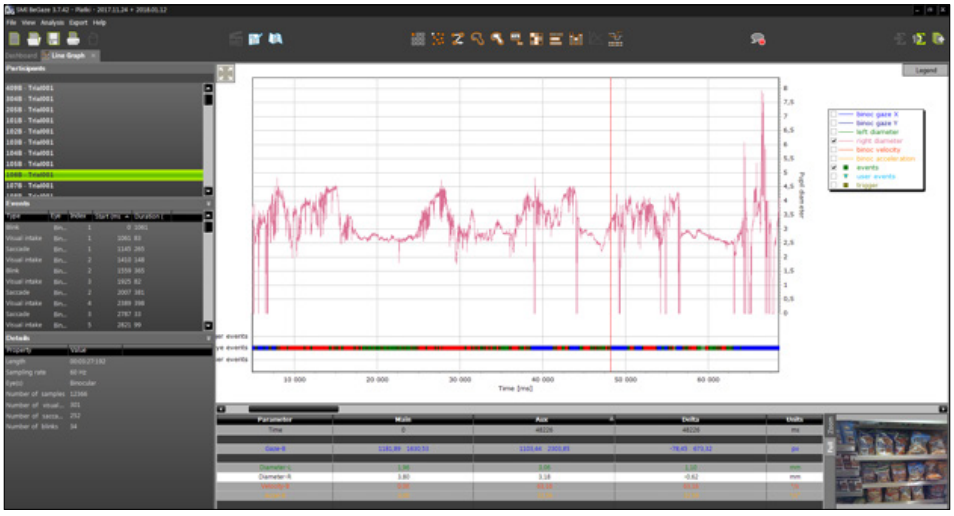


Figure 23. Line Graph

Source: Own elaboration.

1.8. Exporting data for advanced analysis

For more advance analysis, the gathered data regarding the experiment can be downloaded. There is a variety of export settings—template, dataset, metrics or data format can be chosen. It is demonstrated how to export a useful set of data, including the indicators for every fixation in each AOI.

Path: Export – Metrics Export

Select Template – AOI Statistics – Single (fixations only)

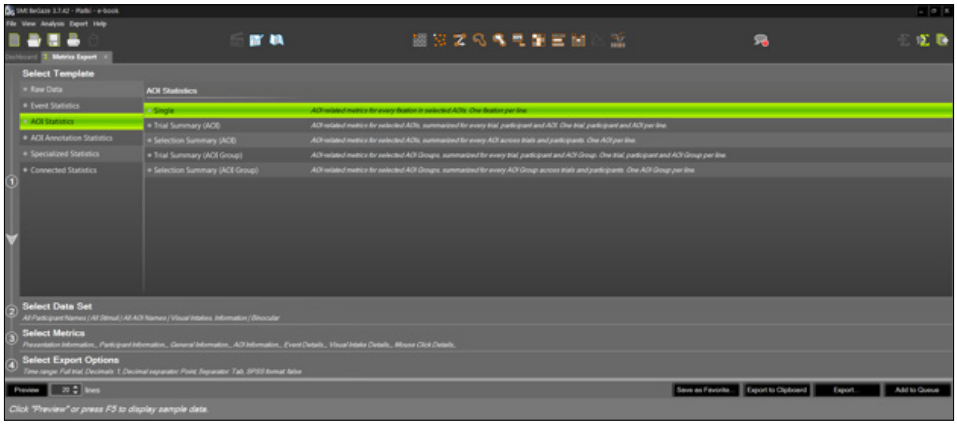
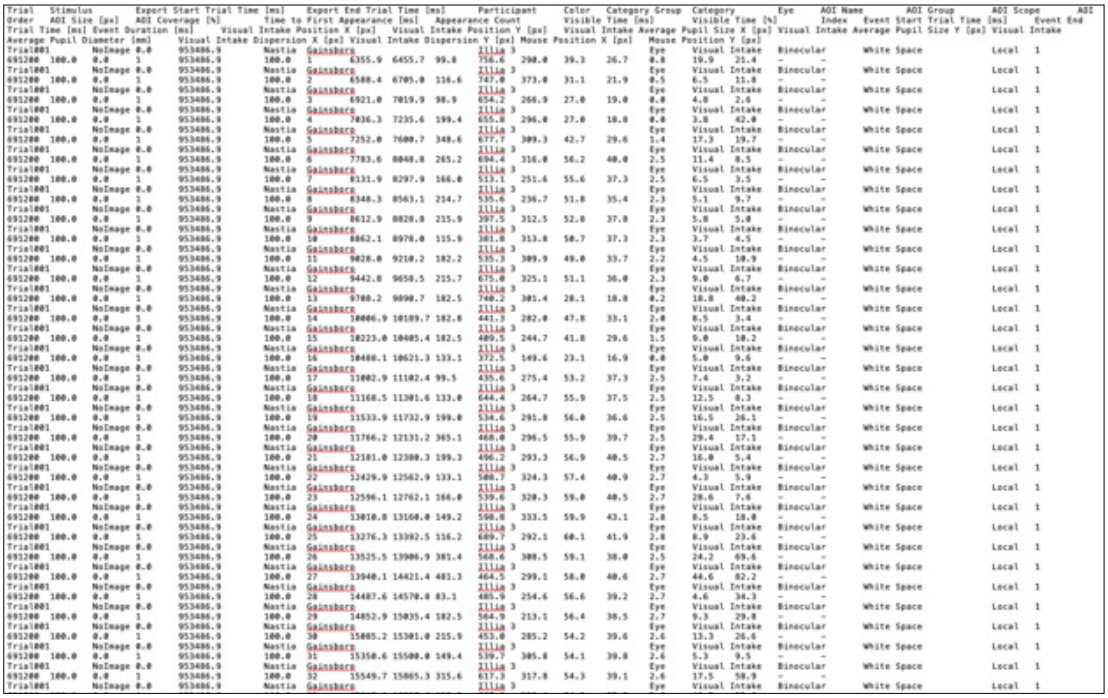


Figure 24. Metrics Export

Source: Own elaboration.

Data are in a .txt format (see Figure 25). Then it can be read in other applications such as Excel or SPSS (Figure 26 and Figure 27).



Trial Order	Stimulus	AOI Size [px]	Export Start Trial Time [ms]	Export End Trial Time [ms]	Participant	Color	Category Group	Category	Eye	AOI Name	AOI Group	AOI Scope	AOI Event End
Trial Time [ms]	Event Duration [ms]	AOI Coverage [%]	Time to First Appearance [ms]	Appearance Count	Visible Time [ms]	Visible Time [%]	Index	Event Start Trial Time [ms]					
Average Pupil Diameter [mm]	Visual Intake Position X [px]	Visual Intake Position Y [px]	Visual Intake Average Pupil Size X [px]	Visual Intake Average Pupil Size Y [px]	Visual Intake Dispersion X [px]	Visual Intake Dispersion Y [px]	Mouse Position X [px]	Mouse Position Y [px]					
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1
691200 100.0	0.0 1	953486.9	100.0 1	6355.9 6455.7 99.8	756.6 290.0	39.3 26.7	0.0	19.9 21.4	- -				
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1
691200 100.0	0.0 1	953486.9	100.0 2	6588.4 6705.0 116.6	747.0 373.0	33.1 21.9	0.5	6.5 11.8	- -				
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1
691200 100.0	0.0 1	953486.9	100.0 3	6921.0 7019.9 98.9	654.2 266.9	27.0 19.0	0.0	4.8 2.6	- -				
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1
691200 100.0	0.0 1	953486.9	100.0 4	7036.3 7235.6 199.4	655.8 296.0	27.0 18.0	0.0	3.8 42.0	- -				
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1
691200 100.0	0.0 1	953486.9	100.0 5	7252.0 7600.7 348.6	677.7 309.3	42.7 29.6	1.5	1.4 17.3	19.7 - -				
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1
691200 100.0	0.0 1	953486.9	100.0 6	7783.6 8048.8 265.2	694.4 316.0	56.2 40.0	2.5	11.4 8.5	- -				
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1
691200 100.0	0.0 1	953486.9	100.0 7	8131.9 8297.9 166.0	513.1 251.6	55.6 37.3	2.5	6.5 3.5	- -				
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1
691200 100.0	0.0 1	953486.9	100.0 8	8348.3 8563.1 214.7	535.6 236.7	51.8 35.4	2.3	5.1 9.7	- -				
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1
691200 100.0	0.0 1	953486.9	100.0 9	8612.9 8828.8 215.9	397.5 312.5	52.8 37.8	2.3	5.8 5.0	- -				
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1
691200 100.0	0.0 1	953486.9	100.0 10	8862.1 8978.0 115.9	381.8 313.8	50.7 37.3	2.3	3.7 4.5	- -				
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1
691200 100.0	0.0 1	953486.9	100.0 11	9028.0 9210.2 182.2	535.3 309.9	49.0 33.7	2.2	4.5 10.9	- -				
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1
691200 100.0	0.0 1	953486.9	100.0 12	9442.8 9650.5 215.7	675.0 325.1	51.1 36.0	2.3	5.0 6.7	- -				
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1
691200 100.0	0.0 1	953486.9	100.0 13	9780.2 9890.7 182.5	740.2 301.4	28.1 18.8	0.2	18.8 40.2	- -				
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1
691200 100.0	0.0 1	953486.9	100.0 14	10046.9 10189.7 182.8	441.3 282.0	47.8 33.1	2.0	8.5 3.4	- -				
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1
691200 100.0	0.0 1	953486.9	100.0 15	10223.0 10405.4 182.5	489.5 244.7	41.8 29.6	1.5	9.0 10.2	- -				
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1
691200 100.0	0.0 1	953486.9	100.0 16	10488.1 10621.3 133.1	372.5 149.6	23.1 16.9	0.0	5.0 9.6	- -				
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1
691200 100.0	0.0 1	953486.9	100.0 17	11002.9 11102.4 99.5	435.6 275.4	53.2 37.3	2.5	7.4 3.2	- -				
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1
691200 100.0	0.0 1	953486.9	100.0 18	11168.5 11301.6 133.0	644.4 264.7	55.9 37.5	2.5	12.5 8.3	- -				
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1
691200 100.0	0.0 1	953486.9	100.0 19	11533.9 11732.9 199.0	536.6 291.8	56.0 36.6	2.5	16.5 26.1	- -				
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1
691200 100.0	0.0 1	953486.9	100.0 20	11766.2 12131.2 365.1	468.0 296.5	55.9 39.7	2.5	29.4 17.1	- -				
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1
691200 100.0	0.0 1	953486.9	100.0 21	12181.0 12380.3 199.3	496.2 293.3	56.9 40.5	2.7	16.0 5.4	- -				
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1
691200 100.0	0.0 1	953486.9	100.0 22	12429.9 12562.9 133.1	580.2 324.3	57.4 40.9	2.7	4.3 5.8	- -				
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1
691200 100.0	0.0 1	953486.9	100.0 23	12596.1 12762.1 166.0	539.6 320.3	59.0 40.5	2.7	28.6 7.6	- -				
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1
691200 100.0	0.0 1	953486.9	100.0 24	13010.0 13160.0 149.2	590.8 333.5	59.9 43.1	2.8	8.5 18.0	- -				
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1
691200 100.0	0.0 1	953486.9	100.0 25	13276.3 13392.5 116.2	689.7 292.1	60.1 41.9	2.8	8.9 23.6	- -				
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1
691200 100.0	0.0 1	953486.9	100.0 26	13525.5 13906.9 381.4	568.6 388.5	59.1 38.0	2.5	24.2 69.6	- -				
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1
691200 100.0	0.0 1	953486.9	100.0 27	13940.1 14421.4 481.3	464.5 299.1	58.0 40.6	2.7	44.6 82.2	- -				
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1
691200 100.0	0.0 1	953486.9	100.0 28	14487.6 14570.8 83.1	485.9 254.6	56.6 39.2	2.7	4.6 34.3	- -				
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1
691200 100.0	0.0 1	953486.9	100.0 29	14852.9 15035.4 182.5	564.9 213.1	56.4 38.5	2.7	9.3 29.8	- -				
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1
691200 100.0	0.0 1	953486.9	100.0 30	15085.2 15301.0 215.9	453.0 285.2	54.2 39.6	2.6	13.3 26.6	- -				
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1
691200 100.0	0.0 1	953486.9	100.0 31	15350.6 15500.0 149.4	539.7 305.8	54.1 39.8	2.6	5.3 9.5	- -				
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1
691200 100.0	0.0 1	953486.9	100.0 32	15549.7 15865.3 315.6	617.3 317.8	54.3 39.1	2.6	17.5 58.9	- -				
Trial001	NoImage 0.0	953486.9	Nastia Gainsbarre		Illia 3				Eye	Visual Intake	Binocular	White Space	Local 1

Figure 25. Data in .txt format

Source: Own elaboration.

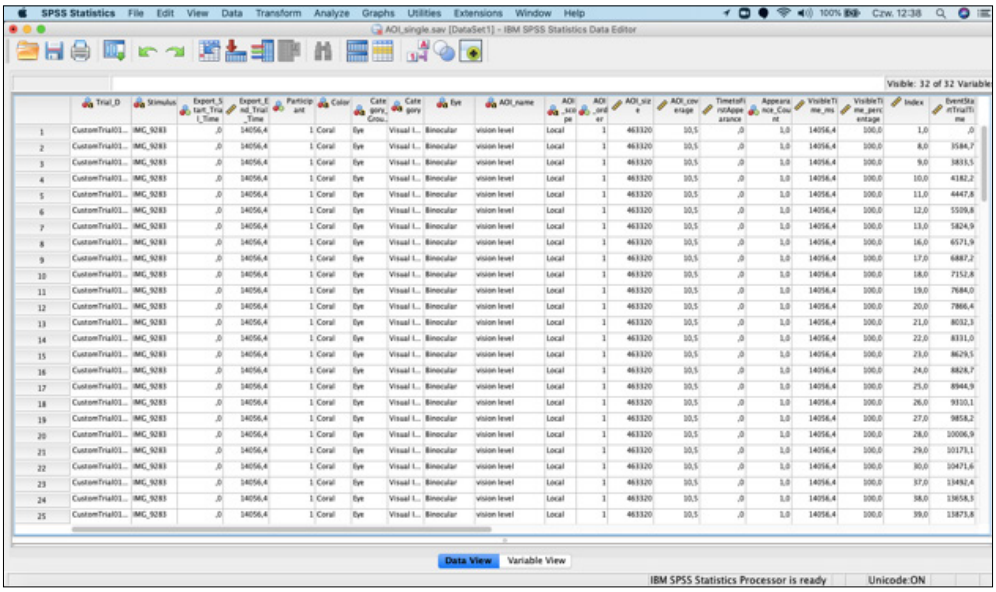


Figure 26. Data in SPSS—part 1

Source: Own elaboration.

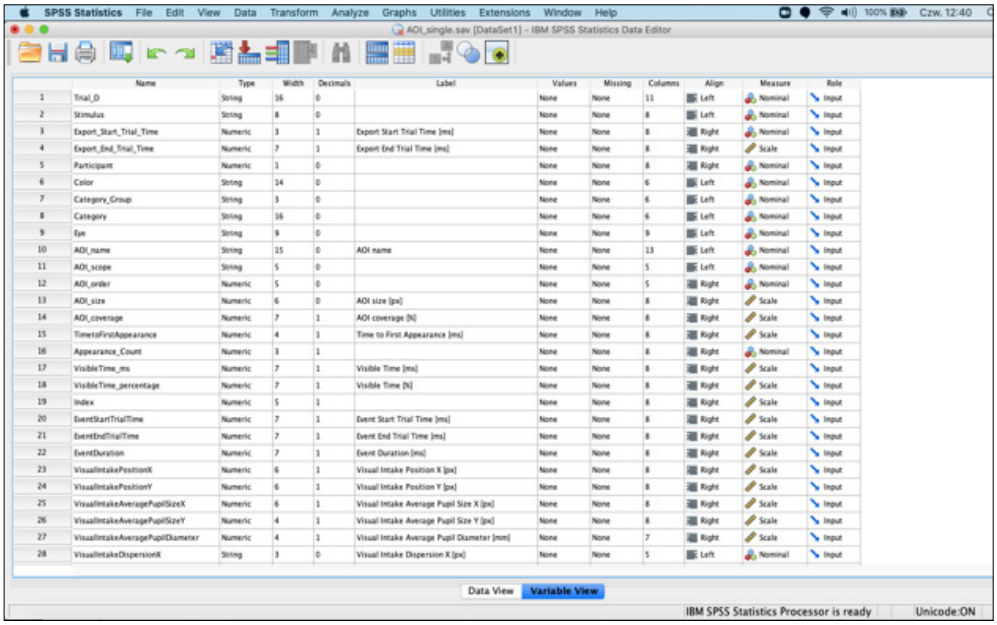


Figure 27. Data in SPSS—part 2

Source: Own elaboration.

References

BeGaze Manual. Version 3.7. (2017). SensoMotoric Instruments.

Białowąs, S., & Szyszka, A. (2019). Eye-tracking in marketing research. *Managing Economic Innovations—Methods and Instruments*, *January*, 91–104. https://doi.org/10.12657/9788379862771-6

Bojko, A. (2013). *Eye tracking the user experience: A practical guide to research*. Rosenfeld Media.

Borys, M., & Plechawska-Wójcik, M. (2017). Eye-tracking metrics in perception and visual attention research. *European Journal of Medical Technologies*, *3*(16), 11–23. Retrieved from http://www.medical-technologies.eu/upload/2_eye-tracking_metrics_in_perception_-_borys.pdf

Bylinskii, Z., & Borkin, M. A. (2015). *Eye fixation metrics for large scale analysis of information visualizations*. ETVIS Workshop on Eye Tracking and Visualization.

Duchowski, A. T. (2007). *Eye tracking methodology*. Springer-Verlag. https://doi.org/10.1007/978-3-319-57883-5

Garczarek-Bąk, U. (2016). Użyteczność badań eye trackignowych w pomiarze utajonych determinant zachowań zakupowych nabywców. *Ekonometria*, *3*(53), 55–71. https://doi.org/10.15611/ekt.2016.3.05

Garczarek-Bąk, U., & Disterheft, A. (2018). Analiza obszarów zainteresowania w oparciu o badania eyetrackingowe na przykładzie produktów marek własnych i producenckich. *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu*, *525*, 211–226. https://doi.org/10.15611/pn.2018.525.18

Hoffman, J. E., & Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Perception & Psychophysics*, *57*(6), 787–795. https://doi.org/10.3758/BF03206794

Pernice, K., & Nielsen, J. (2009). *How to conduct eyetracking studies*. Retrieved from https://media.nngroup.com/media/reports/free/How_to_Conduct_Eyetracking_Studies.pdf

Rojna, W. (2003). *Eye tracking. Metodologia i jej zastosowanie w badaniach percepcji reklamy i zachowań konsumentów*. IV Ogólnopolski Kongres Badaczy Rynku i Opinii.

Schall, A., & Bergstrom, J. R. (2014). *Eye tracking in user experience design*. Oxford: Elsevier Ltd. https://doi.org/https://doi.org/10.1016/C2012-0-06867-6

Tonkin, C., Ouzts, A. D., & Duchowski, A. T. (2011). *Eye tracking within the packaging design workflow: Interaction with physical and virtual shelves*. Conference on Novel Gaze-Controlled Applications, January, 1–8. https://doi.org/10.1145/1983302.1983305

Tullis, T., & Albert, B. (2013). *Measuring the user experience collecting, analyzing, and presenting usability metrics*. Elsevier Inc.

Wąsikowska, B. (2016). Eye tracking in marketing research. *Zeszyty Naukowe Uniwersytetu Szczecińskiego. Studia Informatica*, *36*(863), 177–192. https://doi.org/10.18276/si.2015.36-13

Wawer, R., & Pakuła, M. (2012). Zastosowanie techniki eyetrackingowej do analizy postrzegania historycznej przestrzeni wystawienniczej przez osoby starsze i młodzież: teoretyczne i metodologiczne podstawy badań. *Zeszyty Naukowe Uniwersytetu Szczecińskiego. Ekonomiczne Problemy Usług*, *88*, 698–707. Retrieved from http://www.wzieu.pl/zn/703/ZN_703.pdf

Experimental design and biometric research. Toward innovations, pp. 61-88
https://doi.org/10.18559/978-83-8211-079-1/II2

2.

RESEARCH ON ELECTRODERMAL ACTIVITY

Bartłomiej Pierański Jakub Berčík
Poznań University of Economics and Business

Jakub Berčík
Slovak University of Agriculture in Nitra

Abstract: In this chapter, a method of physiological measurements—that is detection of electrodermal activity based on the sectonic activity of eccrine sweat glands—is discussed. It is believed that the excretion of sweat, which is regulated by the nervous system acting independently of human will, is an indicator of a person's emotional arousal as a result of specific stimuli. Hence, the electrodermal reaction can be used in diagnosing emotional arousal caused by, e.g. specific products, advertisements or elements of the in-store space.

Electrical activity of the skin is caused by two types of stimuli: sustained and one-off. Sustained stimuli have a continuous effect on the body over a relatively long period of time. On the other hand, one-off stimuli have a relatively strong and very short-lasting effect. This type is defined as novel, unexpected, significant or aversive. Electrodermal activity is measured on the skin surface (Strelau, 2006).

Generally speaking, the measurement of electrodermal activity is one of the biometric measurements. Biometrics is a universal term that represents measurements of the body's physiological responses—not directly of the brain—to external stimuli that are felt through the senses (Pradeep, 2010; Berčík & Rybanská, 2017). The electrodermal method allows to measure either electrical resistance or its inverse, i.e. the electrical conductivity of the skin. These measurements are carried out while a small current flows through the skin from an external source.

Electrodermal activity measurement is performed with the use of special electrodes, electrode gels and recording devices. The available equipment for the analysis of electordermal activity is characterised by relatively low cost (compared to other devices for physiological measurements) of purchase and operation. Moreover, the electrodermal activity measurement is non-invasive and carries no risk to the health or life of the test subjects.

Keywords: customer research, electrodermal activity, emotional arousal, measurement devices, physiological measurement.

2.1. What is electrodermal activity and why consumers can be better understood by measuring it?

The electrical activity of the skin is known as electrodermal activity (EDA). Its essence lies in the electrical phenomena generated by the skin. The source of the skin's electrodermal activity are the so-called eccrine sweat glands (Cacioppo, Tassinary, & Berntson, 2007), which are mainly responsible for the secretion of sweat (*secretory theory*). Sweating is a source of conducting electricity.

The functioning of the eccrine sweat glands is regulated by the sympathetic nervous system that is part of the autonomic system (Zhai, Barreto, Chin, & Li, 2005). The centres of this system are located in the spinal cord and work on the basis of the reflex principle. This means that the increase or decrease in skin sweating (and thus the skin's electrical conductivity or resistance) is automatic and subconscious, and therefore, it cannot be influenced by a human (Cacioppo et al., 2007).

The sweat glands are distributed across nearly the entire body surface area (covering practically the entire surface of the body), totalling an amount of approximately 2 million. However, they are particularly concentrated on the forehead, cheeks, hands and feet. Glands play a thermoregulatory role in the human body. Under normal conditions, the glands excrete about 500 ml of sweat from the body per day (Sosnowski & Zimmer, 1993).

Nonetheless, thermoregulation is not the only cause of the sweat glands' work. Increased sweat excretion is also observed during the following situations (Boucsein, 2012):

1) eating meals;
2) physical impact on the skin;
3) taking medication;
4) spontaneous reaction of the glands;
5) and emotional arousal.

The sweat glands are stimulated by eating mainly acidic, very salty and spicy meals. Sweat, the source of which is food, appears primarily on the forehead, the top of the cheeks and the tip of the nose. The amount of sweat produced in this way can be considerable and thus, clearly visible. A local increase in sweating is also observed in areas of physical impact on the skin, for example, due to acupuncture, high temperature or radiation. The work of the sweat glands can also be stimulated pharmacologically (Boucsein, 2012).

However, what is really important within the context of costumer research, is the activity of sweat glands caused by the body's response to a specific type of stimuli coming from the environment. It is believed that the excretion of sweat, which is regulated by the nervous system acting independently of human will, is an indicator of the emotional arousal of a person as a result of specific stimuli. It

ranges from a low-level during sleep to a high level during strong activation. It is assumed that all emotions (both positive and negative) cause increased sweating. **Hence, the electrodermal reaction can be used in diagnosing emotional arousal of consumers caused by, e.g. specific products, advertisements or elements of the in-store environment** (Galvanic Skin Response, 2016). That is why this type of sweating is called 'emotional sweating'. In other words, EDA can be used to **examine implicit emotional responses** that may occur without **conscious awareness** or are **beyond cognitive intent** (i.e., threat, anticipation, salience, novelty).

Emotional sweating, in particular, involves the glands that are located on the hands and feet. Therefore, their function is not strictly thermoregulatory. This function is revealed only at high temperatures, exceeding 30 degrees Celsius. However, in normal room temperatures, and assuming undisturbed thermoregulatory functions of the body, a high correlation was found between the work of the sympathetic nervous system and the electrodermal reactions of the skin (Wallin, 1981). It is for this reason that that their functioning is believed to be more susceptible to psychological stimuli than tasks related to thermoregulation of the body (Edelberg, 1972). An important feature of electrodermal reactions in the context of emotional arousal is their high sensitivity to stimuli of very low intensity (Boucsein, 2012).

Nevertheless, on the basis of increased sweating alone, it cannot be inferred whether the emotions evoked by a given stimuli are negative or positive (Cacioppo et al., 2007). Therefore, it cannot be precisely indicated that these changes are the result of, for example, anger, joy or fear.

2.2. Types of electrodermal activity

The electrical activity of the skin is caused by two types of stimuli (see Figure 1): sustained and one-off. Sustained stimuli have a continuous effect on the body over a relatively long period of time. On the other hand, one-off stimuli have a relatively strong but very short-lasting effect. This type is defined as novel, unexpected, significant or aversive (Cacioppo et al., 2007).

Sustained stimuli affect so-called tonic skin activity (Cacioppo et al., 2007, p. 171). Tonic activity represents relatively constant or slow changes in the electrodermal activity of the skin. This activity is known as SCL (skin conductance level). The so-called tonic activity also includes non-specific reactions (fluctuations), i.e. reactions occurring without the influence of a stimulus (Strelau, 2006). They are known as NS SCR (non-specific skin conductance response) (Cacioppo et al., 2007). It has been found that, for example, SCL is characterised by a gradual decrease in its level when a particular person is resting (i.e., not affected by stimuli and relatively still) (Strelau, 2006). On the other hand, it has been found that the increase

in tonic electrodermal activity (SCL level and NS SCR frequency) is influenced by the performance of a specific task. Supporting studies were carried out by Lacey (Lacey, Kagan, Lacey, & Moss, 1963). As part of them, participants were asked to perform a variety of tasks, ranging from listening to irregular, loud sounds to solving arithmetic tasks. In the case of preparation by the participants for each task, an increase in the level of SCL in relation to the level at rest was noted. On the other hand, the performance of tasks led to a further increase in tonic level (Cacioppo et al., 2007, pp. 171–172).

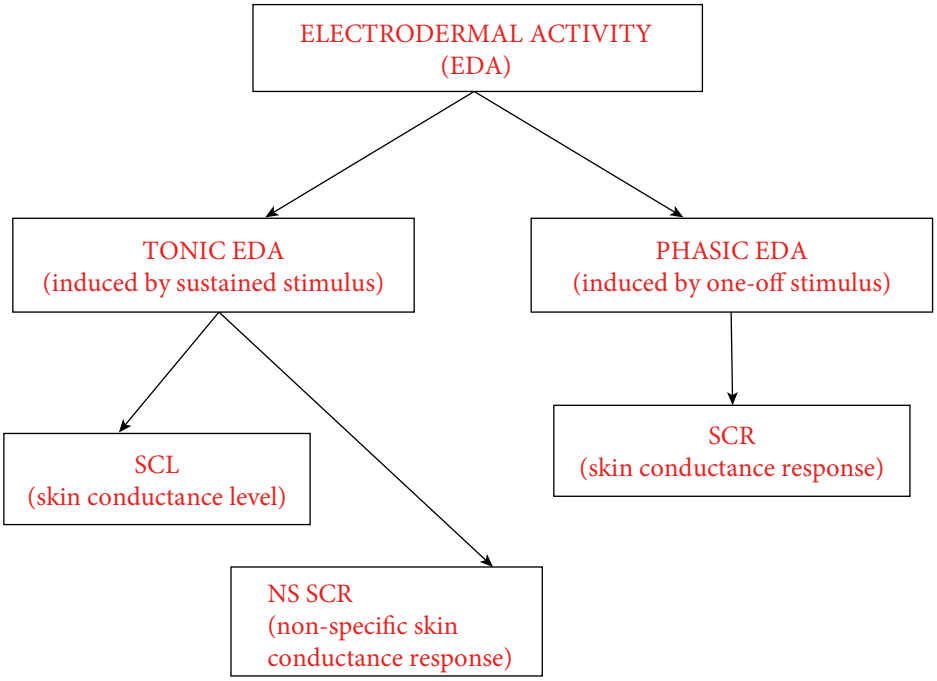


Figure. 1. Types of electrodermal activity

Source: Author's own elaboration based on (Benedek & Kaernbrach, 2010).

The electrical activity of the skin can also be phased. Phase activity reflects a sudden response to a short-term but intense stimuli through a jump in sweating. This activity is known as SCR (skin conductance response) (Cacioppo et al., 2007).

Phase electrodermal activity is a manifestation of both the orientation reflex and its habituation. The orientation response is defined as the body's response to a stimulus. The function of this reaction is to facilitate the reception of the stimulus while stopping other activities that may hinder the perception of the stimulus. The orientation reflex manifests itself simultaneously in several areas. The first of them are changes in external behaviour, manifested by stopping the tasks

being performed, directing the whole body, and thus the sense organs towards the stimulus. The second area of change in physiological systems occur, for example, by increased skin sweating, slowing of heart rate, dilation of blood vessels in the head, pupil dilation, etc. The orientation response is reduced until complete disappearance (habituation) in the event of repeated specific stimulus.

Concluding, it should be stated that tonic activity reflecting slow electrodermal changes is caused by stimuli of sustained nature. On the other hand, phase activity is a relatively violent electrodermal reaction to a relatively short-term intense stimulus.

2.3. Measurement of electrodermal activity

What is actually measured?

Generally speaking, measurement of electrodermal activity is considered a biometric measurement. Biometrics is a universal term representing measurements of the body's physiological responses—not of the brain directly—to external stimuli that are felt through the senses (Pradeep, 2010; Berčík & Rybanská, 2017).

It is worth noting that emotional arousal can be detected in two ways. First, via electrical conductivity of the skin. The higher it is, the greater the sweat secretion and the greater emotional arousal. The second way is based on inverse electrical conductivity, i.e. electrical resistance. In this case, the lower the resistance, the greater the sweat secretion and emotional arousal (Białowąs & Szyszka, 2019).

Furthermore:

1) an increase in skin electrical conductivity means—a decrease in electrical resistance of the skin = emotional arousal;
2) a decrease in skin electrical conductivity means—an, increase in electrical resistance = lack of emotional arousal.

Both conductance and resistance are expressed in specific units. Thus, conductance is expressed in simens, or more often in microsiemens (mS), while resistance—in ohms (more often in kilohms) (Strelau, 2006).

Below, a description of skin conductance measurement is given.

It should be taken into account that the phase and tonic electrodermal reaction (described in the previous chapter)—manifested by an increase in skin conductivity—are measured differently.

Measuring the phase electordermal reaction to a stimulus (and thus, emotional arousal), two groups of parameters are considered: parameters characterising the size (amplitude) and duration of the reaction. The first group includes the amplitude of the reaction. That is the level to which the level of EDA has increased as a result of the influence of a specific stimulus. The second group of time parameters include (Sosnowski & Zimmer, 1993, Cacioppo et al., 2007):

1) latency time, that is, the period from the stimulus onset to the electrodermal response; there is usually a 1-4-second so called 'latency window', hence, any SCR that begins between 1 and 4 s, following stimulus onset, is considered to be elicited by that stimulus;
2) rise time, that is the temporal interval between SCR initiation and SCR peak;
3) recovery time, that is the temporal interval between SCR peak and point of complete SCR amplitude recovery.

As the recovery time is relatively extended over time and thus, there is a risk that the electordermal activity of the skin may not return to the baseline level before the onset of the next stimulus, therefore, in the research, a substitute parameter is widely used—half recovery time, temporal interval between SCR peak and point of 50% SCR amplitude recovery (Sosnowski & Zimmer, 1993, Cacioppo et al., 2007).

The analysed parameters are graphically presented in Figure 2.

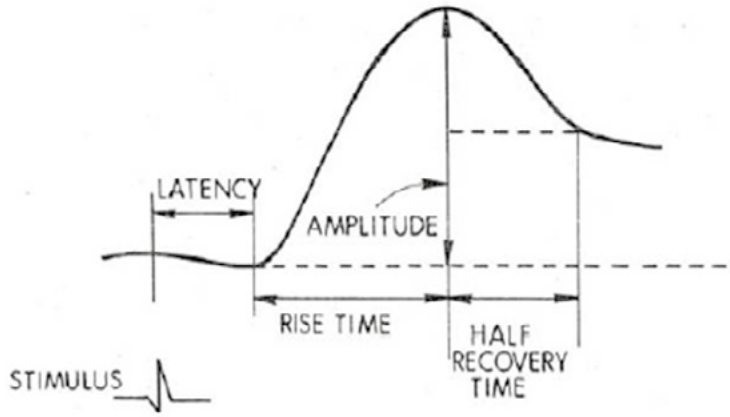


Figure 2. Parameters of phase electrodermal reaction

Source: (Cacioppo et al., 2007, pp. 165–166; Jaśkowski, 2004).

The parameters of the phase electordermal activity indicate certain regularities (Boucsein, 2012):

1) the more important a given stimulus is for a given person, the greater the amplitude of the reaction and the longer its recovery time;
2) the higher amplitude of the phase electrodermal reaction, the stronger emotional arousal is;
3) the longer recovery time a phase reaction is, the more increased the attention to a specific task.

In addition to measurement of the phase electrodermal response caused by the short-term stimuli, it is also possible to measure the response caused by the sustained stimuli (lasting over a long period of time). In this case, the change in tonic level (SCL) requires measurement. The change in tonic level is defined as the difference in its level between at least two points in time.

The measures of the tonic and phase electrodermal activity have specific, typical values (Table 1). It should be noted, however, that the electordermal reaction is very individual. It depends, inter alia, on: age, sex, race or the characteristic properties of the skin regarding the person under study (Cacioppo et al., 2007).

Table 1. Electrodermal measures, definitions and typical values

Measure	Definition	Typical values
Skin conductance level (SCL)	Tonic level of skin electrical conductivity	2–20 microSiemens
Change in SCL	Gradual changes in SCL measured at two or more points in time	1–3 microSiemens
Frequency of NS-SCRs	Number of SCRs in absence of identifiable eliciting stimulus	1–3 per minute
SCR amplitude	Phasic increase in conductance shortly following stimulus onset	0,1–1 microSiemens
SCR latency	Temporal interval between stimulus onset and SCR initiation	1–3 seconds
SCR rise time	SCR rise time	1–3 seconds
SCR half recovery time	Temporal interval between SCR peak and point of 50% SCR amplitude recovery	2–10 seconds

Source: (Cacioppo et al., 2007, p. 165).

Where is electordermal activity measured?

Electrodermal activity is measured on the skin surface (Strelau, 2006). Due to the fact that the highest sweat gland densities are on the hands and feet, these parts of the body are the main place for physiological measurements. However, the clear advantage of the hand in this respect is a consequence of the much easier usage of the measuring equipment. There is no clear suggestion in the literature as to on which hand the skin's electrical activity should be measured. The most often, the non-dominant hand is used for practical reasons. Nonetheless, the areas of the hand on which the measurement should be performed are relatively, precisely defined. These are the distal phalanges and the middle phalanges on the index and middle fingers, as well as the ball of the thumb and the little finger. Alternatively, the measurement can be carried out on the wrist. The measurement is taken by attaching electrodes to skin surface. The areas of the hand on which it is possible to measure electrodermal response (attach electrodes) are shown in Figure 3.

Figure 3. Locations for recording electrodermal activity

Source: Training materials of the NuroDevice company.

When deciding on the places where electrodermal activity is recorded, the following conditions should be taken into account:

1) recording the electrodermal activity from the subject's fingers gives a good signal (good data acquisition) but may prevent the subject from moving his/her hand freely;
2) recording the electrodermal activity from the subject's wrist makes it less difficult for the subject to move the hand, but gives a weaker signal (poorer data acquisition).

What equipment is used to measure electrodermal activity?

Measurements of electrodermal activity is performed while a small current is flowing through the skin from an external source. Therefore, this measurement cannot be done without dedicated equipment. It requires the use of special electrodes, electrode gels and recording devices. Its main element is the so-called biological signal acquisition station. The electrodes are connected to this station by a wire which, in turn, are attached (most often) to the hand of the participant under study. The obtained data is sent from the acquisition station to a computer, on which appropriate software is installed and allows for analysis. Such a set of apparatus allows to conduct research during which the participants are not required to move around.

On the other hand, research conducted in natural conditions, requiring the movement of people (e.g. inside a store), requires a slightly different configuration

of the apparatus. In that case it is impossible to connect the electrodes directly to a small device that is attached to the subject's forearm with a band. It records electordermal activity data. Then, this data is sent to the computer (see: Figure 4).

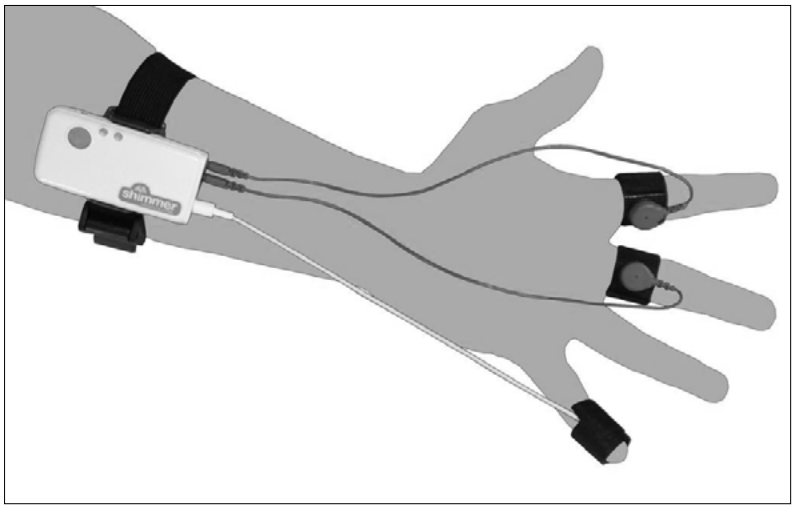


Figure 4. Example of the device used to detect electrodermal activity

Source: (Hernando-Gallego, Artés-Rodríguez, 2015).

The available equipment for analysis of electordermal activity is characterised by a relatively low cost (compared to other devices for physiological measurements) of purchase and operation. After the initial expense related to the acquisition of the measuring equipment itself, further use requires periodic purchases of appropriate consumables (gel or electrodes). Moreover, the EDA measurement is non-invasive and carries no risk to the health or life of the test subjects.

What needs to remembered when conducting electrodermal activity research?
The proper use of psychophysiological methods—including measurement of electrodermal activity—requires the application of several fundamental principles (Białowąs & Szyszka, 2019). First of all, one needs to **design an experiment in such a way that makes it possible do determine whether a given SCR is event-related (experiment related) or non-specific.** If the criteria in the experiment are too loose, one risks including non-specific SCRs into the analysis for event-related SCRs, and erroneously, this could led to misleading results. On the other hand, strict criteria may end in missing many ER-SCRs to meet the adopted criteria by wrongly discarding or misclassifying them as NS-SCRs (Braithwaite, Watson, Jones, & Rowe, 2015). Apart from a proper experiment design, there is also a set of good practices that facilitate electrodermal activity testing. They are the following:

1) the device should be put on the participant a few minutes before the test—this will improve the contact of the electrodes with the skin;
2) the respondent should be asked to perform an exercise, e.g. breathe in and out deeply (this will increase the EDA signal);
3) the right temperature should be set in the room—optimally, 22–24°C;
4) the number of artifacts related to movement should be reduced;
5) the presence of physiological activities of the body should be noted (coughing, deep inhalation, conversation)—they cause the generation of SCR;
6) a larger number of people should be recruited for the research—approx. 10% of the population is hyporesponsive.

After the examination, attention should also be paid to the record of the obtained electrodermal activity. Recordings that raise doubts should be excluded. Below, in Figure 5, a correct record of electrodermal activity is presented. In red, phase reactions are indicated. Each of them are marked with a 'drop'. Whereas in Figure 6, an erroneous record is shown. It results from the loss of contact between the electrodes and the palm of the participant at some point of the test.

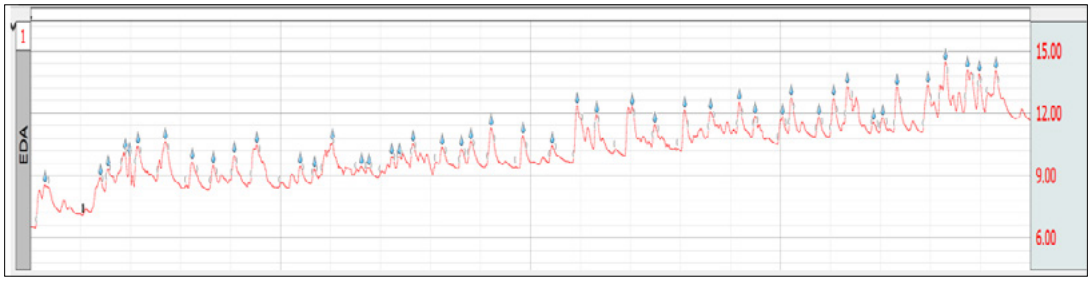


Figure 5. Record of correct tonic and phase electrodermal reaction

Source: (Pierański, 2019, p. 184).

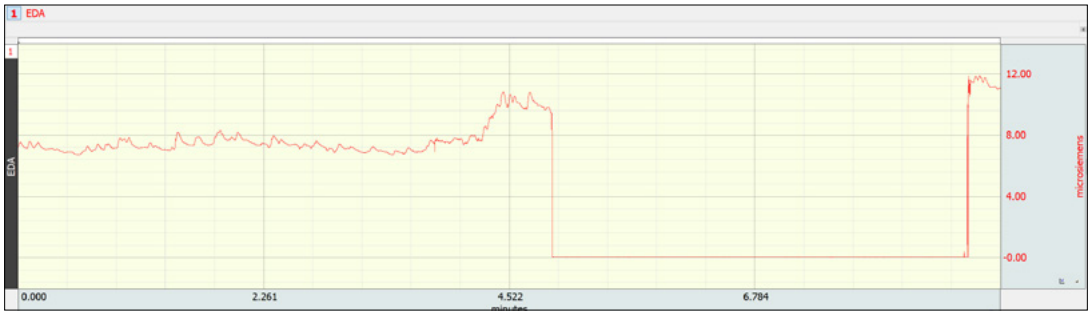


Figure 6. Record of electrodermal reaction indicating loss of contact between electrodes and skin of the participant

Source: (Pierański, 2019, p. 181).

What data can be obtained from measuring electrodermal activity?
Three measurements are used to extract information from events. These event measurements can provide quick summaries of event information, compute mean intervals between event types, and detail other operations.

There are three types of measurements of electrodermal activity:

1) Event Amplitude Measurement—extracts measurement results where events are defined:
   - Sum of amplitudes of all electrodermal reactions—presents the sum of the value for all events within the selected period of time;
   - Mean amplitude from all electrodermal reactions—presents the average amplitude value for all events within the selected period of time;
   - Minimum amplitude from all electrodermal reactions—presents the minimum amplitude value for all events within the selected period of time;
   - Maximum amplitude from all electrodermal reactions—presents the maximum amplitude value for all events within the selected period of time;
   - Median value of amplitude from all electrodermal reactions—presents the median amplitude value for all events within selected period of time;
   - Peak to peak interval of the set of amplitudes from all electrodermal reactions—takes the peak-to-peak difference from the set of amplitudes for all events (max–min);
   - Standard deviation of amplitudes from all electrodermal reactions—presents the standard deviation of the set of amplitudes for all events.
2) Event Count Measurement—evaluates the number of electrodermal reactions within the selected period of time.
3) Event Location Measurement—extracts information about the times of electrodermal reactions.

2.4. How to successfully conduct experiments on EDA (step-by-step guide)

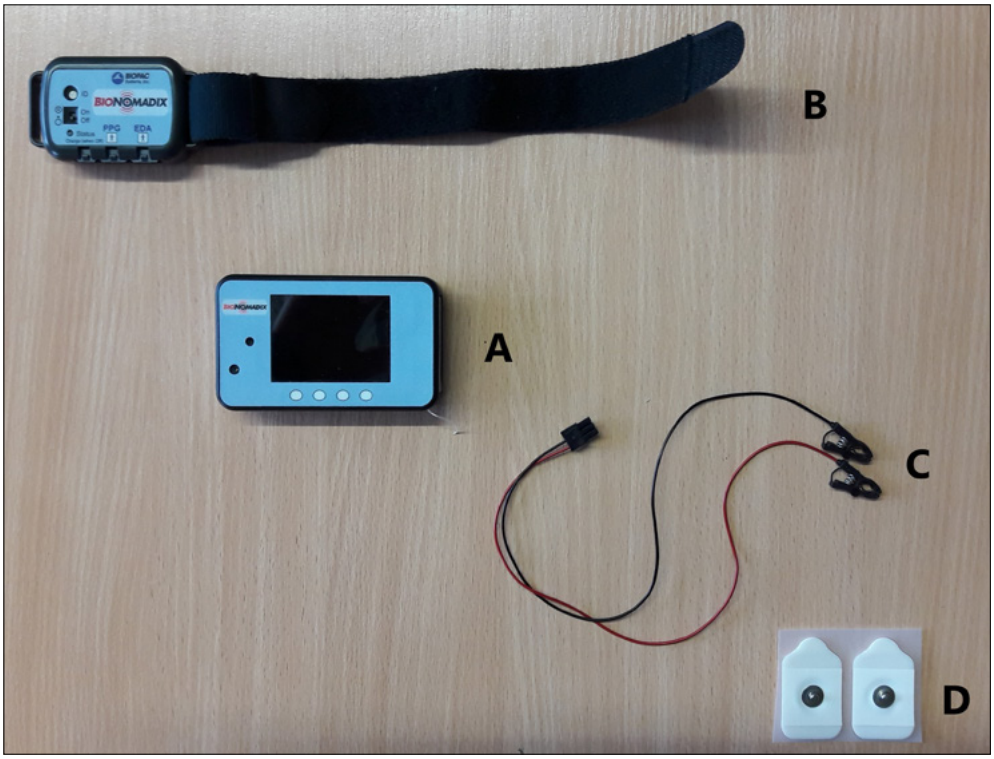
In this chapter, BIOPAC systems and software (AcqKnowledge) for EDA analysis are presented.

2.4.1. Equipment preparation

The first step in conducting experiments is to prepare relevant equipment. The research equipment that is described in the chapter consists of the following elements (see Figure 7):

1) logger—that wirelessly acquires and stores biometric data (EDA included);
2) transmitter—that applies electrical potential between two points of skin contact and measures the resulting current flow between them;
3) wire—that connects electrodes with the transmitter;
4) electrodes—two of them.

This set is suitable for conducting experiments in natural conditions, requiring the movement of people (e.g. inside a store).



A – logger, B – transmitter, C – wire, D – electrodes

Figure 7. Components of equipment used to measure EDA

Source: Own compilation.

At the initial stage of each experiment, a wireless connection between the logger and transmitter needs to be established. In order to do so, please follow the instructions presented in Figures 8–11.

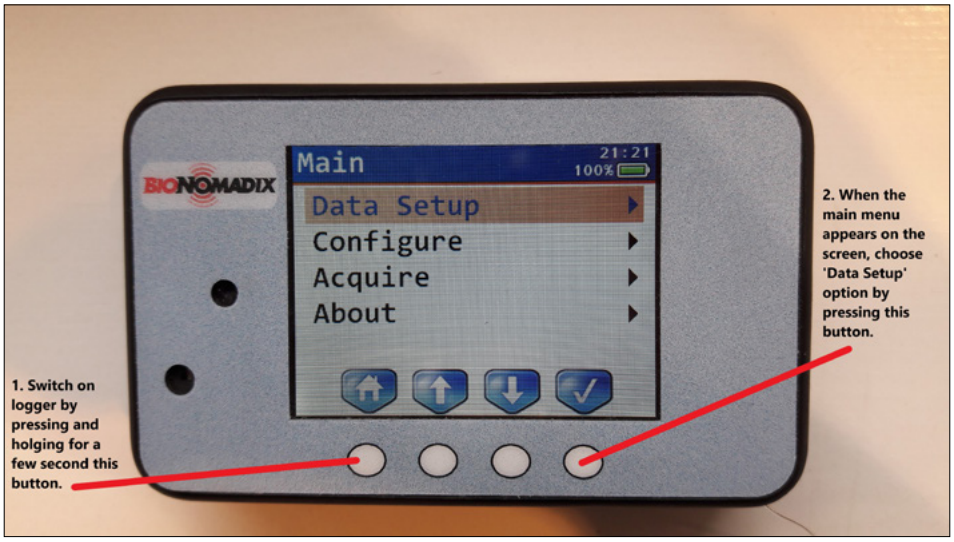


Figure 8. Preparing equipment—step 1

Source: Own compilation.

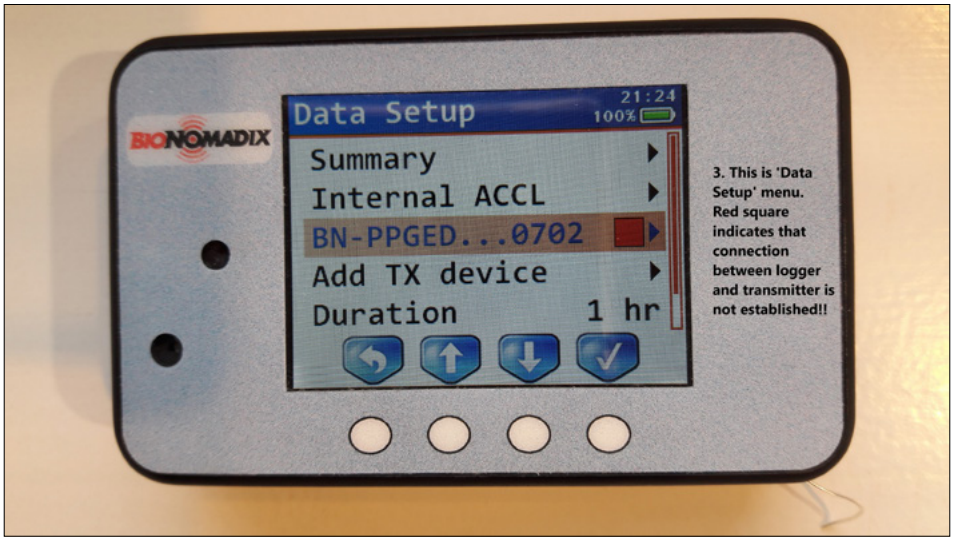


Figure 9. Preparing equipment—step 2

Source: Own compilation.

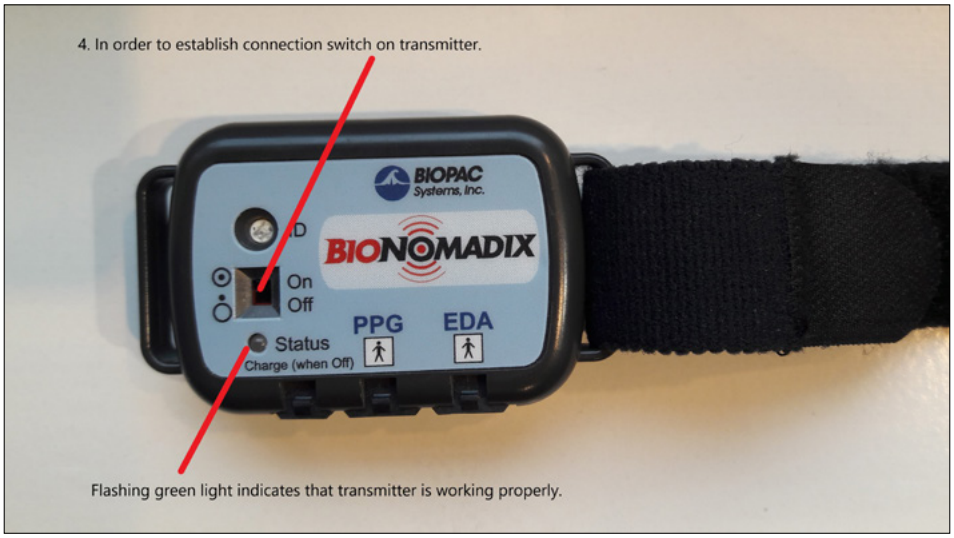


Figure 10. Preparing equipment—step 3

Source: Own compilation.

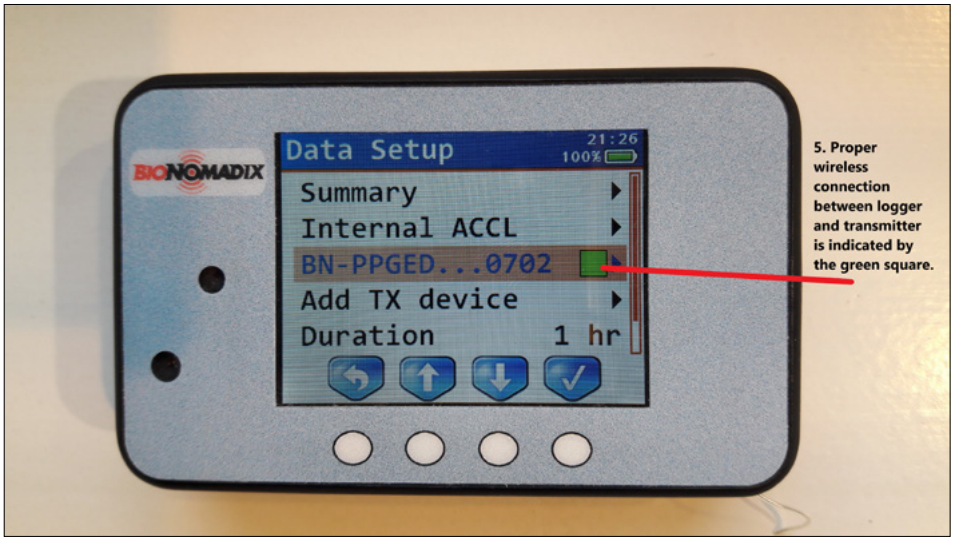


Figure 11. Preparing equipment—step 4

Source: Own compilation.

Once a connection between the logger and transmitter is established, the next step is to put the equipment on the participant's forearm. To do this correctly, please follow the sequence presented in Figure 12.

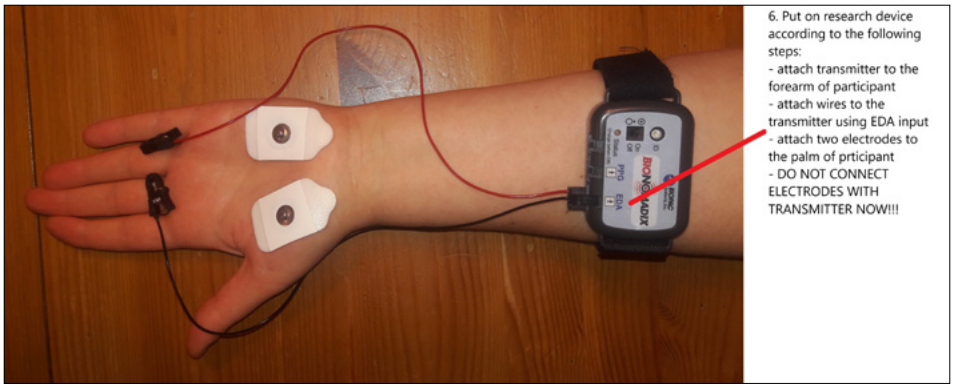
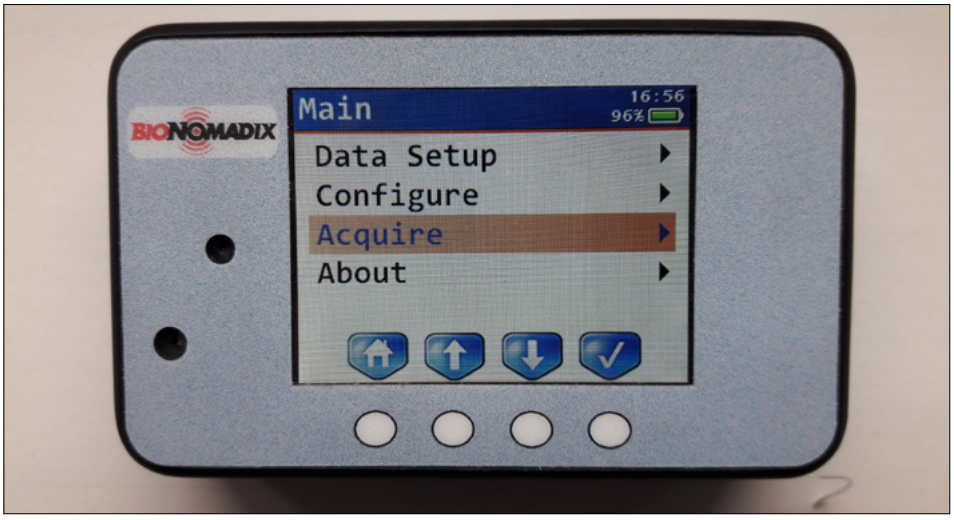


Figure 12. Preparing equipment—step 5

Source: Own compilation.

2.4.2. Acquiring EDA data

After establishing a wireless connection between the transmitter and logger, as well as putting the research device on the participant's forearm, the next part of the experiment is data acquisition. In order to record the required data, please follow the steps described in Figures 13–19.



Note: From the main menu, choose 'Acquire' by pressing the far right button.

Figure 13. Data acquisition—step 1

Source: Own compilation.

Note: From the 'Acquire' menu, choose 'Start' by pressing the far right button.

Figure 14. Data acquisition—step 1

Source: Own compilation.

Note: Confirm data acquisition by pressing the far right button.

Figure 15. Data acquisition—step 3

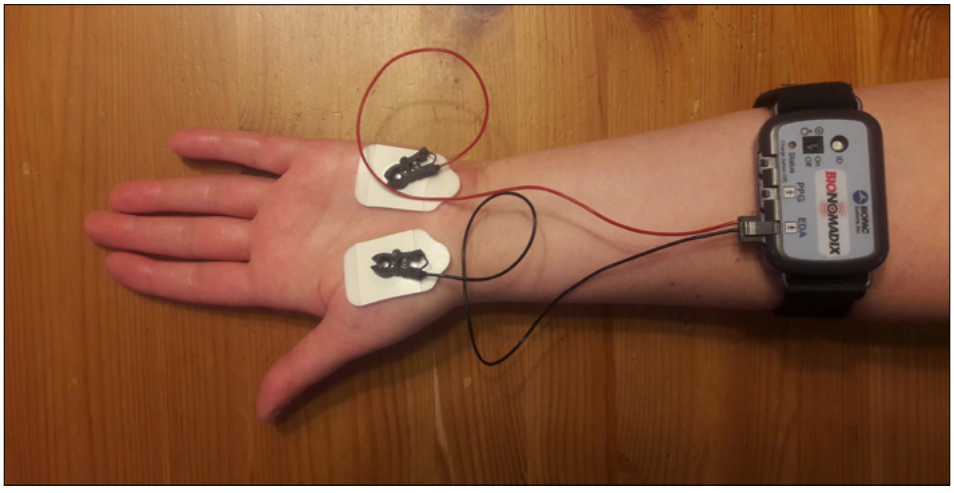
Source: Own compilation.

Note: This is what the logger screen should look like during the process of data recording.

Figure 16. Data acquisition—step 4

Source: Own compilation.

After starting the recording, electrodes can be connected to the transmitter (see: Figure 17).



Note: Proper connection of electrodes to transmitter.

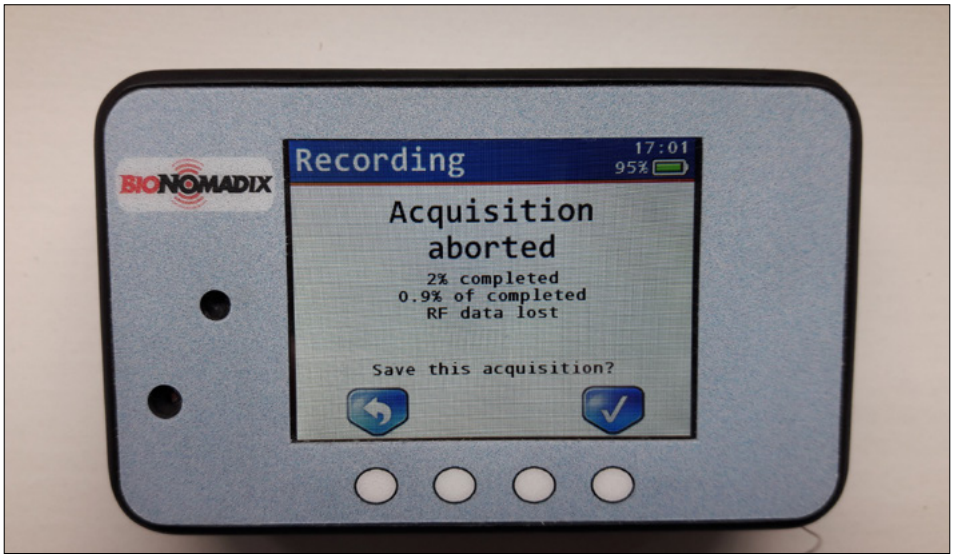
Figure 17. Data acquisition—step 5

Source: Own compilation.

Note: In order to terminate data acquisition, first press and hold the far left button for a few seconds, then press the far right button.

Figure 18. Data acquisition—step 6

Source: Own compilation.



Note: Save acquired data by pressing the far right button.

Figure 19. Data acquisition—step 7

Source: Own compilation.

2.4.3. Analysing EDA data

After acquiring data, the next step is its analysis. The logger must be connected to a computer with downloaded AcqKnowledge software. And then, follow the steps described in Figures 20–25.

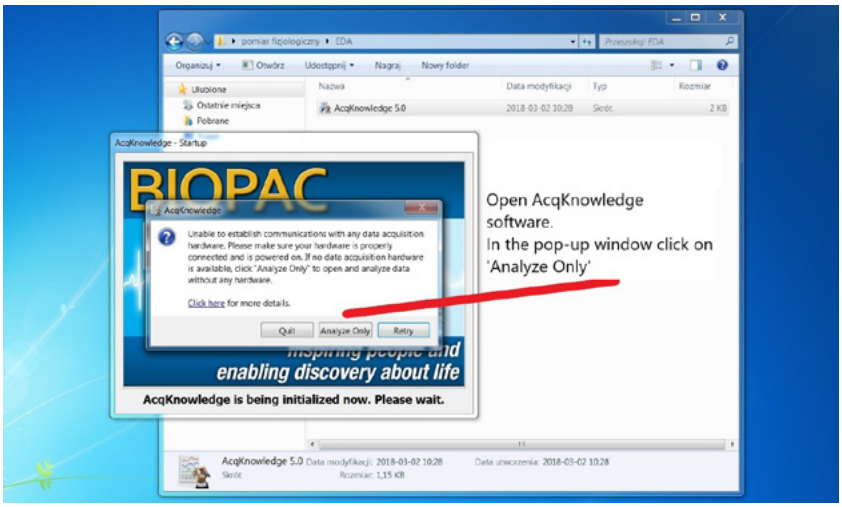


Figure 20. EDA data analysis—step 1

Source: Own compilation.

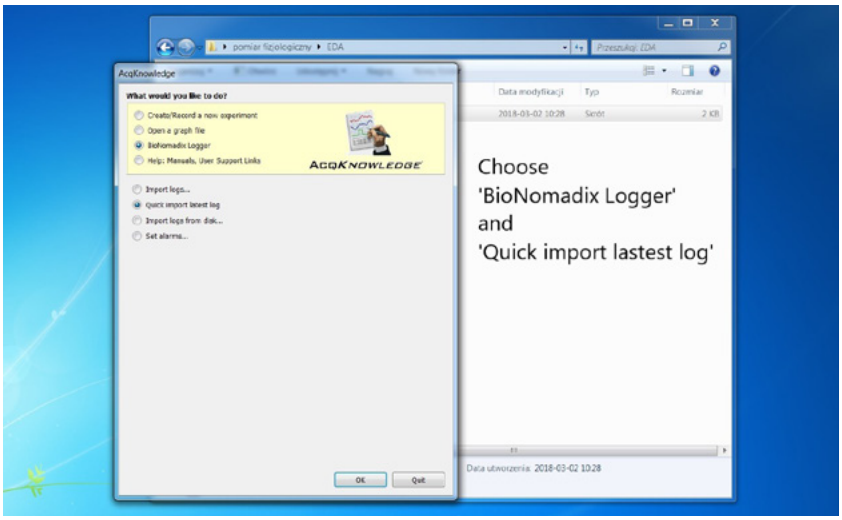


Figure 21. EDA data analysis—step 2

Source: Own compilation.

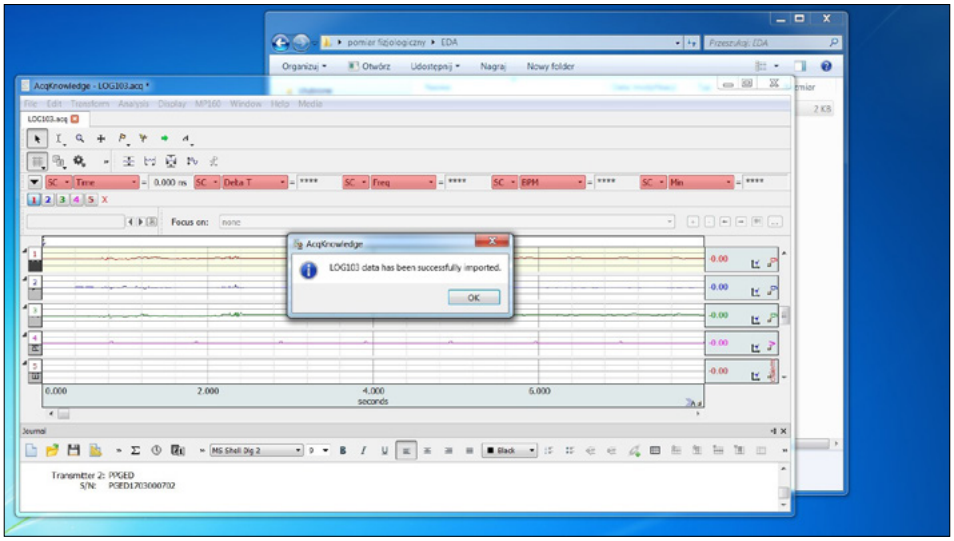


Figure 22. EDA data analysis—step 3: importing EDA data from logger to AcqKnowledge software

Source: Own compilation.

After importing data from logger, several channels (graphs) may be presented in the AcqKnowledge software. In order not to analyse graphs that not relate to electrodermal activity (in Figure 23, graphs: X, Y, Z and PPG), from the pop-up menu, select 'Channels' and unclick unwanted channels (only EDA channel should be marked). In this case, only EDA channel remains on the screen and can be analysed.

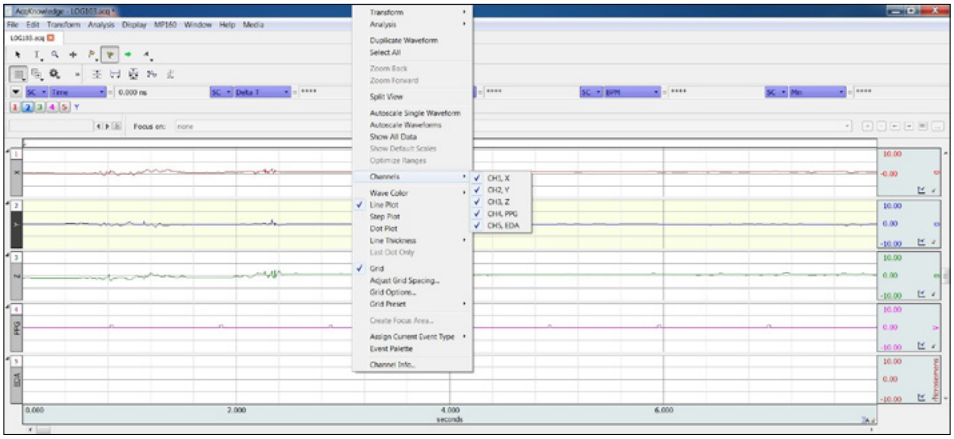


Figure 23. EDA data analysis—step 4: selecting EDA channel

Source: Own compilation.

To establish emotional response to experimental stimuli, Phasic EDA needs to be analysed. This type of electrodermal activity has to be derived from tonic EDA, that is—by default—recorded by logger. In Figure 24, the process of obtaining Phasic EDA (from the EDA channel) is presented.

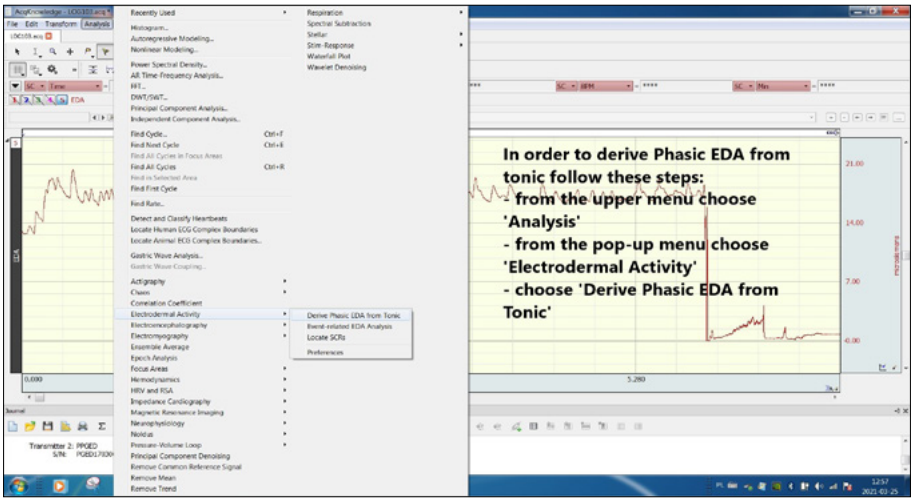


Figure 24. EDA data analysis—step 5: deriving phase EDA from tonic

Source: Own compilation.

What the screen should look like after deriving Phasic EDA from tonic is presented in Figure 25. There must be two channels: EDA and Phasic EDA (Phasic EDA is highlighted in yellow).

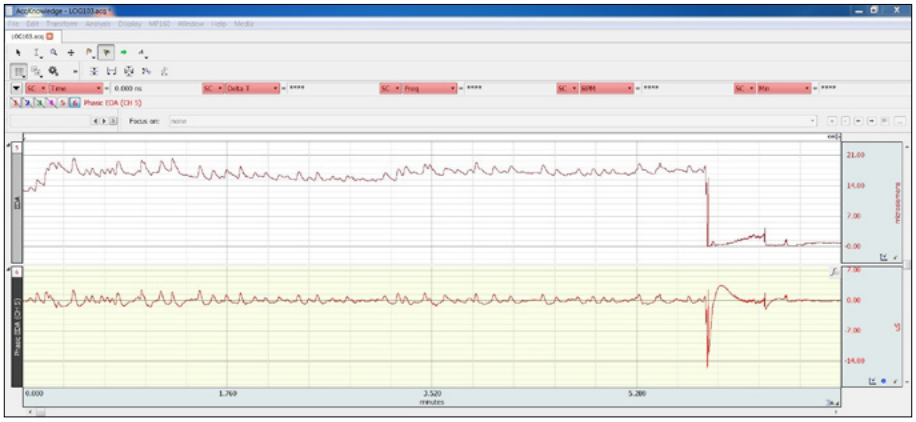


Figure 25. EDA data analysis—derived Phasic EDA (graph highlighted in yellow)

Source: Own compilation.

There are a few main ways of exploring and analysing EDA data. The most commonly used is the so-called I-Beam procedure. I-Beam analysis is based on number of electrodermal activity measurements that can be selected by the researcher. The procedure starts by choosing the required measurements from 'measurement boxes'. In Figure 26, the location of measurement boxes is provided. Please note that each measurement can be assigned to a specific channel (each graph represents one channel). In Figure 26, it can also be seen where the required channel can be set. The best option is to choose 'SC', which stands for 'selected channel'. The channel selection can be done by selecting a specific graph (by clicking it and making it highlighted in yellow). It is recommended to assign all measurements to one channel only. In Figure 27, it can be observed which measurements can be selected for each measurement box.

To provide information in measurement boxes, specific regions of the signal need to be highlighted for analysis using the I-Beam tool. This tool works in conjunction with the measurement boxes (which provide output from the region highlighted by the I-Beam tool) (Braithwaite et al., 2015). In Figures 28 and 29, it is explained how to proceed in this case.

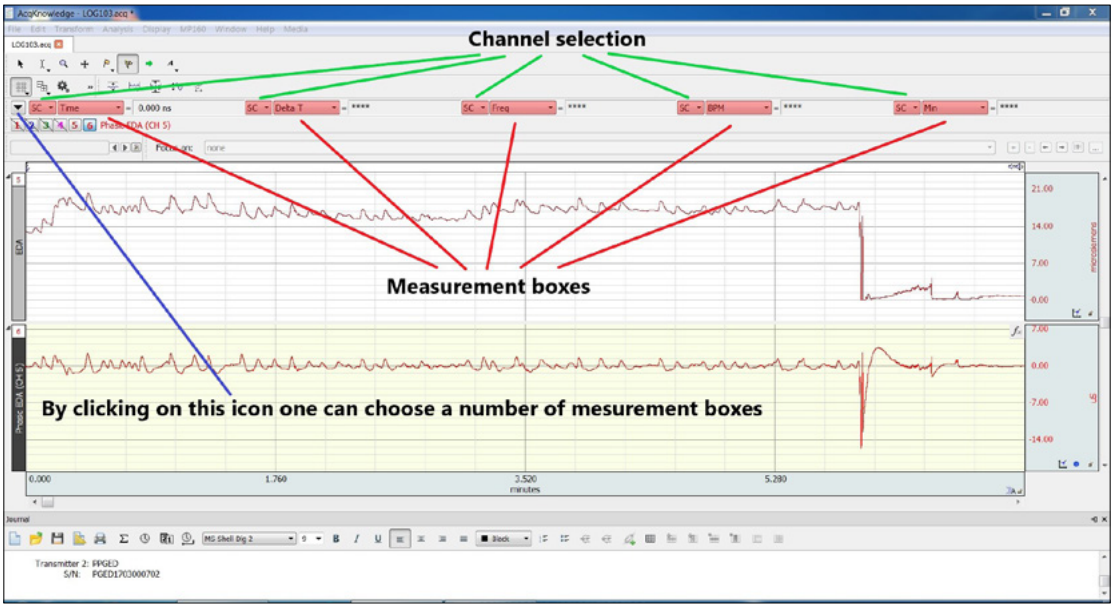


Figure 26. EDA data analysis—step 6: selecting measurement boxes and channels

Source: Own compilation.

Please note: The list pop-ups when clicking on specific measurement box.

Figure 27. EDA data analysis—step 7: choosing EDA measurements

Source: Own compilation.

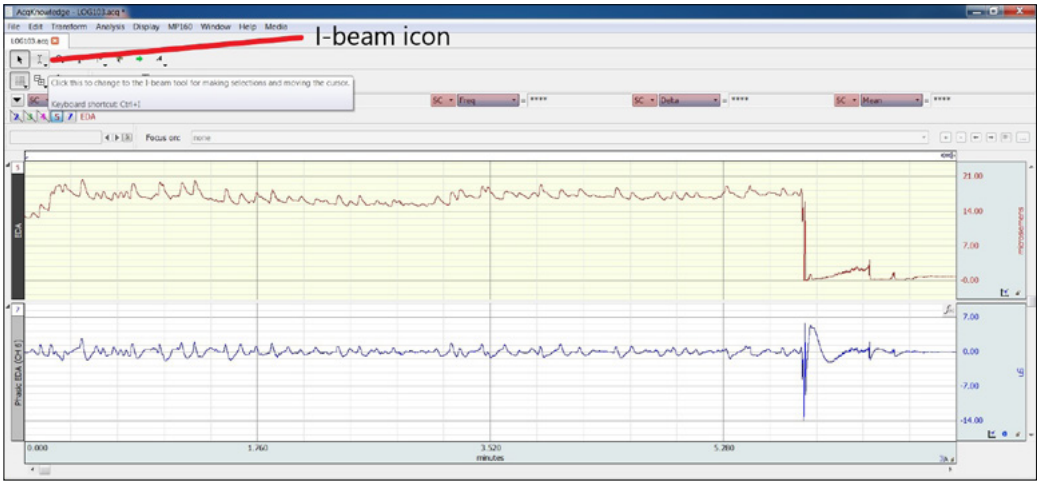


Figure 28. EDA data analysis—step 8: choosing I-Beam analysis

Source: Own compilation.

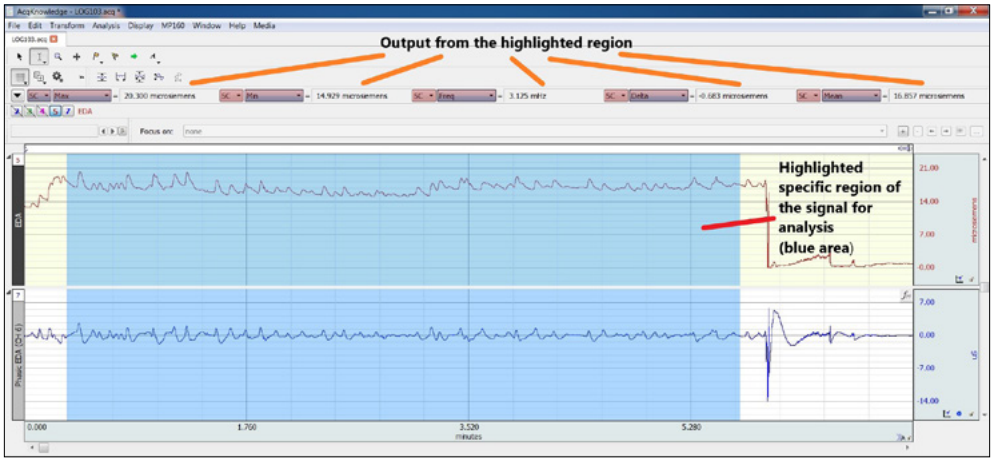
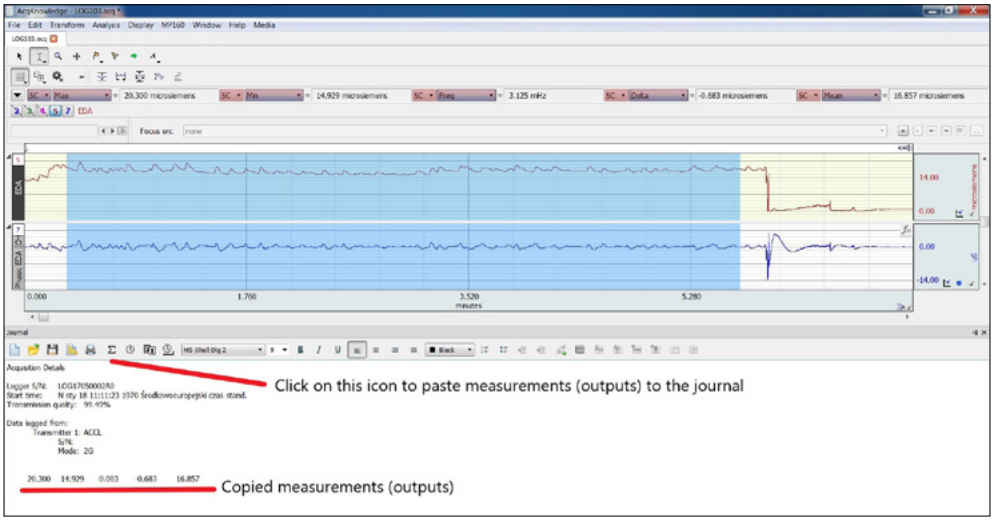


Figure 29. EDA data analysis—step 9: example of highlighted region of signal and corresponding measurement boxes providing output from the highlighted area

Source: Own compilation.

The final part EDA data analysis is to copy selected measurements to the so-called 'Journal'. This is a part of AcqKnowledge software (visible at the bottom of the screen) that allows to export data to statistical software for advanced analysis (see: Figure 30).



Please note: In Figure 30, all measurements refer to the EDA channel that is highlighted in yellow.

Figure 30. EDA data analysis—step 10: copying selected measurements to the journal

Source: Own compilation.

2.5. Case study—Perception of a humanoid robot[1]

In an effort to streamline client services in selected branches, a financial institution operating on the Slovak market wanted to find out how people react to interaction with a humanoid robot. The goal of the humanoid robot was to act as a navigator to guide the client with respect to the problem/service the client needs to solve. The institution decided to carry out a qualitative ad-hoc survey using biometric tools in order to reveal real perception and emotional feedback due to interaction with the robot.

The main objectives of the project were defined as follows:

1) information about real emotional feedback;
2) identification of stressful parts regarding the interaction;
3) comparison of the declarative part through in-depth interviews and unconscious perception.

The testing was performed using in-depth interview and biometric tools (eye-tracking and measurement of electrodermal activity), as well as by implementing the neuroimaging method of mobile electroencephalography (EEG). The experiment included 8 participants, with whom an initial interview was conducted immediately after their arrival and then, they visited the particular branch in order to interact with the robot. During the interaction with the robot, immediately after arriving to the branch, the subjects were monitored for visual and emotional feedback. After completing the practical interaction, a second in-depth interview was conducted with the respondents.

From the graph presented in Figure 31, it is possible to note the average values of skin resistance recorded during interaction with the robot. A more significant decrease in resistance and thus, a higher level of emotional arousal (nervous irritation), can be observed during eye contact with a humanoid robot and subsequent communication with it (event indicated with a red vertical line). These results can be largely influenced by the uncertainty of the respondents as to how the whole process of interaction/solution of the banking operation will take place (entering a request on the display, real communication with the robot, etc.).

The aim of the graphs in Figure 31 is to demonstrate which parts of solving the banking process were more frustrating for participants in comparison to others. The statements of 7 respondents indicating that they would appreciate if this technology became a common and everyday part at the branch of the institution, are also the proof of this.

In Figure 32, the individual values of male skin resistance are shown, in which the technology of the humanoid robot did not work properly (rotation of the head

[1] Please note that in the following case study, skin resistance was measured. In that case, the lower the level—the higher the emotional arousal.

and speech failure of the robot) due to disconnection of the remote control. Based on the observed range, one can state that this moment was very emotional for the respondent (decrease of skin resistance at approximately 1,000 kOhm).

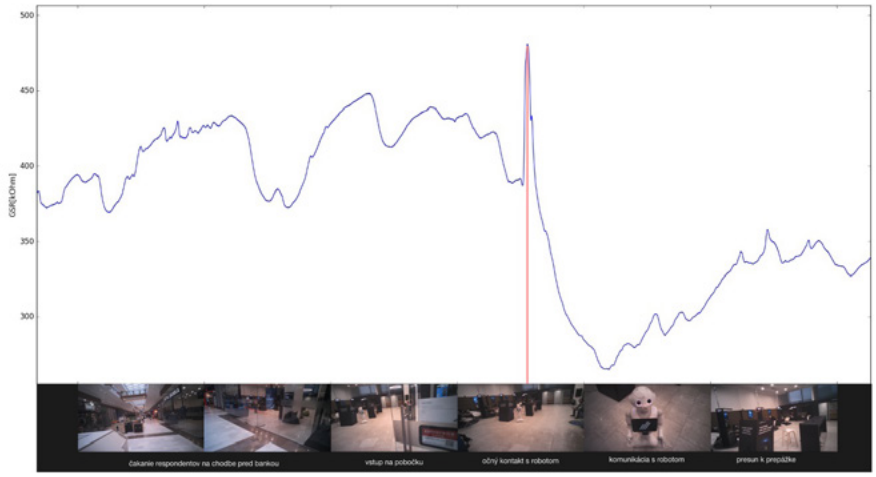


Figure 31. Average value of skin resistance (kOhm) when interacting with a robot

Source: Own compilation based on research from 2019.

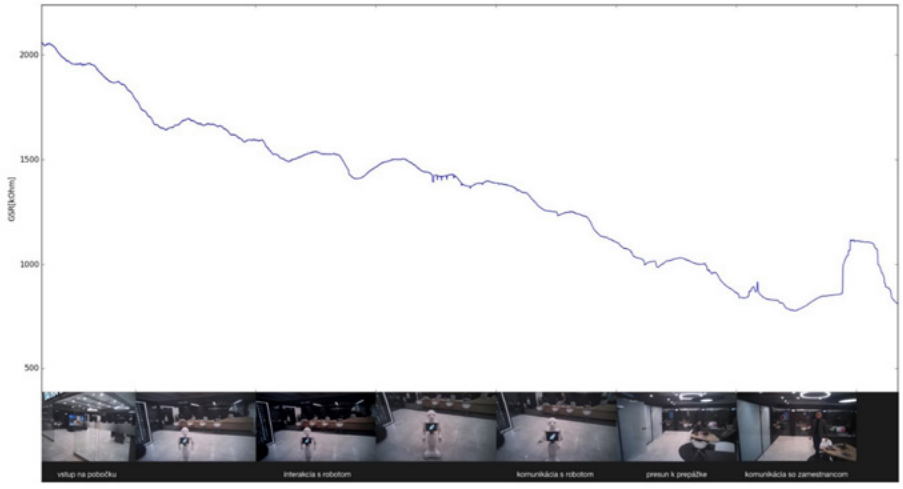


Figure 32. Average value of the skin resistance (kOhm) for an individual when interacting with a robot that did not work properly

Source: Own compilation based on research from 2019.

Although the traditional research tools are effective, there are situations in which other forms of innovative approaches are needed, mainly focused on subconscious perception. The combination of traditional and biometric tools, which include the measurement of skin resistance, appears to be an effective tool for obtaining a realistic image of human perception.

References

Białowąs, S., & Szyszka, A. (2019). Measurement of electrodermal activity in marketing research. In R. Romanowski (Ed.), *Managing economic innovations—methods and instruments*. Poznań: Bogucki Wydawnictwo Naukowe. https://doi.org/10.12657/9788379862771-5

Benedek, M., & Kaernbach, Ch. (2010). A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods*, *190*, 80–91.

Berčík, J., & Rybanská, J. (2017). Methods used in neuromarketing. In *Neuromarketing and food retailing* (pp. 83-101). Wageningen: Wageningen Academic Publishers.

Boucsein, W. (2012). *Electrodermal activity*. 2nd edition. New York: Springer.

Braithwaite, J. J., Watson, D. G., Jones, R., & Rowe, M. (2015). *A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments*. Retrieved from https://www.biopac.com/wp-content/uploads/EDA-SCR-Analysis.pdf

Cacioppo, J. F, Tassinary, L. G. & Berntson, G. G. (2007). *Handbook of psychophysiology*. Cambridge: Cambridge University Press.

Edelberg, R. (1972). Electrical activity of the skin: Its measurement and uses in psychophysiology. In N. S. Greenfield, & R. A. Sternbach (Eds.), *Handbook of psychophysiology* (pp. 367-418). New York: Holt.

Galvanic Skin Response (GSR): *The complete pocket guide*. (2016, March 15). iMotions, Bimetric Research, Simplifed. Retrieved from https://imotions.com/guides/

Hernando-Gallego, F., & Artés-Rodríguez, A. (2015). *Individual performance calibration using physiological stress signals*. Retrieved February 13, 2021 from https://arxiv.org/pdf/1507.03482.pdf

Jaśkowski, P. (2004). *Zarys psychofizjologii*. Warszawa: Wyższa Szkoła Finansów i Zarządzania—School of Business and Finance.

Lacey, J. I., Kagan, J., Lacey, B. C., & Moss, H. A. (1963). The visceral level: Situational determinants and behavioral correlates of autonomic response patterns. In P. H. Knapp (Ed.), *Expression of the emotions in man* (pp. 161-196). New York: International Universities Press.

Pradeep, A. (2010). *The buying brain: Secrets for selling to the subconscious mind*. New Jersey: John Wiley.

Pierański, B. (2019). *Pomiar fizjologiczny w badaniu wewnątrzsklepowych zachowań nabywców*. Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu—Poznań University of Economics and Business Press

Sosnowski, T., & Zimmer, K. (1993). *Metody psychofizjologiczne w badaniach psychologicznych*. Warszawa: PWN—Scientific Publishing House.

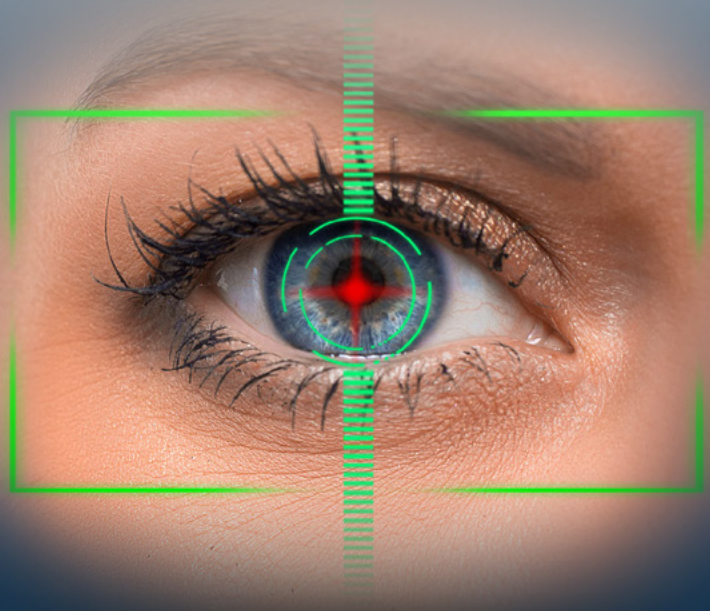
Strelau, J. (Ed.). (2006). *Psychologia. Podręcznik akademicki. Podstawy psychologii*. Gdańsk: Gdańskie Wydawnictwo Psychologiczne.

Wallin, B. G. (1981). Sympathetic nerve activity underlying electrodermal and cardiovascular reactions in man. *Psychophysiology, 18*(4), 470-476.

Zhai, J., Barreto, A. C., Chin, C, & Li, C. (2005). *Realization of stress detection using psychophysiological signals for improvement of human-computer interactions*. Proceedings. IEEE SoutheastCon. https://doi.org/10.1109/SECON.2005.1423280

PART 3.

DATA ANALYSIS



Experimental design and biometric research. Toward innovations, pp. 91-127
https://doi.org/10.18559/978-83-8211-079-1/III1

1.

INDEPENDENT SAMPLES— SINGLE HYPOTHESIS TESTING

Sylwester Białowąs
Poznań University of Economics and Business

Blaženka Knežević
Faculty of Economics and Business, University of Zagreb

Adrianna Szyszka
Poznań University of Economics and Business

Berislav Žmuk
Faculty of Economics and Business, University of Zagreb

Abstract: In this chapter, the "between-subject" is dealt with, as well as the single hypothesis approach. Both parametrical and non-parametrical versions of the tests are described. All tests are introduced, and the full, step-by-step SPSS guidance is presented. The sections regarding effect size and about writing the report are also included.

Keywords: Kruskal-Wallis H test, Mann-Whitney U test, one-way ANOVA, *t*-test.

As described in the first part of the book, the analysis of the experiment results will follow one of two approaches: between-subject and within-subject. This division is reflected in the analytical part. The first two chapters (1 and 2) are devoted to the between-subject approach, first if only one hypothesis is verified, and the second chapter—when more hypotheses are verified. In the last sub-chapter of this part (3), the authors deal with the within-subject approach.

1.1. Independent samples *t*-test

General information

The independent samples *t*-test is one of the most popular statistical tests. It is used to compare the means of two groups (e.g. age, height, balance in a savings account, bio food expenses, exam scores). It is a basic test in experimental designs when one group is a control group (e.g. receives placebo or usual treatment), while the other one is administered what we want to test. In the *t*-test, the means and standard deviations of two groups are computed and it is checked whether there is a statistically significant difference between means. The compared groups may be selected by the researchers while assigning participants to different conditions or may occur naturally (Verma & Abdel-Salam, 2019). It should be borne in mind that the differences between groups may be caused not only by the manipulation of the researcher but also by different aspects that influence variance, such as individual differences or IQ (Field, 2013).

Hypothesis

In order to compare the scores for two groups, the null and alternate hypotheses should be stated. The null hypothesis is that the mean scores in the two groups are equal. This indicates that the observed difference is due to chance alone. The alternate hypothesis is that the means in two groups differ from each other (Lind, Marchal, & Wathen, 2006).

$$H_0: m_1 = m_2$$

$$H_1: m_1 \neq m_2$$

Assumptions

The following assumptions are associated with the independent samples *t*-test:

– the level of measurement should be interval or ratio (what in SPSS is indicated as the scale level of measurement);

- the samples must be disjoint, which means there should be no relationship between the subjects in the groups, the samples should be unrelated to each other;
- samples should be randomly selected, which means that the data constitute a representative portion of the total population and every individual has the same chance to be selected into the sample (Verma & Abdel-Salam, 2019; Waters, 2011);
- the data should follow normal distribution and the dataset should not include outliers. The researcher should check if there are any extreme (unusually high or low) values in the dataset (Verma & Abdel-Salam, 2019);
- the sample should be reasonably large. Although we can technically carry out the *t*-test with a group of any size, the results of the *t*-test are considered stronger with larger samples. It is often recommended that each sample should have about 30 observations, but groups do not necessarily have to include the same number of participants (Lind et al., 2006; Waters, 2011).

Example

Dataset: dwell time of studying information about managing electricity expenses in two groups.

The community managing the apartment blocks has chosen two random groups, each consisting of 105 families living in medium-size flats. (Note, the groups don't have to be equal, they can have different number of cases). Both groups got one page with information on sustainable household management. Electricity management comprised 30% of the page. One of the groups received additional information about the future increase in the price of electricity. The other group was the control (without this info). Using eye-tracking gear, the dwell time in the area of interest (AOI) covering the info about electricity expenses was recorded for every participant.

Data info:

- variable 1: group—nominal, 1—group given the special information, 2—control group;
- variable 2: dwell time—scale, recorded time in seconds spent in the AOI (part about electricity management).

Hypotheses:

H0: There is no difference in dwell time between the groups.

H1: There is a difference in dwell time among both groups.

Testing the assumptions

In the presented example, both groups contain 105 observations, thus the assumption of the group size is met. The size of each group can be read from every SPSS output (e.g. first line of the Kolmogorov-Smirnov test).

Splitting the file—this will cause SPSS to show all the results divided according to the selected groups. In this case, the file division will be carried out according to the variable "group", therefore, all the results will be shown separately for the groups—"control" and "informed". This command is valid until it is revoked. For revoking, please open the dialogue box again and click 'analyze all cases, do not create groups'.

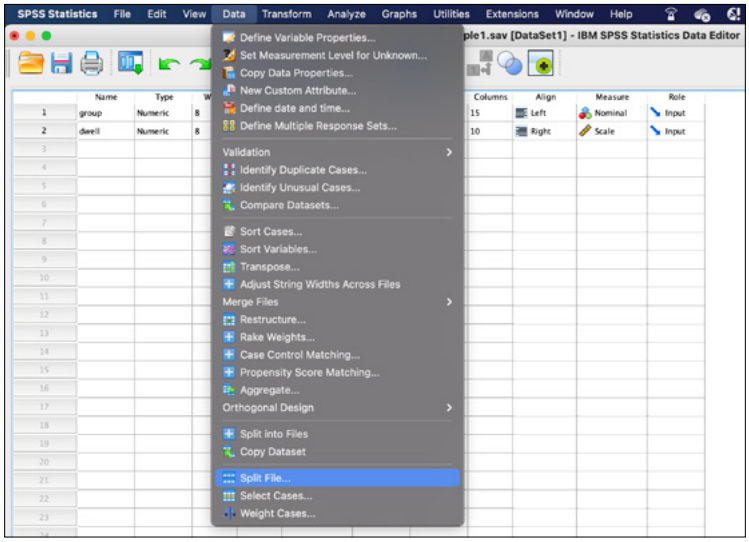


Figure 1. Splitting the file—path

Source: The authors' own elaboration, IBM SPSS screenshot.

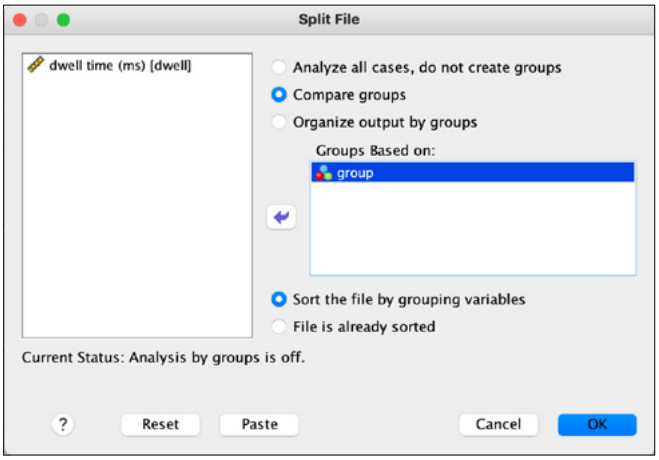


Figure 2. Splitting the file—dialogue box

Source: The authors' own elaboration, IBM SPSS screenshot.

Normality of distribution

The commonly used test for evaluating the normality is the Kolmogorov-Smirnov test. This test allows to compare the set of scores obtained in the study to the normally distributed scores.

Hypotheses for the Kolmogorov-Smirnov test

Null hypothesis (H0): The data follow normal distribution.

Alternate hypothesis (H1): The data significantly differ from normal distribution.

Performing the Kolmogorov-Smirnov test will produce a table with the output for both groups separately (group splitting is still valid).

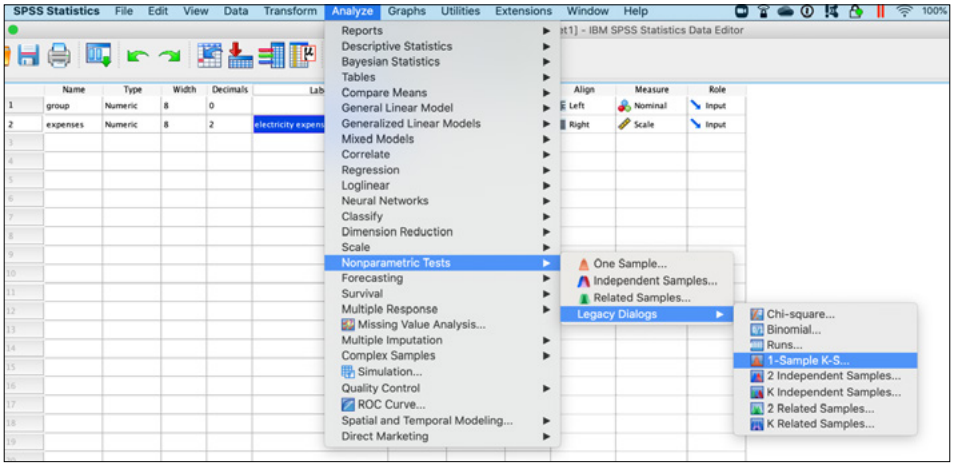


Figure 3. Kolmogorov-Smirnov test for normality of distribution—path

Source: The authors' own elaboration, IBM SPSS screenshot.

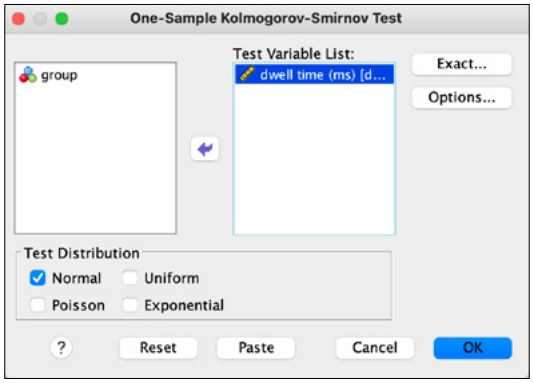


Figure 4. Kolmogorov-Smirnov test for normality of distribution—dialogue box

Source: The authors' own elaboration, IBM SPSS screenshot.

One-Sample Kolmogorov-Smirnov Test

group			dwell time (ms)
1 informed	N		105
	Normal Parameters[a,b]	Mean	6250.686
		Std. Deviation	577.2935
	Most Extreme Differences	Absolute	.058
		Positive	.058
		Negative	-.046
	Test Statistic		.058
	Asymp. Sig. (2-tailed)		.200[c,d]
2 control	N		105
	Normal Parameters[a,b]	Mean	5859.429
		Std. Deviation	522.4182
	Most Extreme Differences	Absolute	.068
		Positive	.051
		Negative	-.068
	Test Statistic		.068
	Asymp. Sig. (2-tailed)		.200[c,d]

Figure 5. Kolmogorov-Smirnov test for normality of distribution—results

Source: The authors' own elaboration, IBM SPSS screenshot.

The hypothesis is determined by interpreting the *p*-value. If the test is significant ($p < .05$), this means that the data do not follow normal distribution. If the test is non-significant ($p > .05$), the distribution of the obtained scores is normal (Field, 2013; Verma & Abdel-Salam, 2019). In this case, for both groups $p = .200$, which indicates that the assumption of normality is fulfilled.

The next step is performing the *t*-test itself. Firstly, the splitting of the groups needs to be revoked by clicking in the command "analyze all cases, do not create groups" in the dialogue box (see Figure 2).

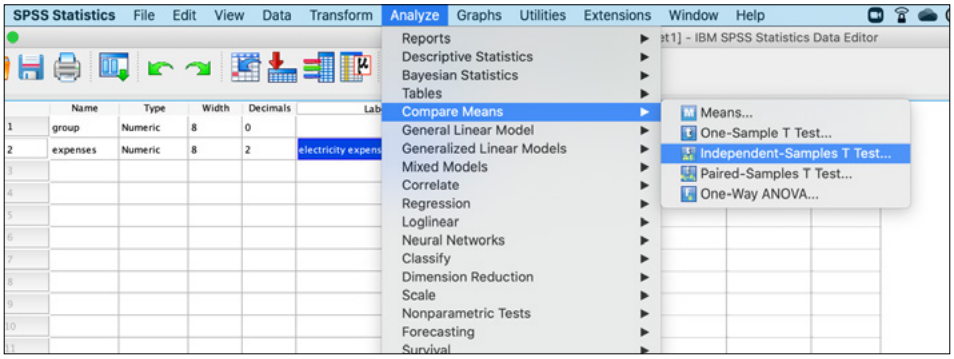


Figure 6. Independent samples *t*-test—path

Source: The authors' own elaboration, IBM SPSS screenshot.

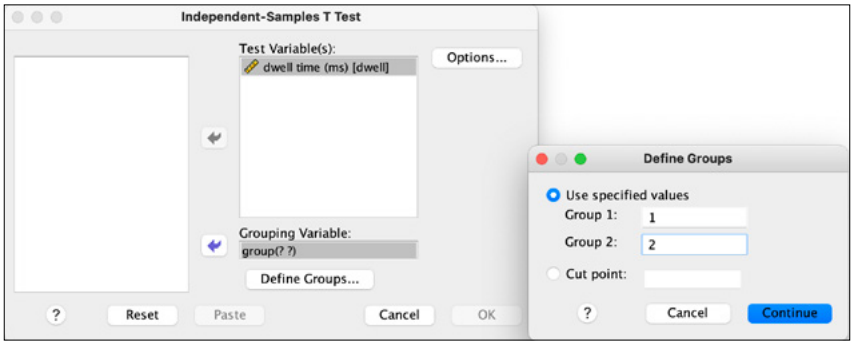


Figure 7. Independent samples *t*-test—dialogue box

Source: The authors' own elaboration, IBM SPSS screenshot.

Group Statistics

	group	N	Mean	Std. Deviation	Std. Error Mean
dwell time (ms)	1 informed	105	6250.686	577.2935	56.3381
	2 control	105	5859.429	522.4182	50.9828

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
dwell time (ms)	Equal variances assumed	1.663	.199	5.149	208	.000	391.2571	75.9817	241.4641	541.0502
	Equal variances not assumed			5.149	205.959	.000	391.2571	75.9817	241.4554	541.0589

Figure 8. Independent samples *t*-test—results

Source: The authors' own elaboration, IBM SPSS screenshot.

Results

The results are interpreted from the lower table (Independent Samples Test). First, the Levene's test of homogeneity in the second column is read (Sig.):

- if $p > .05$, the results are interpreted from the upper row (equal variances assumed);
- if $p < .05$, the results are interpreted from the lower row (equal variances not assumed).

Now, a decision can be made about the significance of the *t*-test. In this case, it equals $p = .199$. This value is greater than the critical value of $p = .05$, indicating that the results will be read from the upper row (equal variances assumed).

In the lower table, it can be checked if the difference is statistically significant by interpreting the *p*-value from the 5th column (Sig. 2-tailed). It can be found that $p < .001$, which is lower than the critical value of $p = .05$. This means that the null

hypothesis can be rejected and the results interpreted as the statistically significant difference between the groups.

In the upper table of the outcome (group statistics), it can be noted that the mean for the informed group is 6250.7, while for the control group it totals 5859.4.

The independent t-test hypotheses resolution:
$p < .05$—there is a significant difference between the groups; reject H0;
$p > .05$—there is no significant difference between the groups; do not reject H0.

Effect size

In order to examine whether the observed difference is significant, the effect size can be calculated. This enables determining the size of the observed effect in a standardised way, which makes the results easy to compare (e.g. with different studies). For the independent samples *t*-test, a popular measure is Cohen's *d*:

$$d = \frac{\left|\overline{x}_1 - \overline{x}_2\right|}{s_{control}}$$

$\overline{x}_1$, $\overline{x}_2$—means of both groups;

$s_{control}$—standard deviation of the control group (Cohen, 1988; Field, 2013).

Cohen's *d* has the following interpretation:
Below 0.2—no effect;
< 0.2–0.5)—small effect;
< 0.5–0.8)—moderate effect;
0.8 and above—large effect.

$$d = \frac{\left|62.51 - 58.59\right|}{5.77} = 0.68$$

In this case, a moderate effect can be observed ($d = 0.68$).

Summary

The community managing the apartment blocks has randomly chosen two groups, each consisting of 105 families living in medium-size flats. Both groups received one page with information on sustainable household management and 30% of the page was devoted to electricity management. One of the groups got additional information about future increases in the price of electricity. The other group was the control (without this info). Using eye-tracking gear, dwell time in the AOI, covering the info about electricity expenses, has been recorded for every participant.

Data info:

- variable 1: group—nominal, 1—group given the special information, 2—control group;
- variable 2: dwell time—scale, recorded time in seconds spent in the area of interest (part about electricity management).

The dwell time for the electricity expenses regarding AOI in the group without extra information was at an average of 5859.4 milliseconds (SD = 522.4), whereas in the informed group—at an average of 6250.7 milliseconds (SD = 577.3). The t-test revealed that the difference of 391.3 milliseconds is statistically significant ($p < .001$, $\alpha = .05$), suggesting that the informed participants focused their attention on the electricity part longer than the control group. The effect size for this analysis (d = 0.68) was found to be moderate.

More info about the *t*-test

The independent samples t-test enable comparison of the scores in two separate groups (populations), and test if there are differences between them. Precisely, the t-test is commonly used in statistics to examine whether the means of two populations are the same (Field, 2013; Verma & Abdel-Salam, 2019; Waters, 2011). Before performing the t-test, the above-mentioned assumptions should be fulfilled (see the 'Assumptions' part in this chapter). However, there is another assumption that has not been mentioned in this chapter so far—the homogeneity of variance regarding the data. This means that the samples should be selected from populations that have equal variance with reference to some criterion. The reason for not mentioning the homogeneity of variance is because performing the t-test in SPSS enables interpretation of the results even if this assumption is violated. Specifically, together with the t-test output, this generates Levene's test and calculates the results for both equal and unequal variances (in case of lack of homogeneity between groups, the results can be read from the lower row). What also should be emphasized is that violating the homogeneity of variance assumption applies only if the sizes of tested groups are unequal (Field, 2013). However, the other assumptions may sometimes be violated as well and there are certain ways to deal with some of them. If the assumption of normal distribution is not fulfilled, there are techniques to convert the data distribution into at least quasi-normal (e.g. log, root, or Box-Cox transformation). If this is not possible, non-parametric tests should be used (Verma & Abdel-Salam, 2019).

It should be noted that the Kolmogorov-Smirnov test is not the only way of checking the normality of distribution. Another popular test that is usually used for this purpose is the Shapiro-Wilk test (considered as better for smaller samples). The hypotheses and interpretation of test statistics are analogical. However, among researchers, there are discussions about the necessity of testing normality before using the independent sample t-test. That is because if tests depend on the hypothesis testing, this may consequently show significant effects for large samples, even in case

of irrelevant influences. On the other hand, for smaller samples, the test results may not indicate that the assumption is violated (Field, 2013). For more experienced investigators, the normality of distribution can be assessed using histograms or by assessing skewness and kurtosis.

For effect size calculation, the simplified formula of Cohen's *d* was proposed with standard deviation of the control group in the denominator. This approach is justified because it can be assumed that the treatment in the study may affect not only the mean, but also dispersion in the dataset. However, there are other possibilities of standard deviation calculations used as a standardiser. The commonly accepted formula has pooled standard deviation in the denominator that is given by the following equation:

$$s^{\star} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

This formula finds its application especially when there is a remarkable difference between standard deviations of both groups. Its advantage also depends on including the sizes of samples (Borenstein, Hedges, Higgins, & Rothstein, 2009; Cumming, 2012; Dean & Illowsky, 2013; Field, 2013).

Furthermore, when calculating effect size, there is inconsistency in terminology. The formula with no pooling (with control standard deviation in the denominator) is also referred to as Glass' *d* or Glass' *Δ*. The researchers sometimes refer to Hedge's *g* as the measure with pooled standard deviation as a standardiser. Recently, using the *d* for all different formulas prevails. However, it is crucial to explain how the effect size was calculated (Cumming, 2012).

References

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Effect sizes based on means. In *Introduction to meta-analysis*. John Wiley & Sons.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge Taylor & Francis Group.

Dean, S., & Illowsky, B. (2013). *Introductory statistics*. OpenStax College.

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics* (5th ed.). Sage edge.

Lind, D. A., Marchal, W. G., & Wathen, S. A. (2006). *Basic statistics for business & economics* (5th ed.). McGraw-Hill.

Verma, J. P., & Abdel-Salam, G. A.-S. (2019). *Testing statistical assumptions in research*. John Wiley & Sons, Inc.

Waters, D. (2011). *Quantitative methods for business* (5th ed.). Pearson Education Limited.

1.2. Mann-Whitney U test

General information

The Mann-Whitney U test is a nonparametric test that is an alternative for the independent sample *t*-test. Generally, this test can be carried out when the assumptions for using the *t*-test are not met. The Mann-Whitney U test is used particularly in two cases—when at least one variable is ordinal or when the continuous data do not follow normal distribution. The Mann-Whitney U test assesses whether samples are drawn from the same population.

Assumptions

The following assumptions associated with the Mann-Whitney U test can be put forward:

- the measurement level of the dependent variable should be at least ordinal;
- the samples must be disjoint—there should be no relationship between the subjects in both groups, the samples should be unrelated to each other (Verma & Abdel-Salam, 2019).

Example

Dataset: The company managing sharing bicycles decided to check the impact of the station location on the use of bicycles. Two comparable high-schools were chosen. In the case of one of them (control group), the location of the station was 200 m from the entrance and in the other (test group), the station was located just in front of the entrance.

After two months of experiment, two random samples of students from each school were selected. Respondents declared the frequency of using the shared bicycles.

Data info:

- variable 1: group—nominal (1—distant location, 2—close location);
- variable 2: freq.—ordinal (declared frequency of using the shared bicycles; 1—more than once a day; 2—every day; 3—2–4 times a week; 4—once a week; 5—once a month; 6—less than once a month; 7—never).

Hypotheses:

H0: There is no difference in the frequency of using shared bicycles between the groups.

H1: The frequency of using shared bicycles differs in both groups.

Testing the assumptions

There are two unrelated groups and the frequency is measured on an ordinal scale.

In the first step, the medians are computed for both groups. In order to obtain the output, the file is split (procedure described in 1.1.) and descriptive statistics are run.

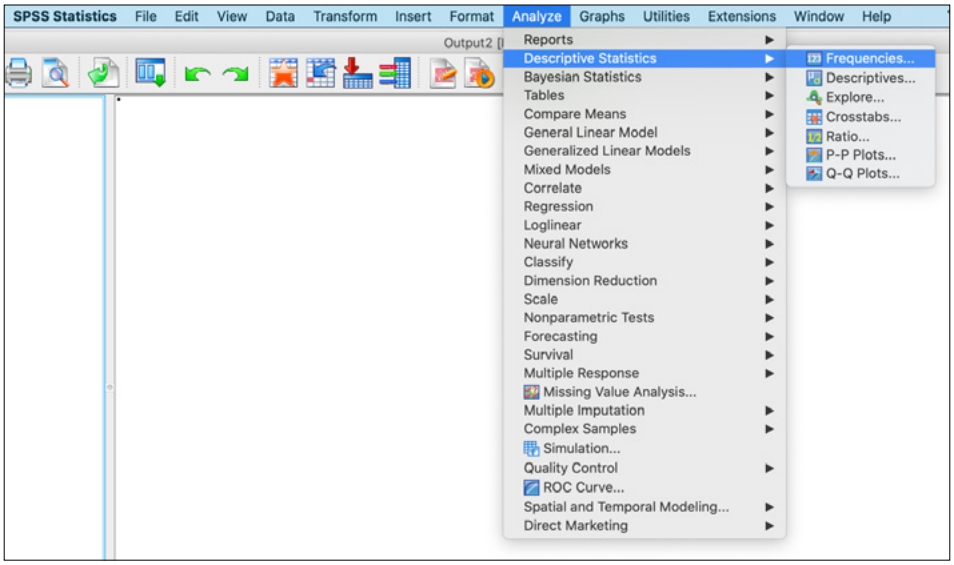


Figure 9. Descriptive statistics—path

Source: The authors' own elaboration, IBM SPSS screenshot.

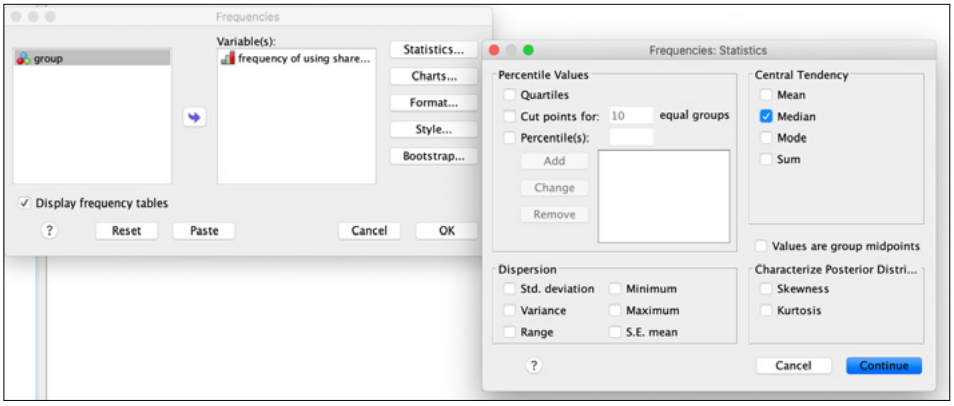


Figure 10. Descriptive statistics—dialogue box

Source: The authors' own elaboration, IBM SPSS screenshot.

Statistics

frequency of using shared bicycles

1 control	N	Valid	45
		Missing	0
	Median		5.00
2 test	N	Valid	39
		Missing	0
	Median		3.00

Figure 11. Descriptive statistics—results

Source: The authors' own elaboration, IBM SPSS screenshot.

The medians are 3 and 5 for the test and control groups, respectively. The number of observations can be seen as well. In the next step, the Mann-Whitney U test is performed. It will be compared whether the difference between groups is statistically significant. Before running the test, it must be remembered to split the groups by using the command "analyse all cases, do not create groups" in the dialogue box (procedure described in 1.1).

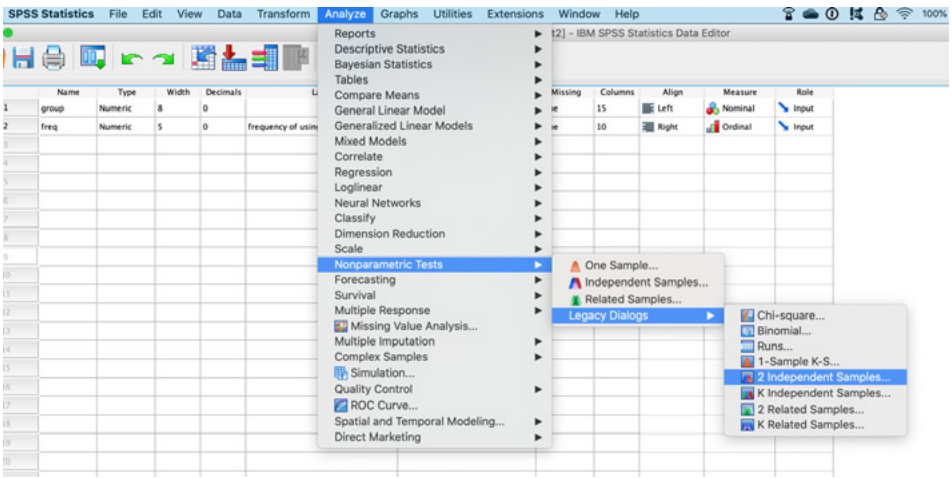


Figure 12. The Mann-Whitney U test—path

Source: The authors' own elaboration, IBM SPSS screenshot.

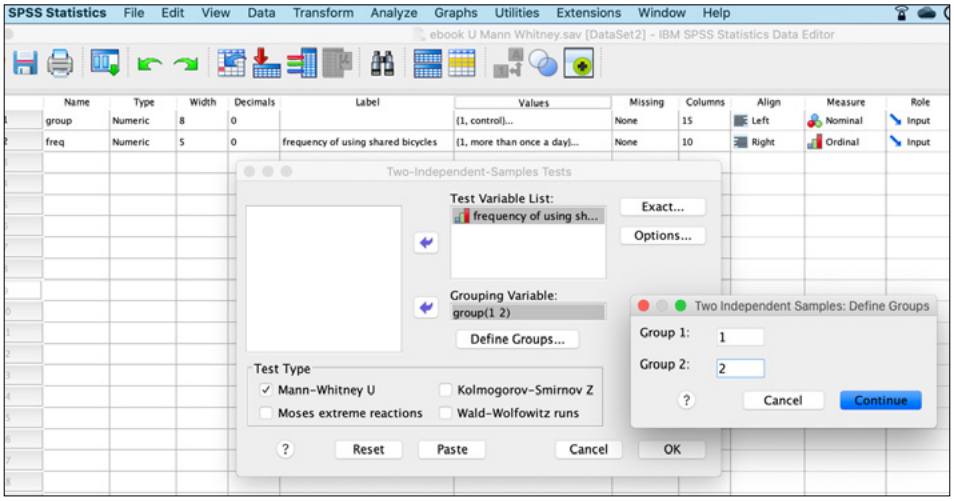


Figure 13. Mann-Whitney U test—dialogue box

Source: The authors' own elaboration, IBM SPSS screenshot.

Mann-Whitney Test

Ranks

	group	N	Mean Rank	Sum of Ranks
frequency of using shared bicycles	1 control	45	48.37	2176.50
	2 test	39	35.73	1393.50
	Total	84		

Test Statistics[a]

	frequency of using shared bicycles
Mann-Whitney U	613.500
Wilcoxon W	1393.500
Z	-2.407
Asymp. Sig. (2-tailed)	.016

a. Grouping Variable: group

Figure 14. Mann-Whitney U test—results

Source: The authors' own elaboration, IBM SPSS screenshot.

Results

The results are interpreted from the last row in the lower table (test statistics). The significance equals $p = .016$, which is lower than the critical level of $p = .05$. This indicates that there is a significant difference in frequencies of using shared bicycles between the groups.

In Figure 11 (descriptive statistics results), it can be noted that the median for the control group is five (once a month) and for the test group it totaled three (2–4 times a week).

Mann-Whitney U test hypotheses resolution:
$p < .05$—there is a significant difference between the groups; reject H0;
$p > .05$—there is no significant difference between the groups; do not reject H0.

Effect size
The effect size measure for the Mann-Whitney U test is the *r* (*do not confuse with Pearson's r*), which is calculated using the statistic *Z* value and n_1, n_2 being the total number of observations in both groups:

$$r = \frac{|Z|}{\sqrt{n_1 + n_2}}$$

The *r* has the following interpretation:
Below .1—no effect;
< .1-.3)—small effect;
< .3-.5)—moderate effect;
.5 and above—large effect (Pallant, 2011; Field, 2013).

$$r = \frac{|-2.407|}{\sqrt{84}} = .26$$

In this case, a small effect ($r = .26$) can be observed.

Summary
Dataset: The company managing sharing bicycles decided to check the impact of the station location on use of the bicycles. Two comparable high-schools were chosen. In the case of one of them (control group), the location of the station was set 200 m from the entrance and in the other (test group), the station was located just in front of the entrance.

After two months of the experiment, two random student samples from each school took part in the study. Respondents declared the frequency of using the shared bicycles.

Data info:
- variable 1: group—nominal (1—distant location, 2—close location);
- variable 2: freq.—ordinal (declared frequency of using the shared bicycles; 1—more than once a day; 2—every day; 3—2–4 times a week; 4—once a week; 5—once a month; 6—less than once a month; 7—never).

Using the bicycles in the test group was more frequent (*Mdn* = 3; once a month) than in the control (*Mdn* = 5; 2–3 times a week). The Mann-Whitney U test allows to indicate that this difference is statistically significant: $U(N_{control} = 45, N_{test} = 39) =$ $= 613.50$, $Z = -2.41$, $p = .016$.

It can be assumed that the location of the station has significant impact on the frequency of using the bicycles. This effect is considered small ($r = .26$).

References

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics* (5th ed.). Sage edge.

Pallant, J. (2011). *SPSS survival manual: a step by step guide to data analysis using SPSS* (4th ed.). Allen & Unwin.

Verma, J. P., & Abdel-Salam, G. A.-S. (2019). *Testing statistical assumptions in research*. John Wiley & Sons, Inc.

1.3. One-way analysis of variance (ANOVA)

Theoretical background

One-way analysis of variance (ANOVA) is used to determine if there is a significant difference between means of several subpopulations (groups) dependent on one factor. In ANOVA, independent variables are organised in categorical groups (Dean & Illowsky, 2013; Field, 2013; Fraser, 2016). For example, if the difference in one's average daily income in January, February, March and April is to be tested, then there will be four groups of data (according to particular month), and daily income expressed in some currency will be the dependent variable. If it is to be tested whether there is a difference in sales when merchandise is displayed in a window, in the centre of the shop or at some point behind sales person, there will be three groups: "window", "centre", "behind", and for one particular product, sales will be measured in some period according to those positions. The value of the daily sales will be the dependent variable. Also, ANOVA is useful when wanting to observe if there is a significant difference in consumer behaviour regarding various socio-demographic characteristics. In addition, ANOVA can be useful when wanting to analyse effectiveness of sales force in different locations.

One-way ANOVA is usually utilised when comparing three or more categorical independent groups to establish whether there is a statistically significant difference between them (Field, 2013; Barrow, 2017). One-way ANOVA can be used in the case of just two categorical independent groups, but in that case, the independent sample *t*-test is more frequently used. It is recommended each category (group) contain at least two units or two measurements in order to be able to calculate variance.

Hypothesis
The null hypothesis is that the means of all groups are equal, i.e. that the observed difference between groups is due to chance. On the other hand, the alternate hypothesis is that there is at least one pair of groups where the mean between them is significantly different.

Assumptions
The following assumptions can be associated with one-way ANOVA independent samples (Dean & Illowsky, 2013; Field, 2013; Randolph & Myers, 2013):
- the independent variable should consist of two or more categorical, independent groups. Typically, one-way ANOVA is used when there are three or more groups, but it can also be used for only two groups (even though the independent sample *t*-test is more commonly used in that case);
- the samples are disjoint, there is no relationship between the observations in each group or between the groups themselves. For example, one participant has to be exclusively in one group;
- the dependent variable should be measured at the interval or ratio level (i.e. they have to be continuous);
- the dependent variable should be approximately normally distributed for each category of the independent variable and there should be no significant outliers;
- homogeneity of variance is required. Therefore, it is recommended to perform Levene's test for homogeneity before application of one-way ANOVA.

Example
Dataset: quantity of food waste measured in grams per month, per person observed in four groups of consumers, according to age groups: 18–25; 26–40; 41–60; above the age of 60. Food waste as a problem is growing in the modern world. There are some studies in which it is shown that age might be the crucial factor when explaining difference in consumer behaviour regarding food waste. Thus, it is enquired whether there is a difference between generations of consumers regarding food waste on a monthly basis. Therefore, research was carried out in which the respondents were asked to assess the quantity of wasted food on a personal level within one month in grams. The survey was carried out using a random sample of 200 respondents.

Data info:
- variable 1: groups—nominal (1—age 18–25, 2—age 26–40, 3—age 41–60, 4—above the age of 60);
- variable 2: food waste quantity—numeric (grams of wasted food in grams per person in a month).

Hypotheses:
H0: There is no difference in mean food waste quantities between the groups.
H1: There is at least one group at which mean food waste quantity is different than in the other groups.

Testing the hypotheses in SPSS

In this example, 200 respondents were studied, and for each respondent, two types of data were collected: (1) age, (2) level of food waste in grams per month, per person. Three questionnaires (observations) were not valid, thus the dataset was based on 197 valid questionnaires (or observations).

SPSS does not require grouping the collected survey data, but data is entered as an observation per row. In Figure 15, in the first row—in the "Generation" column, data on the generation of the respondent is entered and in column "Foodwastegr" data on food waste for this respondent is entered, therefore, in row 40, it can be observed that the respondent's age is 18–25 (generation group numbered as 1) and respondent wastes 195 grams of food per month (see Figure 15).

Prior to analysis, the type of loaded data has to be checked, and it is recommended that numerical denomination of categories is used for the independent variable (Field, 2013). That means instead of the text "Group 1 (18–25)", 1 should be used to denominate this particular generation of consumers, similarly—"Group 2 (26-40)" will be coded as 2, etc. It is important to emphasize that with introducing numeric codes for the variable does not strenghten its measurement level. It is still categorical (in our case: ordinal). In Figure 15, see column = "Generation".

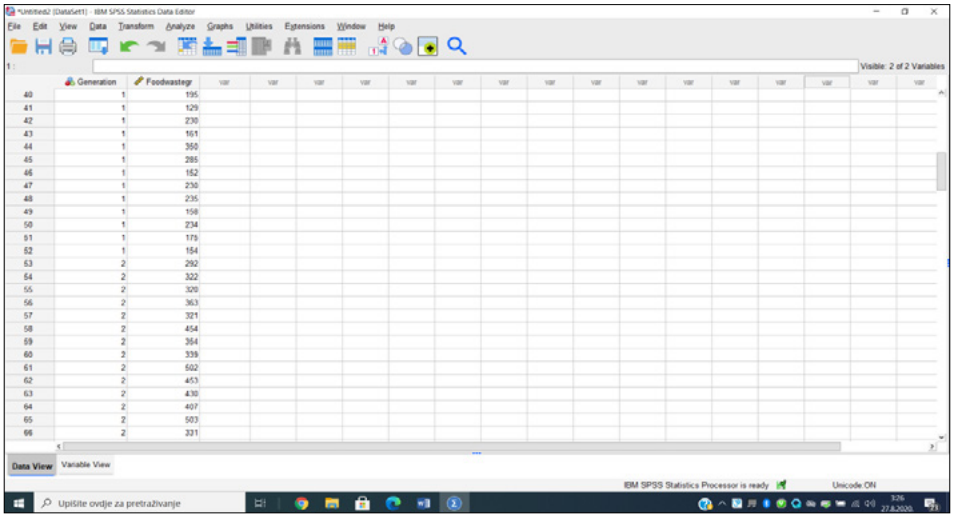


Figure 15. Excerpt from dataset in SPSS

Source: The authors' own elaboration.

Before conducting ANOVA, it is recommended to calculate means of groups (Note: This step can be skipped, because the latter in the one-way ANOVA procedure, the option to display descriptive statistics can be chosen, which will show the summarised descriptive statistics data for each group).

In Figure 16, the screenshot shows the command for calculation of means, while in Figure 17, it is demonstrated how to set options in order to calculate means for various age groups (generations) of consumers based on the dependent variable "Food waste in grams" from the dataset.

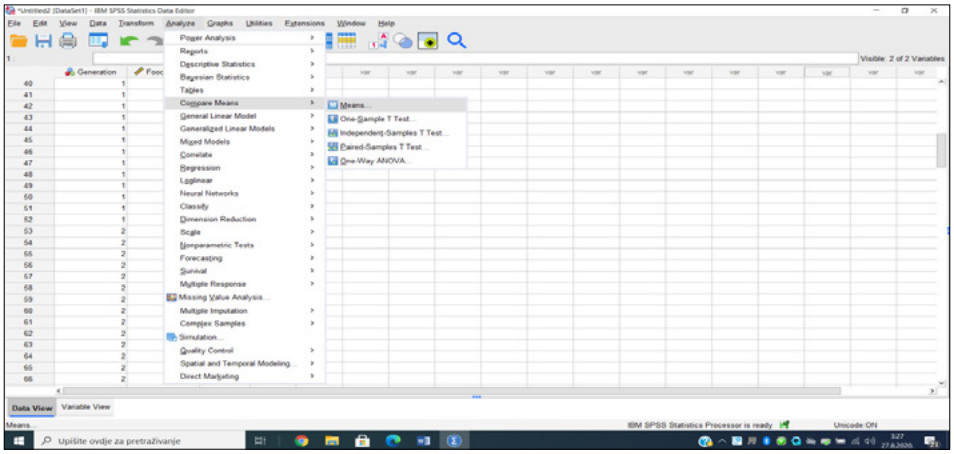


Figure 16. SPSS Command to calculate means

Source: The authors' own elaboration.

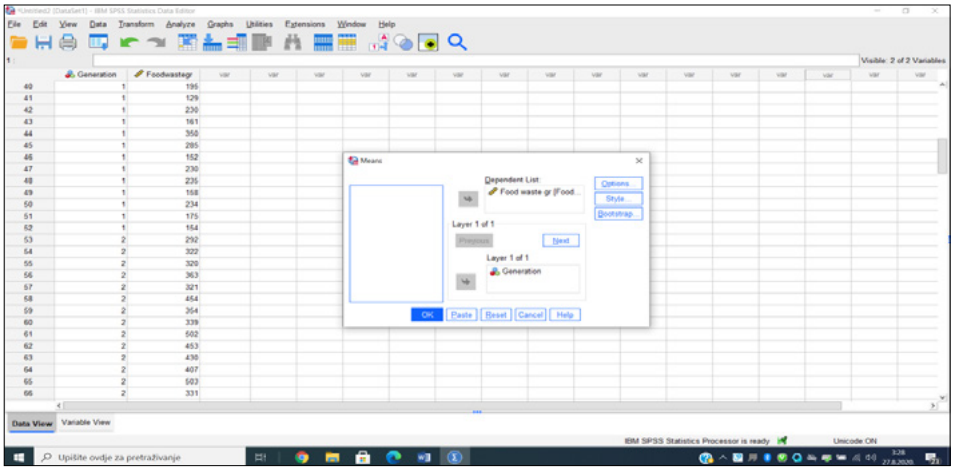


Figure 17. Setting variables to calculate means for groups in the dataset

Source: The authors' own elaboration.

Output of means calculation is given in Figure 18. The first part of the output is the summary based on the total sample from which it can be read how many cases from the dataset are included or excluded from means calculation. In this case, all observations were correct, therefore, all data is included in the calculation of means. The second part of the report are means, number of cases (observations) and standard errors for each group in the sample, i.e. for each generation of consumers. For instance, for generation 4 or "Group 4 (60+)", it can be observed that average monthly food waste per person is 290.52 grams, the result based on 50 cases (observations) with the standard deviation of 89.30. Compared to the total sample, this generation has a lower average of food waste. Namely, the average monthly food waste, taking all 197 respondents into account, is 376.85 grams per person.

Case Processing Summary						
	Cases					
	Included		Excluded		Total	
	N	Percentage	N	Percentage	N	Percentage
Food waste (gr) * Generation	197	100.0%	0	0.0%	197	100.0%

Report			
Food waste (gr)			
Generation	Mean	N	Std. Deviation
1	261.88	52	80.236
2	363.18	50	99.787
3	620.82	45	12.660
4	290.52	50	89.300
Total	376.85	197	169.944

Figure 18. SPSS means report

Source: The authors' own elaboration.

After that, the procedure for one-way ANOVA will be started. Selection of SPSS required command is shown is Figure 19, while in Figure 20, the dialogue for tuning up settings in the presented example is given.

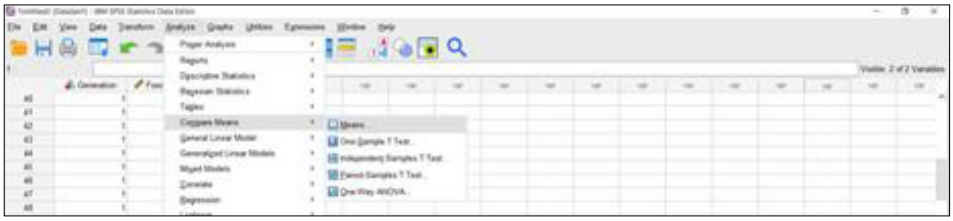


Figure 19. SPSS Command for one-way ANOVA

Source: Authors' own elaboration.

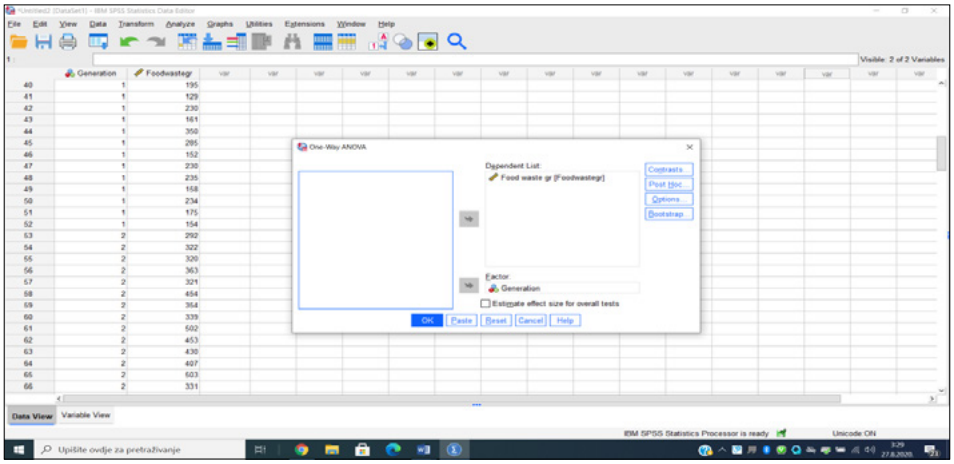


Figure 20. Settings of one-way ANOVA for groups in the dataset

Source: The authors' own elaboration.

Together with basic settings, SPSS can be set to perform post hoc analysis in the same run. Therefore, the 'Post Hoc' button should be clicked, and the dialogue box shown at Figure 21 will appear. Usually, it is enough to do Tukey's post hoc analysis at the confidence level of .05 (necessary settings are shown in Figure 21). When everything is set up, the analysis will be run.

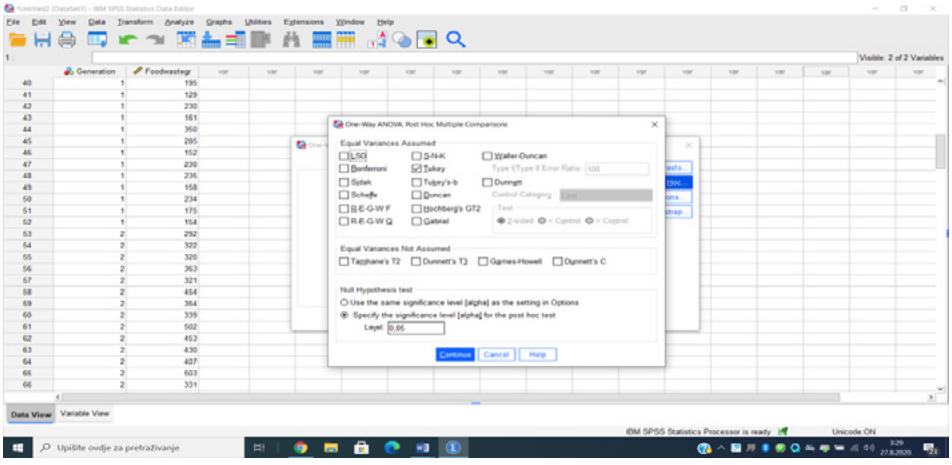


Figure 21. Settings of Tukey's post hoc analysis

Source: The authors' own elaboration.

In Figure 22, the output of one-way ANOVA is shown. It has to be emphasized that the significance value on the output is written as equal to 0.000 (p = .000), but

that does not mean that the significance value equals zero. That is just the way SPSS tells us that the significance value is below .001. Thus, in accordance with the output, it can be concluded that the significance value is very small and, for sure, lower than .05. Therefore, at the significance level of .05, the null hypothesis of the test that there is no difference in mean food waste quantities between the groups can be rejected. So, it can furthermore be concluded that there is at least one group in which mean food waste quantity is different than in the other age groups (generations).

Post hoc analysis provides scrutinized insight into differences between pairs of groups. As a result, the significance value (see Sig. column in Post hoc analysis) can be observed for each age group compared to other age groups. In the presented example, it can be seen that all significance values are less than .05, except for the value use to compare age groups (generation) 1 and 4 (p = .469). Therefore, for instance, it can be assumed that the average quantity of food waste per month, per person from generation 1 is statistically different compared to generations 2 and 3, respectively. However, at the significance level of .05, the hypothesis cannot be rejected that there is no statistically significant difference between generations 1 and 4 regarding the average quantity of food waste per month, per person.

ANOVA
Food waste gr

	Sum of squares	df	Mean square	F	Sig.
Between groups	3747782.985	3	1249260.995	126.044	.000
Within groups	1912881.745	193	9911.304		
Total	5660664.731	196			

Post hoc tests

Multiple comparisons
Dependent variable: Food waste gr
Tukey HSD

(I) Generation	(J) Generation	Mean difference (I -J)	Std. Error	Sig.	95% Confidence interval	
					Lower bound	Upper bound
1	2	−101.295*	19.719	.000	−152.40	−50.19
	3	−358.938*	20.270	.000	−411.47	−306.41
	4	−28.635	19.719	.469	−79.74	22.47
2	1	101.295*	19.719	.000	50.19	152.40
	3	−257.642*	20.457	.000	−310.66	−204.63
	4	72.660*	19.911	.002	21.06	124.26
3	1	358.938*	20.270	.000	306.41	411.47
	2	257.642*	20.457	.000	204.63	310.66
	4	330.302*	20.457	.000	277.29	383.32
4	1	28.635	19.719	.469	−22.47	79.74
	2	−72.660*	19.911	.002	−124.26	−21.06
	3	−330.302*	20.457	.000	−383.32	−277.29

Figure 22. Output of one-way ANOVA in SPSS

Source: The authors' own elaboration.

Testing hypotheses in Excel

In order to perform analysis of the same dataset in Excel, collected data has to be prepared for analysis, i.e. collected data has to be classified into columns that represent groups (Balakirshnan, Render, & Stair, 2007; Winston, 2016; Fraser, 2016). In our case columns will represent groups by age—generations of consumers. Therefore, in this case, the collected data will be classified into four columns and each column will be labelled according to consumer generation (in Figure 23, see title of columns in row 3). Then, all observed values will be entered for each generation of consumers. For instance, if a certain respondent is from generation 2 (age 26-40) and wastes 407 grams of food per month, his/her data is entered into the second column—'Group 2 (26-40)' (in Figure 23, see row 15). In the SPSS dataset, data on this respondent was entered as a simple observation in a single row as 2 and 407 (see Figure 15, row 64).

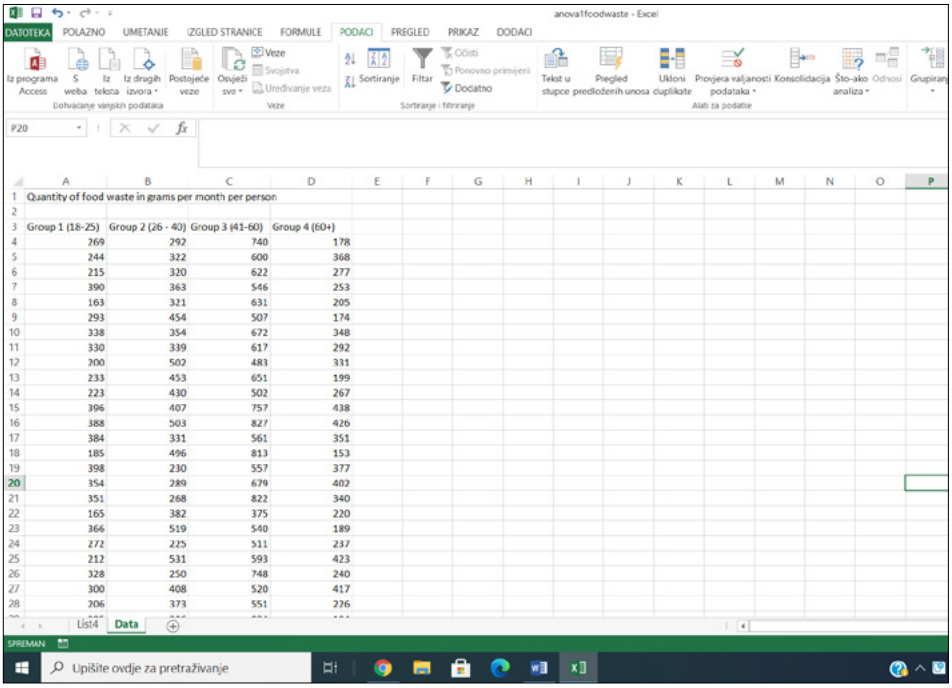


Figure 23. Excerpt from dataset for one-way ANOVA of food waste according to age

Source: The authors' own elaboration.

Then, 'Data tab' has to be selected and 'Data Analysis' (within Analysis group of commands) clicked. (Note that Data Analysis pack is not defalult package, you have to install it in your Excel). From among the list of methods, 'Anova: Single Factor' is chosen (see Figure 24).

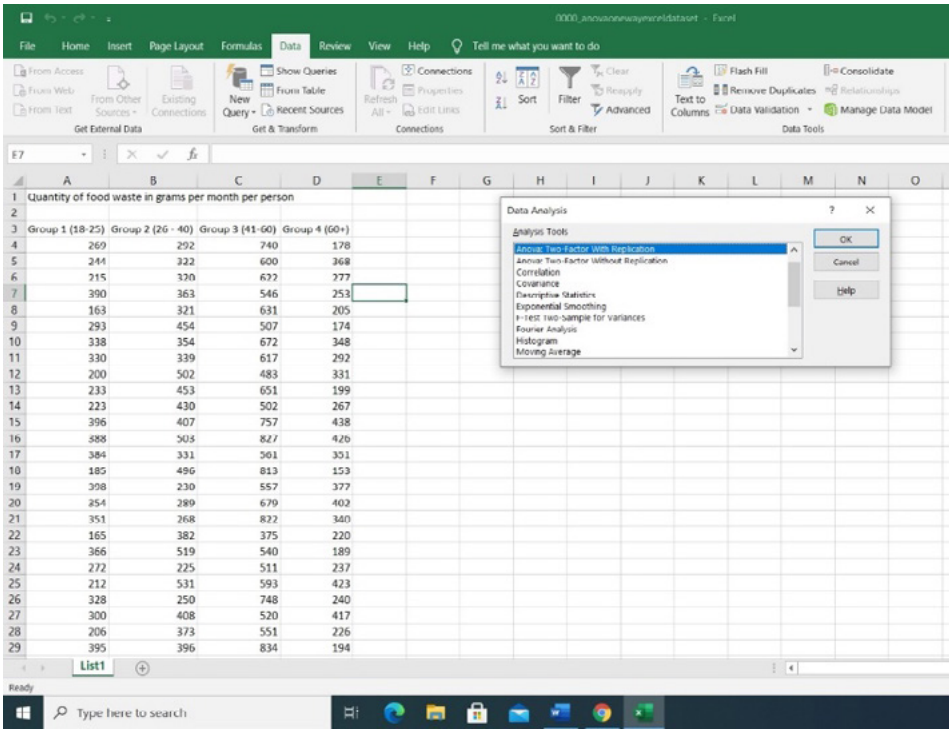


Figure 24. Data analysis tab in Excel—selection of ANOVA method: Single Factor

Source: Authors' own elaboration.

In the dialogue box of Anova: Single Factor—configuration has to be carried out as follows (see Figure 25):

- input range of the dataset including labels, in this example—A3:D55;
- position of data labels, in this example—First Row (there are names of the observed groups);
- way of organising groups of data, in this case, data is organised in columns, therefore, 'Columns' is chosen;
- output range—data can be choosen to be shown at some position in the active worksheet. Then, the exact cell, from which our results are going to be presented (such as F3), has to be specified; but in this case, we rather specified 'New worksheet' was indicated as the location for results. A name for the output can be specified (in this example—'Anova1');
- finally, the level of significance, i.e. alpha value. The default value, already set to .05, can be used.

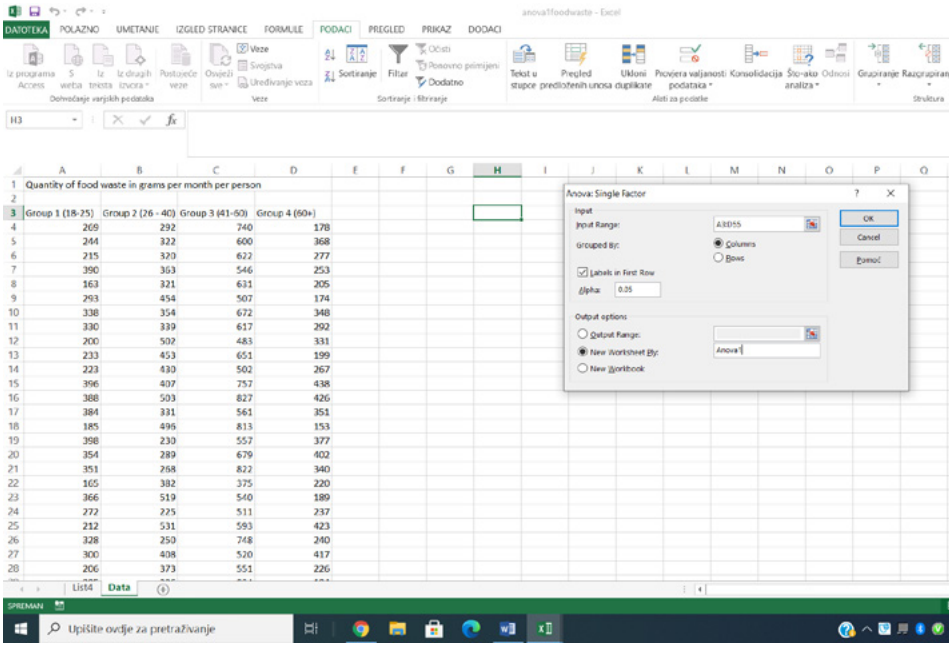
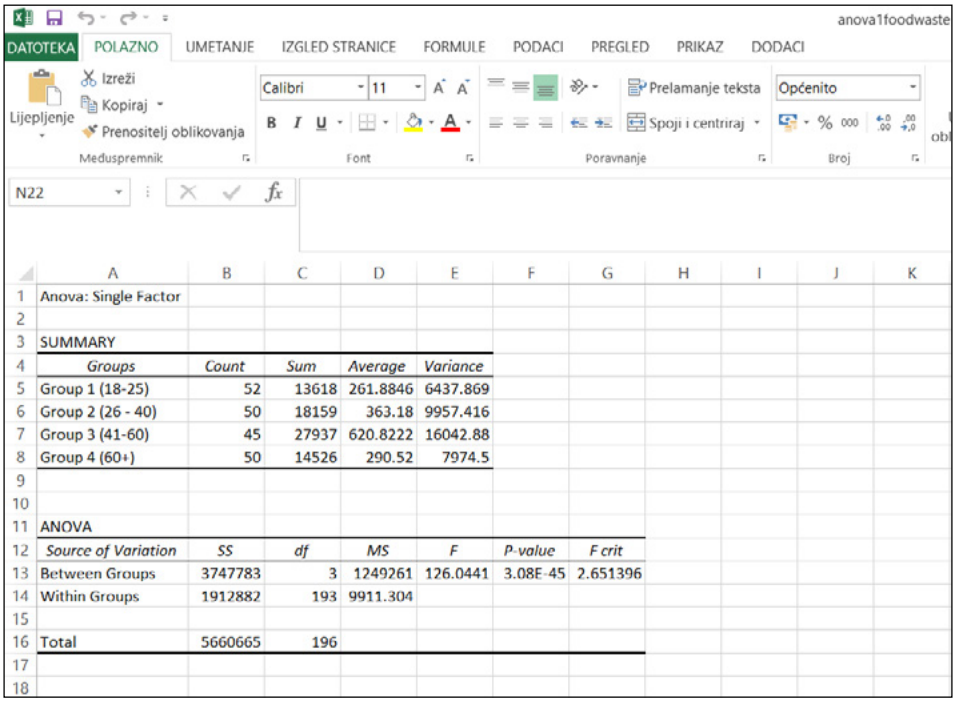


Figure 25. Dialog box—ANOVA: Single Factor

Source: The authors' own elaboration.

In Figure 26, the results of data analysis are shown, and the results can be interpreted. First of all, basic descriptive statistical data on each age group is obtained (see SUMMARY). From this part, it can be read how many respondents are in which group, then, what the average food waste in each group is, as well as the variance within each group. For instance, the lowest average of 261.88 grams of food waste per person, per month is shown in 'Group 1' (aged 18–25). The highest average value is in 'Group 3' (aged 41–60) and amounts to 620.82 grams a month, per person. In addition, ANOVA results are shown. In this table, the most important reading is *p*-value, because using this value, it can be decided not to reject or to reject the null hypothesis. In this case, the *p*-value is $3.08*10^{-45}$, or if rounded and truncated to four decimal points, the *p*-value is: .0000. However, the more precise would be if it were said that the *p*-value is lower than .0001 (*p*-value <.0001). In this way, it can be concluded that the significance value is much lower than that of .05. Consequently, that result means that the null hypothesis H_0 can be rejected and that there is no difference in mean food waste quantities between groups. In other words, at a significance level of .05, it may be concluded that there is at least one group in which mean food waste quantity is statistically different than in the other age groups (generations).



	A	B	C	D	E	F	G	H	I	J	K
1	Anova: Single Factor										
2											
3	SUMMARY										
4	*Groups*	*Count*	*Sum*	*Average*	*Variance*						
5	Group 1 (18-25)	52	13618	261.8846	6437.869						
6	Group 2 (26 - 40)	50	18159	363.18	9957.416						
7	Group 3 (41-60)	45	27937	620.8222	16042.88						
8	Group 4 (60+)	50	14526	290.52	7974.5						
9											
10											
11	ANOVA										
12	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*				
13	Between Groups	3747783	3	1249261	126.0441	3.08E-45	2.651396				
14	Within Groups	1912882	193	9911.304							
15											
16	Total	5660665	196								
17											
18											

Figure 26. One-way ANOVA results

Source: The authors' own elaboration.

If ANOVA shows that there is a statistically significant difference between observed groups, post hoc analysis has to be carried out by comparing pairs of groups in order to explain which groups differ in comparison to the other groups. For this purpose, several *t*-tests can be performed in Excel.

In the presented example, the *t*-test will be performed between Group 1 and Group 2 as an example. This kind of comparison is then done to compare Group 1 to Group 3, Group 1 to Group 4, Group 2 to Group 3 and Group 2 to Group 4. The *t*-tests have to be repeated accordingly to investigate differences between all possible pairs of groups in the dataset.

Steps for performing the *t*-test in Excel are the following: first, it must be specified which type of *t*-test it to be performed. This is done via the 'Data analysis' tab (see Figure 27).

During this step, the *t*-test: 'Two Samples Assuming Equal Variance' is chosen.

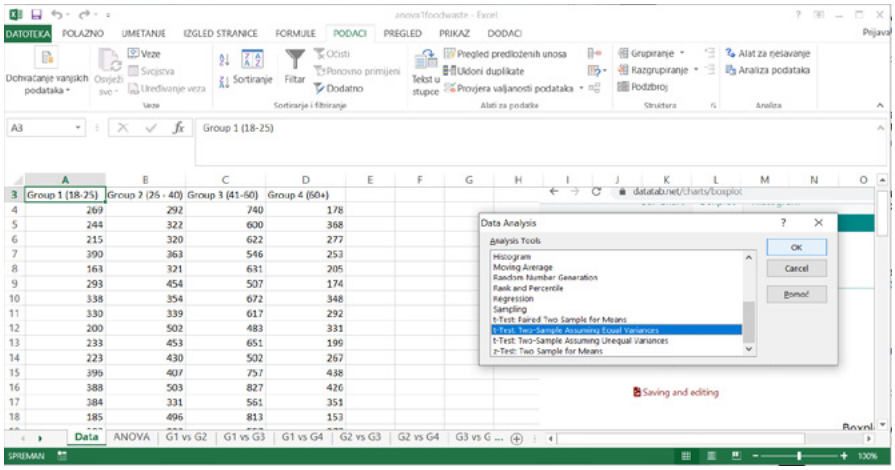


Figure 27. Data analysis tab in Excel—selection of *t*-test type

Source: The authors' own elaboration.

Then, in the *t*-test dialogue box (see Figure 28), it has to be specified which data is to be compared. The first pair of data comprises Group 1 (18–25) and Group 2 (26–40). Therefore, the range of data for Group 1 in 'Variable 1 Range' is specified, and the same is done for 'Variable 2 Range', giving the range of data from Group 2. Moreover, the data has data labels in the first row of selected data range, thus, 'Labels' have to be checked. Finally, the location for the output or results are specified. In this case, it was decided to have a new worksheet named 'G1 vs. G2'.

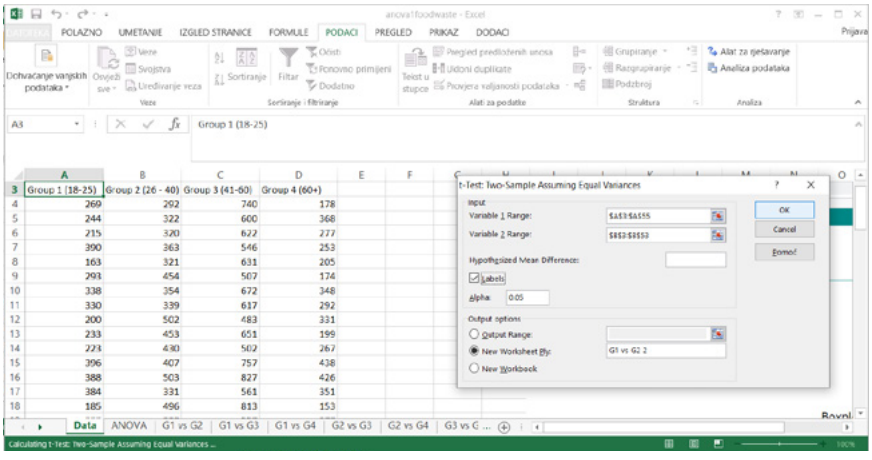


Figure 28. Dialogue box for *t*-test: Two Sample Assuming Equal Variances

Source: The authors' own elaboration.

After clicking 'OK', the output of the *t*-test is shown (see Figure 29) and interpretation can be carried out on the basis of the analysed pair of variables (in this test—Group 1 and Group 2).

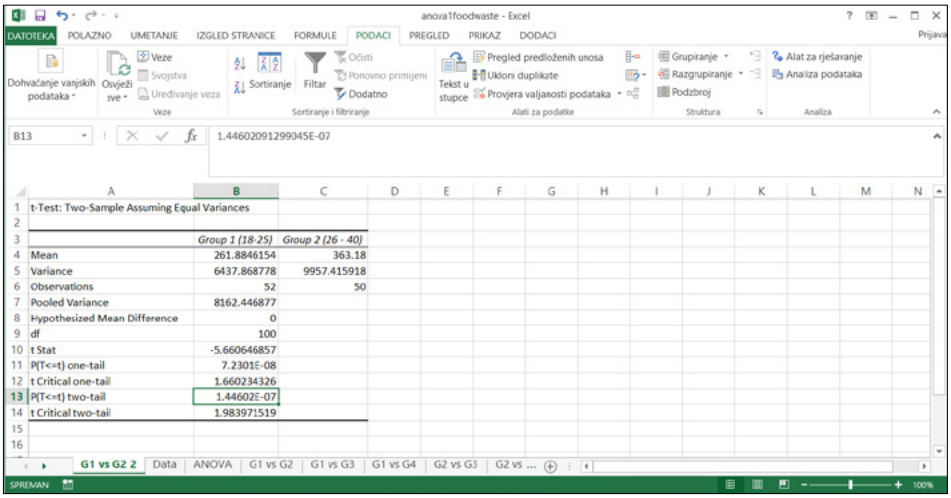


Figure 29. *T*-**test results for pair of groups**

Source: The authors' own elaboration.

If the post hoc *t*-test results are to be interpreted, the *p*-value for two-tail comparison is used (see Figure 21). Based on the level of significance of .05, it can be concluded that the given *p*-value (in this case: $1.4607*10^{-7}$ or truncated to four decimal digits: 0.0000), is lower than .05 and that there is a statistically significant difference in means between Group 1 (18–25) and Group 2 (26–40). However, by doing so, an erroneous conclusion could be drawn. Therefore, as suggested in literature on the subject, before final conclusions, the significance level of .05 has to be adjusted according to number of groups of data in the ANOVA. As in this case there are 4 groups of data (according to the age of respondents), the relevant value for comparison would be .05 divided by 4, i.e. .0125. Thus, in order to carry out the correct interpretation and reach adequate conclusions, the given *p*-value of the *t*-test has to be compared for each pair of groups in the dataset to .0125, not to .05! In this case, .0000 is lower than .0125, and it may be concluded that there is a statistically significant difference between the means for Group 1 (18–25) and Group 2 (26–40).

After this, the *t*-test is iteratively repeated for all pairs of data in a similar way. In Table 1, the totalled *t*-test *p*-values relevant for each pair of groups is shown The *p*-value is given in a default calculated format (scientific) and then in numeric

format truncated to 4 decimal digits. From the table, it may be concluded that at the level of .05, there is a statistically significant difference between all observed groups, except for Group 1 (18-25) and Group 4 (60+). For that pair of groups, the calculated *p*-value is higher than the adjusted significance level of .0125. Therefore, we cannot reject the hypothesis that there is no difference between the means of those groups.

Table 1. ***T*-test relevant *p*-values for given example**

Pair of groups		*t*-test *p*-value (scientific)	*t*-test *p*-value (numeric)	Decision (according to adjusted significance 0.0125)
Group 1 (18-25)	*Group 2 (26 - 40)*	1.44602E-07	0.0000	reject H0 (there is difference)
Group 1 (18-25)	*Group 3 (41 - 60)*	2.23E-30	0.0000	reject H0 (there is difference)
Group 1 (18-25)	*Group 4 (60+)*	0.091316825	0.0913	not reject H0 (there is no difference)
Group 2 (26 - 40)	*Group 3 (41-60)*	1.16784E-18	0.0000	reject H0 (there is difference)
Group 2 (26 - 40)	*Group 4 (60+)*	0.000220728	0.0002	reject H0 (there is difference)
Group 3 (41-60)	*Group 4 (60+)*	3.46163E-26	0.0000	reject H0 (there is difference)

Source: The authors' own elaboration.

It must be borne in mind that such post hoc analysis is performed only in the case when ANOVA indicates that there is a difference between means of several groups of data in the dataset in order to interpret data more accurately and precisely (Fraser, 2016; Winston, 2016).

Summary of the example

Dataset: the food waste quantity in city A is inspected. In the conducted survey, a total of 200 respondents participated. However, three questionnaires have been declared invalid. Consequently, in the analysis, 197 data units about monthly food waste quantity of the respondents are used. In order to get better insight into monthly food waste quantity, the respondents have been divided into four categories according to age.

Data info:

- variable 1: groups—nominal (1—age 18–25, 2—age 26–40, 3—age 41–60, 4—above the age of 60);
- variable 2: food waste quantity—numeric (wasted food in grams per person, per month).

The one-way ANOVA approach was used to inspect whether the average monthly food waste quantity can be considered the same across all four age groups. However, the results of one-way ANOVA have shown that there was a statistically significant difference between age groups ($F(3,193) = 126.044$, $p < .001$). Tukey's post hoc test revealed that the average monthly food waste quantity for people aged 18–25 was

statistically significantly lower than the average monthly food waste quantity for those aged 26–40 ($p < .001$), while the average monthly food waste quantity for individuals aged 41–60 ($p < .001$). However, there was no statistically significant difference in the average monthly food waste quantity for people aged 18–25 or for those above the age of 60 ($p = .469$).

More info about one-way ANOVA

One-way ANOVA is used to inspect whether there are any statistically significant differences between the means of two or more independent groups. Despite the fact that one-way ANOVA can be used for comparing means between two independent groups, it is more often applied in cases where there are three or more independent groups, whereas in the cases of two independent groups, the *t*-test for independent samples is applied.

In order for one-way ANOVA to be used, six assumptions have to be fulfilled. Three of them can be checked without any computer software use: independent variable should consist of two or more categorical independent groups; independence of observations; dependent variable should be measured at the interval or ratio level. Those assumptions are straightforward and they can be verified very quickly. The other three assumptions should be checked using a computer program.

The fourth assumption is that dependent variable should be approximately normally distributed. The normality of data can be tested, for example, by use of the Shapiro-Wilk test for normality of distribution, and Kolmogorov-Smirnov test. The normality of data can be inspected graphically as well by using, for example, the normal Q-Q plot. In case of not normal distribution, the data should be converted into that normal by applying certain techniques. Technically spoken, there should be at least two units in each group to apply one-way ANOVA. However, the more units there are in each group, the larger the sample size. Consequently, it is more likely that the normality assumption will be fulfilled.

Because outliers have huge impact on the mean values, their presence has certain influence on the results of one-way ANOVA. Therefore, outlier analysis should be performed before conducting one-way ANOVA. The most straightforward approach to detect outliers is to standardise all values and then to check whether any of them deviate from the mean value more than three standard deviations. The outliers can be detected by using different graphical approaches as well. It has to be emphasized that outliers may have different sources. They can appear due to certain characteristics of the observed unit, but can also be the product of technical error (for example, data is mistyped).

The final assumption of one-way ANOVA application is homogeneity of variance between the groups. This assumption can be checked by Levene's test for homogeneity of variance. The null hypothesis of the test contains the assumption

that the observed groups all have equal population variances. In the given case this assumption is not met, thus, Welch's ANOVA should be used instead of this classic one-way ANOVA approach.

References

Balakirshnan, N., Render, B., & Stair, R. M. (2007). *Managerial decision modeling with spreadsheets*. Pearson Prentice Hall.
Barrow, M. (2017). *Statistics for economics, accounting and business studies*. Pearson.
Dean, S., & Illowsky, B. (2013). *Introductory statistics*. OpenStax College.
Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics* (5th ed.). Sage edge.
Fraser, C. (2016). *Business statistics for competitive advantage with Excel 2016*. Springer.
Randolph, K. A., & Myers, L. L. (2013). *Basic statistics in multivariate analysis*. Oxford: University Press.
Winston, W. L. (2016). *Microsoft Excel 2016: Data analysis and business modelling*. Microsoft Press.

1.4. Kruskal-Wallis H test

General information

The Kruskal-Wallis H test is a commonly used nonparametric alternative to one-way ANOVA. It can be used when one-way ANOVA assumptions are violated—for example, when the dependent variable is measured on an ordinal scale. The test is similar to the Mann-Whitney U test, but it is used to compare scores in three or more groups. Since the Kruskal-Wallis H test does not require normality of data distribution, it does not allow comparison of means but ranks. The procedure includes ordering the observations from lowest to highest, and giving them ranks (Pallant, 2011; Verma & Abdel-Salam, 2019).

Hypotheses:
H0: There is no difference between the scores.
H1: There is at least one difference between the scores.

Assumptions

The following assumptions are associated with the Kruskal-Wallis H test:
- the measurement level of the dependent variable should be at least ordinal;
- there should be one independent variable divided into three or more groups;
- groups do not have common elements.

Example

Dataset: The company managing sharing bicycles decided to check the impact of the station location on the use of the bicycles. Three comparable high-schools were cho-

sen and for each of them, a different proximity of the station was set. The first school had a distant location, 200 m from the entrance; the second school had a middle location (100 m); while the third had the station set exactly in front of the entrance.

After two months of experiment, three random samples of students from each school have been selected (39, 44 and 45 students). Respondents declared the frequency of using the shared bicycles.

Data info:

- variable 1: group—nominal (1—close location (N = 39), 2—middle location (N = 44), 3—distant location (N = 45));
- variable 2: freq.—ordinal (declared frequency of using the shared bicycles; 1—more than once a day; 2—every day; 3—2–4 times a week; 4—once a week; 5—once a month; 6—less than once a month; 7—never).

Hypotheses:

H0: There is no difference in the frequency of using shared bicycles between the groups.

H1: The frequency of using shared bicycles differs among the groups, at least one group is different from the other.

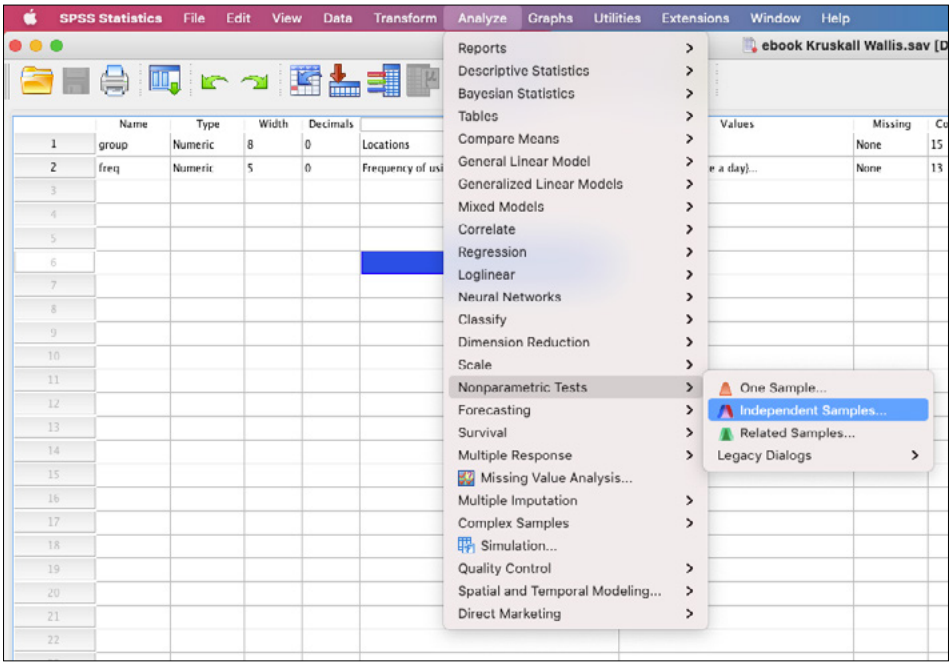


Figure 30. Kruskal-Wallis H test—path

Source: The authors' own elaboration, IBM SPSS screenshot.

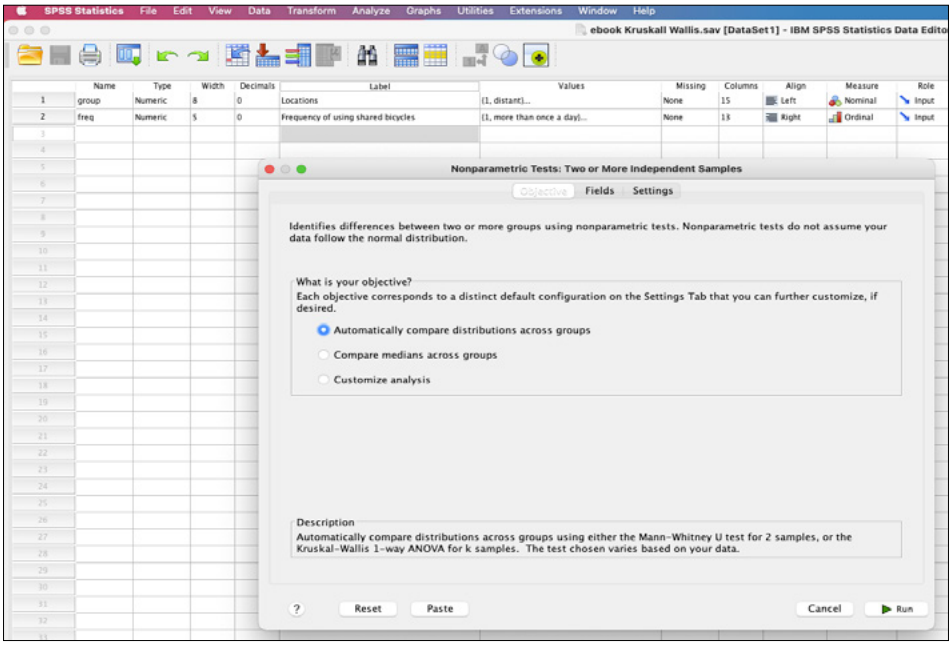


Figure 31. Kruskal-Wallis H test—dialogue box (1)

Source: The authors' own elaboration, IBM SPSS screenshot.

In the first dialogue box, three tabs can be seen—'Objectives', 'Fields' and 'Settings'. The objective of the analysis is defined by choosing the default option—'Automatically compare distributions across groups'.

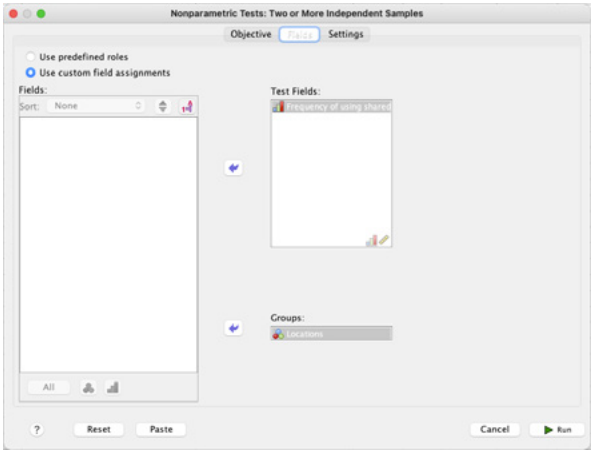


Figure 32. Kruskal-Wallis H test—dialogue box (2)

Source: The authors' own elaboration, IBM SPSS screenshot.

In the next step, we move to the tab 'Fields' where the analysed variable ('Test fields') and grouping variable ('Groups') are chosen.

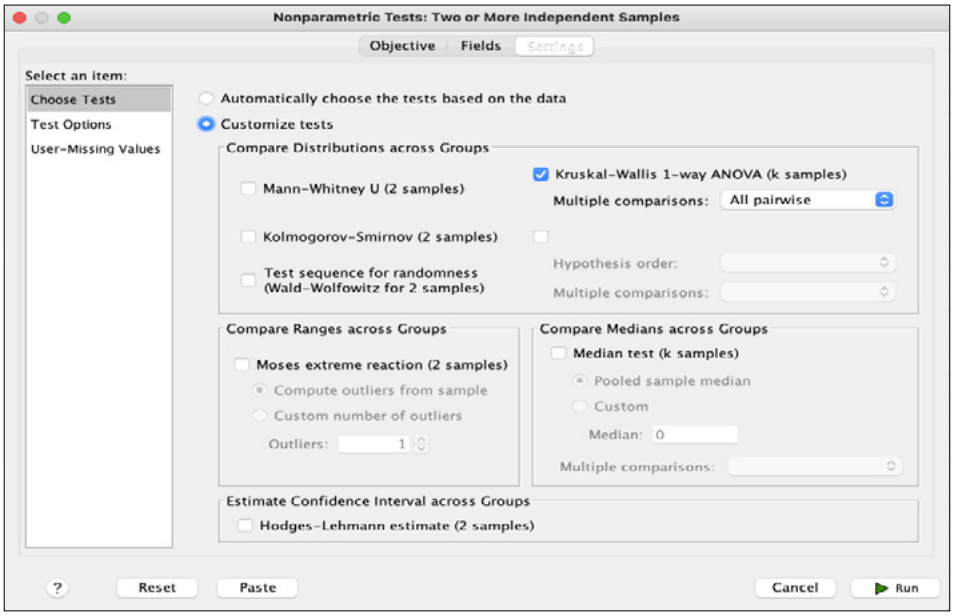


Figure 33. Kruskal-Wallis H test—dialogue box (3)

Source: The authors' own elaboration, IBM SPSS screenshot.

In the last step, we choose 'Customize tests' and select 'Kruskal-Wallis 1-way ANOVA' (k samples) with multiple comparisons: 'All pairwise'.

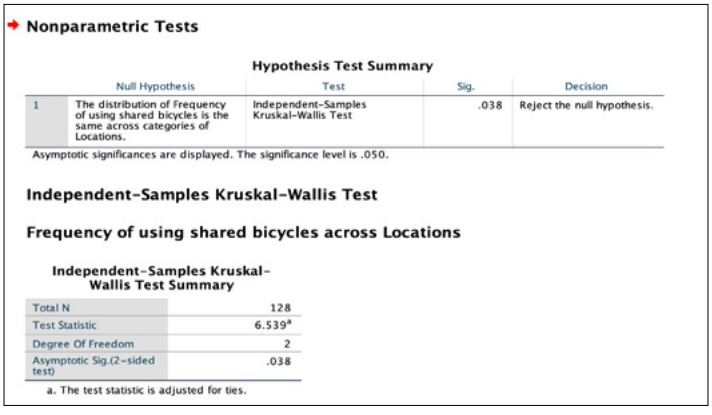


Figure 34. Kruskal-Wallis H test—results

Source: The authors' own elaboration, IBM SPSS screenshot.

Results

The hypothesis is decided upon by interpreting the 'Asymptotic Sig. (2-sided test)' value in the lower table. In this case, it equals $p = .038$. This value is lower than the critical value of $p = .05$, which indicates that there is at least one significant difference in scores across the various groups. The first dialogue box presents only a general result of the Kruskal-Wallis H test—which of the groups is significantly different from the other ones is still unknown.

Pairwise Comparisons of Locations

Sample 1–Sample 2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.[a]
1.00 close–2.00 middle	-13.939	8.037	-1.734	.083	.249
1.00 close–3.00 distant	-20.103	7.995	-2.514	.012	.036
2.00 middle–3.00 distant	-6.165	7.748	-.796	.426	1.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2–sided tests) are displayed. The significance level is .05.
a. Significance values have been adjusted by the Bonferroni correction for multiple ...

Figure 35. Kruskal-Wallis H test—pairwise comparisons (1)

Source: The authors' own elaboration, IBM SPSS screenshot.

In order to identify the differences between the groups, pairwise comparisons are examined. 'Adj. Sig.' value for the last column is interpreted. In the presented example, the *p*-value is lower than the critical value of $p = .05$ when comparing only the close and distant locations ($p = .036$). This means that there is a significant difference in the frequency of using shared bikes between these groups. The *p*-values for other comparisons: $p = .249$ and $p = 1.000$, mean that there is no significant difference in the frequency of using bikes.

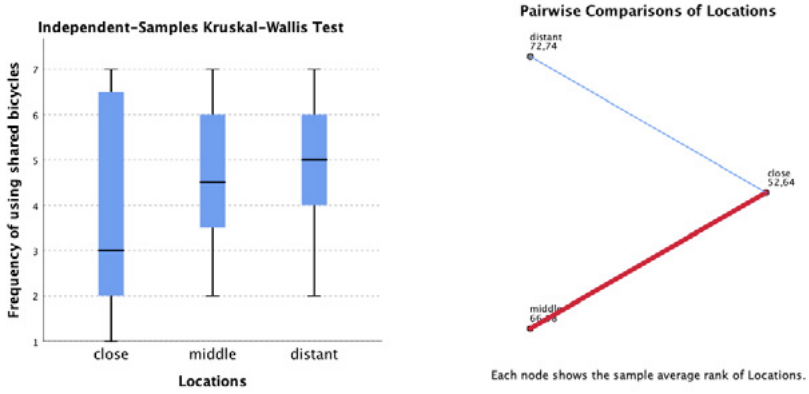


Figure 36. Kruskal-Wallis H test—pairwise comparisons (2)

Source: The authors' own elaboration, IBM SPSS screenshot.

The results can be compared by looking at the box-and-whiskers graph and at the visual representation of pairwise comparisons. The blue line in the graph on the right indicates a significant difference in the frequency of using shared bikes between close and distant locations. The red line indicates an insignificant difference.

Results and post hoc tests

Kruskal-Wallis H test hypotheses resolution:
$p < .05$—there is at least one significant difference in scores across different groups; reject H0;
$p > .05$—there is no significant difference in scores across different groups; do not reject H0.

Effect size

The effect size measure for Kruskal-Wallis H test is calculated following the procedure for the Mann-Whitney U test (Pallant, 2011).

The effect size measure (r) is based on the statistic Z and N values which is total number of observations in both groups:

$$r = \frac{|Z|}{\sqrt{N}}$$

The effect size can only be calculated for significant differences between groups. The Z value for each comparison is expressed as 'Std. Test Statistic' in the 'Pairwise Comparisons of Locations' table.

The r has the following interpretation:
Below .1—no effect;
< .1-.3)—small effect;
< .3-.5)—moderate effect;
.5 and more—large effect.

$$r = \frac{|-2.514|}{\sqrt{84}} = 0.27$$

In this case, a small effect size (r = .27) can be observed.

Summary

Dataset: The company managing sharing bicycles decided to check the impact of the station location on the use of the bicycles. Three comparable high-schools were chosen, and for each of them a different proximity of the station was set. The first school had a distant location, 200 m from the entrance, the second one had moderate location (100 m), while the third school was set exactly in front of the entrance.

After two months of the experiment, three random samples of students from each school were selected (39, 44 and 45 students). Respondents declared the frequency of using the shared bicycles.

Data info:
- variable 1: group—nominal (1—close location (N = 39), 2—middle location (N = 44), 3—distant location (N = 45));
- variable 2: freq.—ordinal (declared frequency of using the shared bicycles; 1—more than once a day; 2—every day; 3—2–4 times a week; 4—once a week; 5—once a month; 6—less than once a month; 7—never)

The Kruskal-Wallis H test allowed to reveal that the frequency of using shared bikes differed statistically significantly across different locations. Pairwise comparisons indicated that there is a difference in the frequency of using shared bikes between the students from school with close location of the station and with distant location of the station (*Gc*, $n = 39$, *Gd*, $n = 45$, $Z = -2.514$; $p = .036$). Students from schools close to the station used bikes more often ($Mdn = 3$) than students from those with distant locations ($Mdn = 5$). However, this effect was rather small ($r = .027$). The analysis did not show any significant differences between other groups.

More information

The result of Kruskal-Wallis H test does not inform us about the between-group comparisons. In order to compare separate groups pairwise, Bonferroni adjustment needs to be applied. This involves multiplying the significance by the number of tests (significance level equal to $p = .012$ after multiplication is shown as adjusted significance (('Adj. Sig.') $p = .036$). The same result may be obtained by dividing the alpha level of .05 by the number of tests that are intended to be used, and by implementing the initial significance level ('Sig.'). While interpreting the group comparisons the revised alpha level should be used as the criteria for determining significance. However, the described procedure shows the results applying Bonferroni adjustment.

References

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics* (5th ed.). Sage edge.

Pallant, J. (2011). *SPSS Survival Manual: a step by step guide to data analysis using SPSS* (4th ed.). Allen & Unwin.

Verma, J. P., & Abdel-Salam, G. A.-S. (2019). *Testing statistical assumptions in research*. John Wiley & Sons, Inc.

Experimental design and biometric research. Toward innovations, pp. 129-151
https://doi.org/10.18559/978-83-8211-079-1/III2

2.

INDEPENDENT SAMPLES— MORE HYPOTHESES TESTING

Blaženka Knežević
Faculty of Economics and Business, University of Zagreb

Berislav Žmuk
Faculty of Economics and Business, University of Zagreb

Abstract: Two-way analysis of variance (ANOVA) without replication is called a factorial ANOVA with two factors. It is used to test if there is a significant difference between means of several sets of data (groups) dependable on two independent factors. It is applied when we have one measurement variable and two nominal variables (usually called 'factors' or 'main effects'). In this chapter hypotheses and assumptions of the method are given. Then the example of the procedure of two-way analysis of variance (ANOVA) without replication is described in details. The two-way analysis of variance (ANOVA) with replication is utilized to simultaneously test the effects of varying two variables for a sample which consists of more than one respondent per a certain combination of variables. The example of the procedure of two-way analysis of variance (ANOVA) with replication is described in details in this chapter. For both procedures the easy to follow examples shows the procedure step-by-step. The practical part includes the guidance for SPSS and for Excel.

Keywords: analysis of variance, ANOVA, two-way analysis of variance without replication, two-way analysis of variance with replication.

2.1. Two-way analysis of variance (ANOVA) without replication

General information

Two-way analysis of variance (ANOVA) without replication is used to determine if there is a significant difference between means of several subpopulations (groups) dependable on two independent factors (Balakirshnan, Render, & Stair, 2007; Fraser, 2016; Randolph & Myers, 2013). In other words, a two-way ANOVA (also called factorial ANOVA, with two factors) is applied when we have one measurement variable and two nominal variables (usually called 'factors' or 'main effects'). For instance, we could apply ANOVA with two factors without replication when explaining differences of revenue generation in different stores for different seasons where store would be one factor and particular season other factor by which we test differences in revenue generation. Or we can apply ANOVA without replication when we want to test in-field results of promotional activities of various sales representatives and various location where activities are applied (in this case representatives are one factor, locations are second factor and results or effect of promotion is variable which is tested for differences according to those two factors).

Hypothesis

In two-way ANOVA without replication, there is a single observation for each combination of the nominal variables, therefore we have only two null hypotheses:

H0(1): There is no difference between means of observations grouped by one factor.

H0(2): There is no difference between means of observations grouped by other factor.

In two-way ANOVA without replication we assume that there is no interaction between factors.

Assumptions

There are the following assumptions associated with the two-way ANOVA without replication (Dean & Ilowsky, 2013; Field, 2013; Winston, 2016):

- dependent variable—should be continuous;
- two independent variables (factors)—should be in two or more categorical, independent groups;
- each sample has to be drawn independently of the other samples (the samples are disjoint);
- the variance of data in the different groups should homogenous;

– each sample should be taken from a normally distributed population;
– there is no dependence or interaction between factors;
– there are no significant outliers.

Example
Dataset: In Retail Company Tradex we collected data on value of expired or spoiled merchandize in EUR per week. Now we want to analyze if this value varies according to sales region and according to product categories.

Data info:
– variable 1: product category—nominal (1—Fruits and vegetables, 2—Diaries, 3—Meat);
– variable 2: sales region – nominal (1—East, 2—West, 3—North, 4—South);
– variable 3: value of expired or spoiled merchandize in EUR per week—numeric.

Hypotheses:
H0(1): There is no difference between means of value of expired or spoiled merchandize grouped by product category.
H1(1): There is a difference between means of value of expired or spoiled merchandize grouped by product category.

H0(2): There is no difference between means of value of expired or spoiled merchandize grouped by sales region.
H1(2): There is a difference between means of value of expired or spoiled merchandize grouped by sales region.

Testing the hypotheses in SPSS
In our example, we observed value of expired or spoiled merchandize in EUR per week in a particular sales region and in particular product categories. In Table 1, dataset is shown.

Table 1. Dataset for two-way ANOVA without replication

	1—Region East	2—Region West	3—Region North	4—Region South
1—Fruits and vegetables	130	150	280	140
2—Diaries	250	320	330	230
3—Meat	120	130	250	180

Source: The authors' own elaboration.

For analysis in SPSS we will form three variables "Type", "Region" and "FW", then we will proceed to enter data. In the first row (see Figure 1, Row 1) we will enter 1, 1, 130 (meaning: 1—Fruit and vegetables, 1—Region East, 130 EUR of

expired and spoiled merchandize per week). Then we will proceed to enter 1, 2, 150 (see Figure 1, row 2) for Fruits and vegetables in Region West—value 150 EUR.

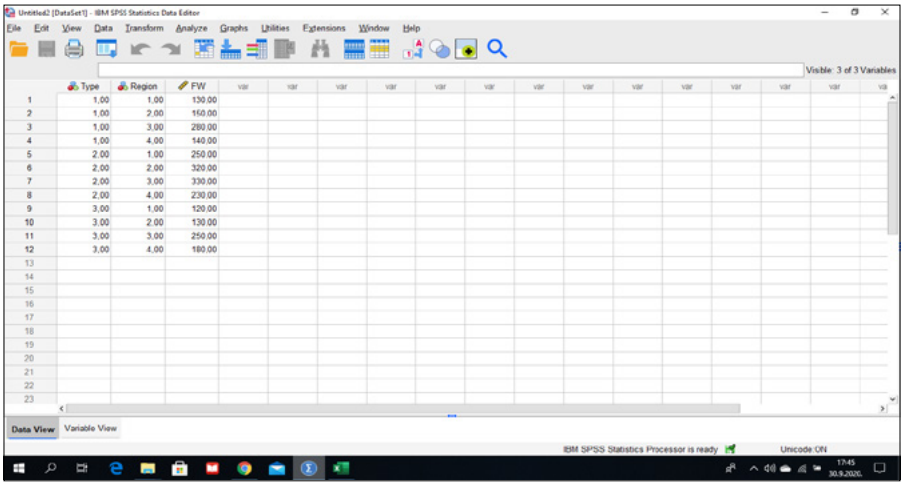


Figure 1. Dataset prepared for two-way ANOVA without replication analysis in SPSS

Source: The authors' own elaboration.

When all data is entered, we will choose type of analysis (see Figure 2). For two-way ANOVA without replication, we will select General Linear Model, Univariate option and then we will specify further options.

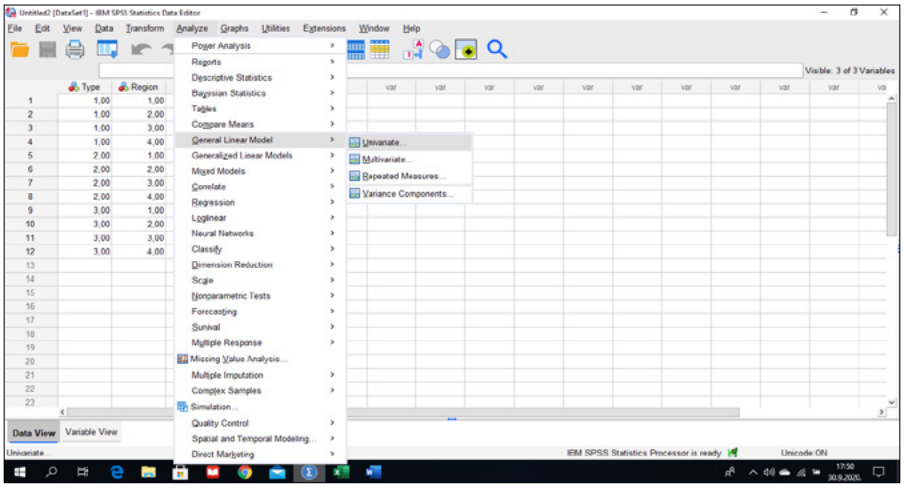


Figure 2. Choosing type of analysis—General Linear Model—Univariate

Source: The authors' own elaboration.

Firstly, we will have to set variables as shown in Figure 3. Our dependent variable is "FW" and our Fixed Factors are "Type" and "Region".

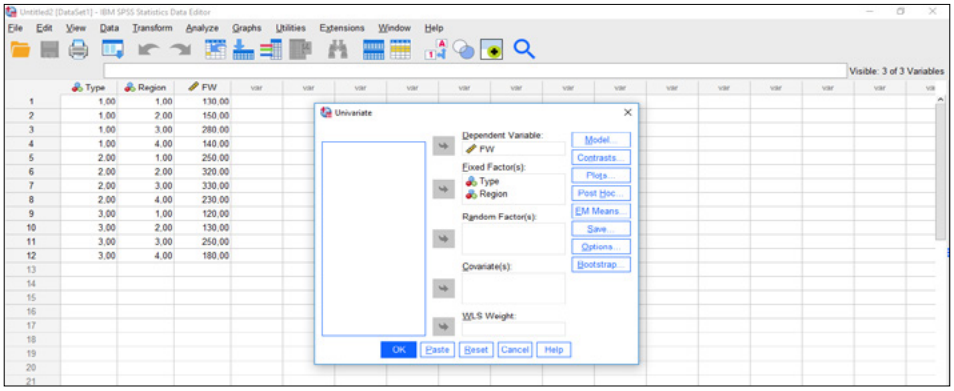


Figure 3. Setting options for analysis—variable specification

Source: The authors' own elaboration.

Secondly, we have to customize type of analysis. Therefore, we will have to specify our model. Therefore we select option "Model". We will use option "Build terms" and we will observe "Main effects" (see the central part of the screen at Figure 4). Moreover, by using arrow at the middle of the screen we have to transfer our factors from column "Factors & Covariates" to column "Model" (see right part of Figure 4). Then we will click to "Continue" and by clicking on "OK" at the previous screen, we will perform our analysis.

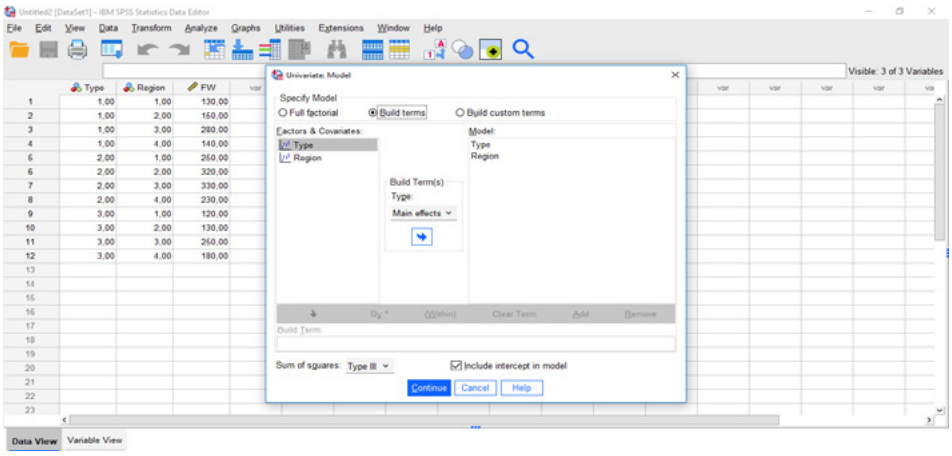


Figure 4. Setting options—Specify Model—customization of analysis

Source: The authors' own elaboration.

In Figure 5 results of analysis are shown. For quick interpretation, we will pay attention to the column "Sig." and we will search for value lower than .05 in order to not reject or reject our hypotheses. At the factor Type the significance value is .006 whereas at the factor Region the significance value is .022. Because both significance values are lower than .05, we can reject H0 hypotheses for both factors (Type and Region).

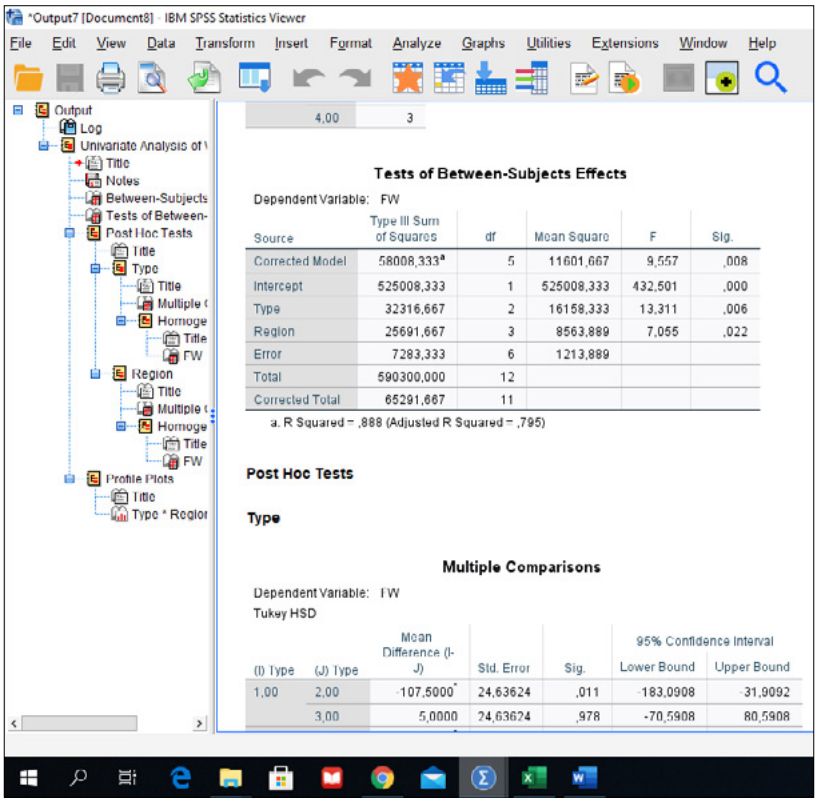


Figure 5. Result of two-way ANOVA without replication analysis in SPSS

Source: The authors' own elaboration.

Based on this we conclude that we can conclude that there is the difference between means of value of expired or spoiled merchandize grouped by product category and that there is difference between means of value of expired or spoiled merchandize grouped by sales region.

Testing the hypotheses in Excel

Same dataset is entered to Excel. In Figure 6 collected data is shown in format suitable for analysis in Excel.

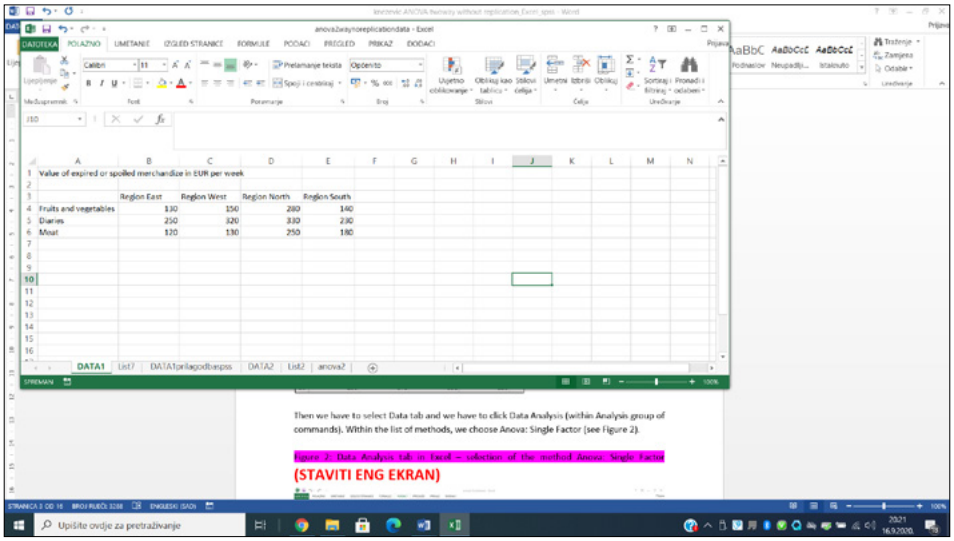


Figure 6. Dataset for two-way ANOVA (without replication) analysis in Excel

Source: The authors' own elaboration.

Then we have to select Data tab and we have to click Data Analysis (within Analysis group of commands). Within the list of methods, we choose ANOVA: two factor without replication (see Figure 7).

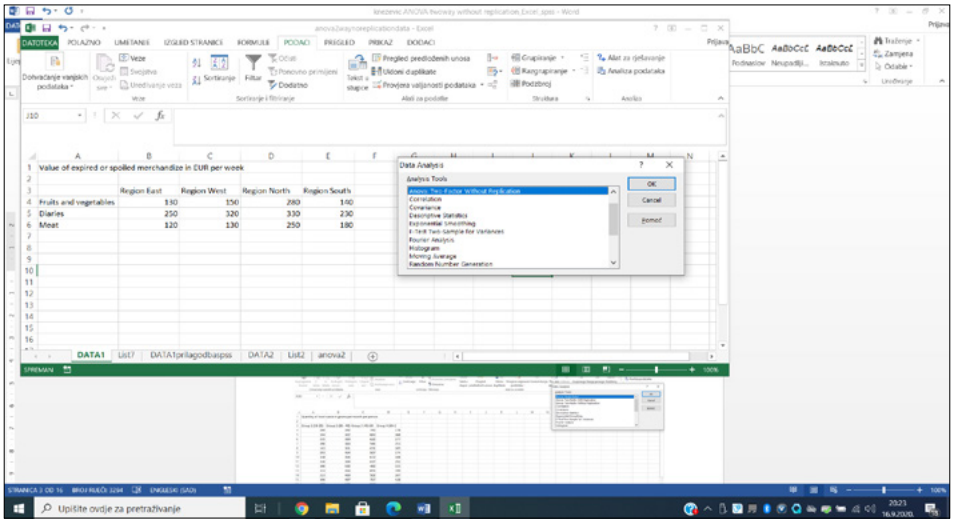


Figure 7. Data Analysis tab in Excel—selection of the method ANOVA: two-factor without replication

Source: The authors' own elaboration.

In the dialog box of ANOVA: two-factor without replication we have to configure as follows (see Figure 8):

- input range of the dataset including labels, in our example it is A3:E6;
- check existence of data labels (see Labels);
- output range, we can choose to show data at some position at the active Worksheet, then we have to specify exact cell from which our results are going to be presented (such as F3); but in our case we rather specified New Worksheet as the location of our results, also we can specify a name for our output (in our example "ANOVA");
- the last thing is the level of significance, i.e. alpha value. There we can use default value as it is already set to .05.

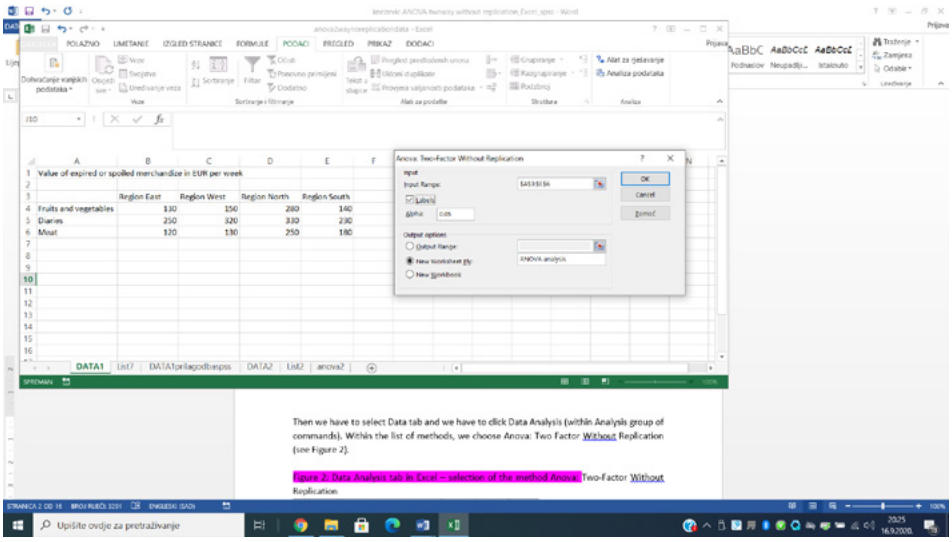


Figure 8. Dialog box ANOVA: two-factor without replication

Source: The authors' own elaboration.

In Figure 9 results of data analysis is shown and we can interpret our results. First of all, we have basic descriptive statistical data grouped by both factors. From this part we can read how many observations we had in which group (see column Number), then we can see what the average value of expired or spoiled merchandize in EUR per week in each group is, and what the variance within each group is. For instance, when we observe product categories, the lowest average of 170 EUR is expired or spoiled in Meat category and if we observe regions, the lowest average value of spoiled or expired merchandize is in region East and it is 166.6667 EUR.

In addition, ANOVA results are shown. In this table, the most important reading is *p*-value because by it we can decide not to reject or to reject the null hypothesis. In

our case the *p*-value for rows .006222 (remember, in our dataset product categories are entered to Excel, see Figure 1). In this case *p*-value is lower than *alpha* value of .05 and it means that we can reject the null hypothesis H0(1) and we can conclude that there is difference between means of value of expired or spoiled merchandize grouped by product category. So, this difference is statistically significant at the level of .05.

In addition, we can observe that in our case the *p*-value for columns (in our case, Regions) is .021524 which is lower than .05. Therefore, we can reject the null hypothesis H0(2), and we can conclude that there is difference between means of value of expired or spoiled merchandize grouped by sales region. So, the test results have shown that this difference is statistically significant at the level of .05.

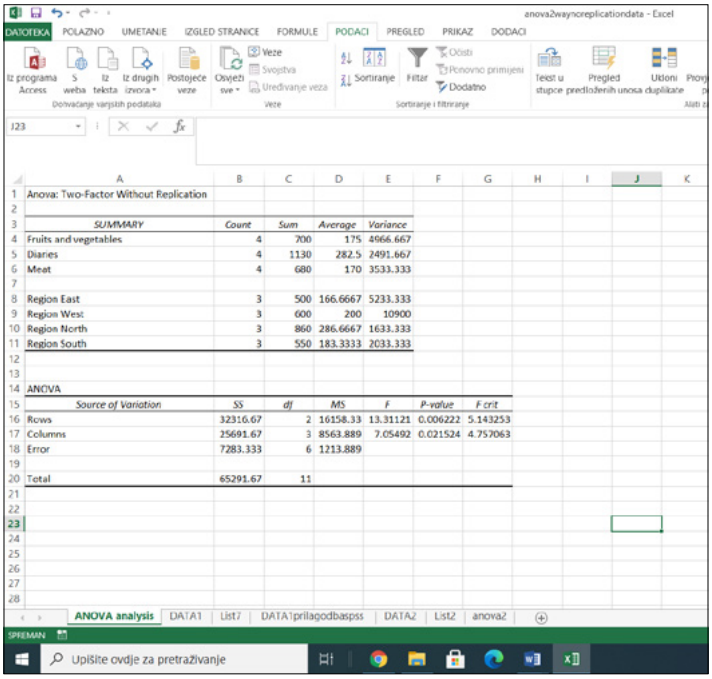


Figure 9. Two-way ANOVA without replication results

Source: The authors' own elaboration.

If ANOVA shows us that there is a statistically significant difference between observed groups, we have to do post hoc analysis by comparing pair of groups in order to explain which group differs in comparison to other group. For this purpose in Excel we can perform several *t*-tests. Procedure of *t*-tests is already explained in details in chapter about one-way ANOVA.

Summary of the example

Dataset: The value of expired or spoiled merchandize in EUR per week in different sales regions is observed. In addition, the value of expired or spoiled merchandize in EUR per week is observed according to the product category as well. For the purpose of the analysis four sales regions and three product categories are defined. For each combination of a sales region and a product category the value of expired or spoiled merchandize in EUR per week is collected. On that way, 12 data values of expired or spoiled merchandize in EUR per week were on disposal for our analysis needs.

Data info:
- variable 1: product category—nominal (1—Fruits and vegetables, 2—Diaries, 3—Meat);
- variable 2: sales region—nominal (1—East, 2—West, 3—North, 4—South);
- variable 3: value of expired or spoiled merchandize in EUR per week—numeric.

The two-way ANOVA approach was used to inspect whether the average value of expired or spoiled merchandize grouped by product category can be considered the same across all three product categories. Also, in the same time, the two-way ANOVA approach was used to inspect whether the average value of expired or spoiled merchandize grouped by sales categories can be considered the same across all four sales categories. The results of the two-way ANOVA have shown that there was a statistically significant difference between product categories groups ($F(2,6) = 13.311$, $p = .0062$). On the other side, the results of the two-way ANOVA have also shown that there was a statistically significant difference between sales region groups ($F(3,6) = 7.055$, $p = .0215$).

More info about two-way ANOVA without replication

The two-way ANOVA without replication can be observed as an extension of the one-way ANOVA. Whereas at the one-way analysis just one factor is observed, here at two-way ANOVA without replication two factors are inspected in the same time. Despite the fact that the two-way ANOVA without replication analysis is more complex than the one-way ANOVA they are sharing the same assumptions with additional assumption that there is no dependence or interaction between factors (Barrow, 2017; Randolph & Myers, 2013).

In the analysis, at both observed factors statistically significant differences are found. However, we do not know whether all means between all three product categories or all means between all four sales regions are different or the difference is statistically significant just between some groups. In order to find out that, in the following step in the analysis Tukey post hoc tests can be conducted to observe differences between pairs of categories at given factors. The Tukey post hoc tests procedure and interpretations are analogous to those explained at the one-way ANOVA.

References

Balakirshnan, N., Render, B., & Stair, R. M. (2007). *Managerial decision modeling with spreadsheets*. Pearson Prentice Hall.
Barrow, M. (2017). *Statistics for economics, accounting and business studies*. Pearson.
Dean, S., & Illowsky, B. (2013). *Introductory statistics*. OpenStax College.
Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics* (5th ed.). Sage edge.
Fraser, C. (2016). *Business statistics for competitive advantage with Excel 2016*. Springer.
Randolph, K. A., & Myers, L. L. (2013). *Basic statistics in multivariate analysis*. Oxford University Press.
Winston, W. (2016). *Microsoft Excel 2016: Data analysis and business modelling*. Microsoft Press.

2.2. Two-way analysis of variance (ANOVA) with replication

General information

The two-way analysis of variance (ANOVA) with replication simultaneously tests the effects of varying two variables (such as gender and age or wealth and geographic area, and their interaction) for a sample which consists of more than one respondent per a certain combination of variables. While in two-factor ANOVA without replication there was only one sample item (observation) for each combination of factors.

Replication refers to the number of cases observed within the same combination of factors (Field, 2013; Fraser, 2016; Winston, 2016). Usually, we use this method in a case of a balanced research design when the size of each subgroup according to two factor is equal. Because then, we can calculate the mean square for each of the two factors, for their interaction, and for each combination of factors.

Hypotheses

In two-way ANOVA with replication, there are more than one observation for each combination of the nominal variables, therefore it is possible to examine interaction between factors as well. So, we will have three null-hypotheses:

H0(1): There is no difference between means of observations grouped by one factor.
H0(2): There is no difference between means of observations grouped by other factor.
H0(3): There is no interaction between factors.

The alternative hypothesis to every above stated is its negation.

The interaction test shows us if the effects of one factor depend on the other factor (Balakirshnan et al., 2007; Dean & Illowsky, 2013).

Assumptions

There are the following assumptions associated with the two-way ANOVA with replication (Barrow, 2017; Field, 2013; Randolph & Myers, 2016):

– dependent variable is continuous;
– two independent variables should consist of at least two or more categorical, independent groups;
– dependent variable should be approximately normally distributed for each combination of independent variables;
– independence of observations, i.e. no relationship between the observations in each group or between the groups;
– no significant outliers because they can have a negative effect on the two-way ANOVA results;
– homogeneity of variances for each combination of the groups of the two independent variables.

Example

Situation: Recently we finished our series of seminars in area of Sustainable consumption and Social responsibility in our local community center. So, we want to test effects of this education in real-life environment. We had 3 educations on various topics which included topics on Eco-friendly products and Fair trade. Therefore, together with the owner of the local supermarket we are collecting data on consumer will to purchase sustainable products in everyday life. We want to test is our education effective or not.

Dataset: observed number of products in a shopping cart at the checkout in a local supermarket according to product labels, number of educations taken (no education, one education, two educations or three educations). For each out of 4 levels of education we collect same number of observations (5), i.e. we have same number of respondents (5). Sample consisted of 20 respondents. For each respondent we count number of products in a shopping cart according to 3 types of product labels.

Data info:
– variable 1: number of educations taken—nominal (1—No education, 2—One education, 3—Two educations, 4—Three educations);
– variable 2: label of product—nominal (1—Eco-friendly, 2—Fair trade, 3—No label);
– variable 3: number of items (products) in the shopping cart—numerical.

Hypotheses:
H0(1): There is no difference between means of number of items in shopping cart grouped by number of educations taken.
H1(1): There is a difference between means of number of items in shopping cart grouped by number of educations taken.

H0(2): There is no difference between means of number of items in shopping cart grouped by label of products.

H1(2): There is a difference between means of means of number of items in shopping cart grouped by label of product.

H0(3): There is no interaction between factors.
H1(3): There is interaction between factors.

Testing the hypotheses in SPSS

In our example we observed how many items (product) labeled with labels in field of sustainable business some customer had and we observed how many educations in our local community this person attended in the field of sustainability and social responsibility, then we entered data into SPSS (see Figure 10). For instance, in the first row we have recorded data for respondent who did not attended seminars (see column "Education" where 1 means No education), and this person had only 2 items (see "Numberofitems" column where we recorded quantity 2) labeled as Eco-friendly products (see column "Label" where 1 meaning Eco-friendly is entered). For the same respondent we recorded 1, 2, 3—see row 6 at Figure 10 (as he/she did not attend education (1), and had 3 items labeled as 2—Fair trade) and we concluded entering data on contents of his/her shopping cart by entering 1, 3, 7—see row 11 at Figure 10 (as he/she did not attend education (1), and had 7 items without any label in field of sustainability, 3 is entered for No label in column "Label"). Then we entered data for all other respondents and items in their shopping carts.

As there were five respondents from each group according to 4 education levels, and for each respondent we counted number of items in their shopping carts for 3 labels, our dataset in SPSS at the end had 60 rows.

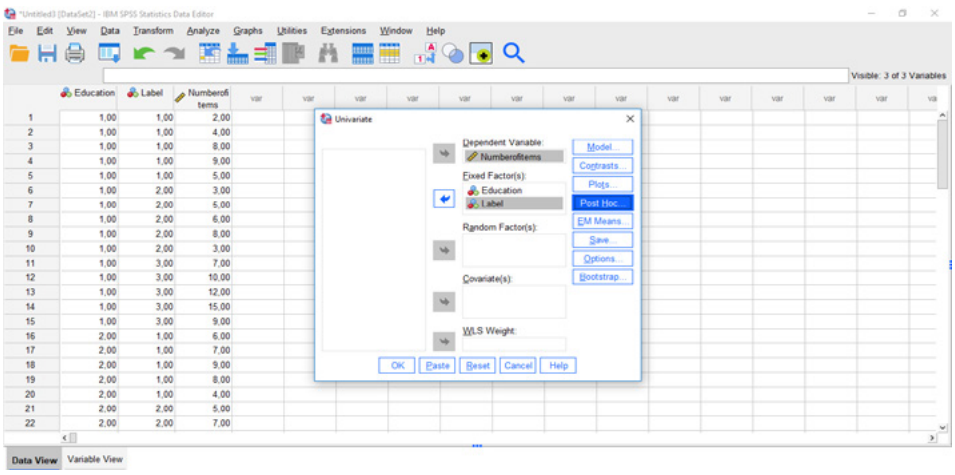


Figure 10. Excerpt of the dataset suitable for two-way ANOVA with replication

Source: The authors' own elaboration.

When data is entered, we proceed to Method selection (see Figure 11). For two-way ANOVA with replication as a method we will choose Analyze—General Linear Model and select Univariate.

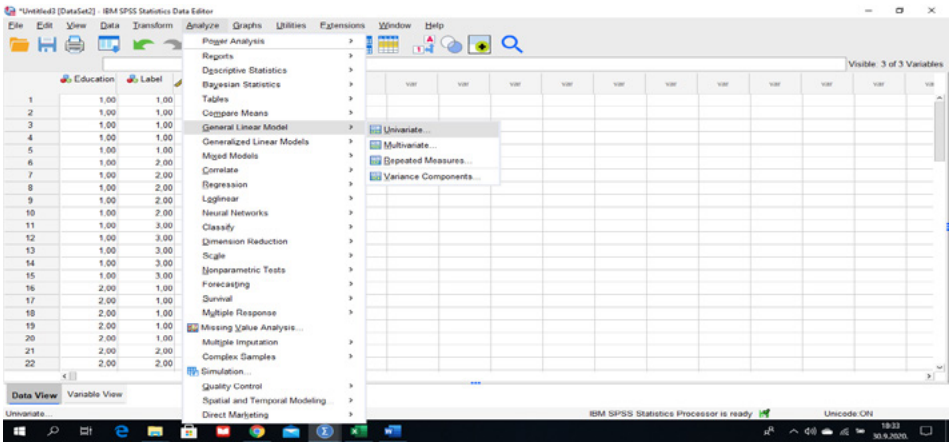


Figure 11. Choosing method—General Linear Model—Univariate

Source: The authors' own elaboration.

After that, we will define variables as shown at Figure 12. Dependent variable is "Numberofitems", while fixed factors are "Education" and "Label". Now we can specify several options which will enable us to interpret results of the analysis. In this chapter we will set options for Plot diagram and Post hoc analysis.

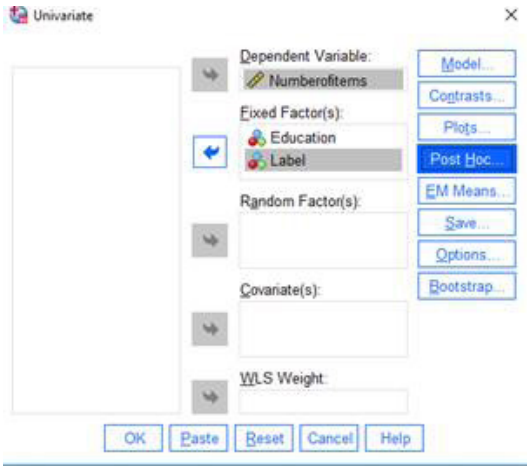


Figure 12. Setting basics options for the analysis—variables

Source: The authors' own elaboration.

Plot diagram settings are shown at Figure 13. Plot diagram will enable us to visualize results of ANOVA in quick and intuitive manner. It will show us inter-relation between factors in a graphic mode. We will set Education to be shown at horizontal axis and Label as a separate lines, then we will click Add and we will just check if in the text box under word "Plots" "Education*Label" is shown and if the checkbox is selected for "Line Chart". If everything is correct, we can confirm everything by clicking the button "Continue".

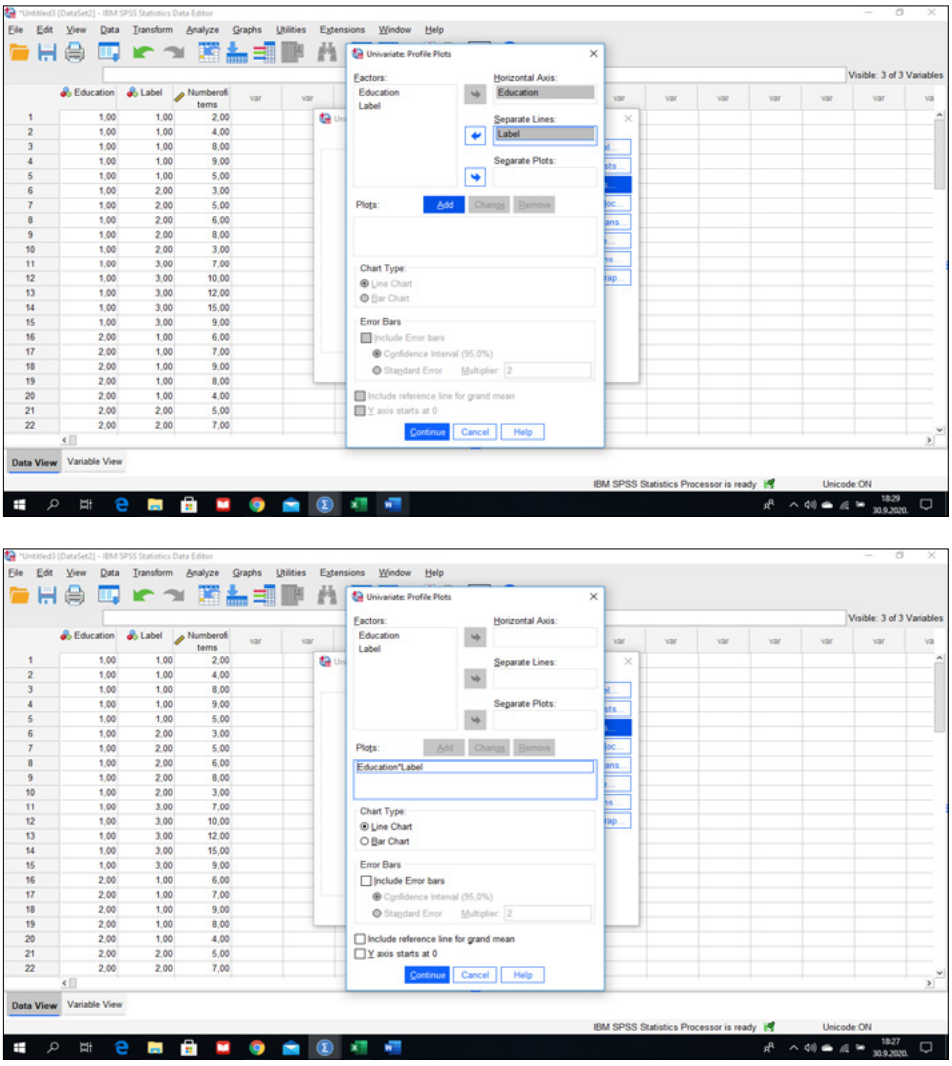


Figure 13. Settings for Plot diagram

Source: The authors' own elaboration.

Finally, we will define options for the post hoc analysis. We will perform post hoc tests for Education and Label and we will mark Tukey post hoc analysis (see Figure 14). Again, we will click button "Continue" and then at previous screen confirm analysis by clicking to "OK" button.

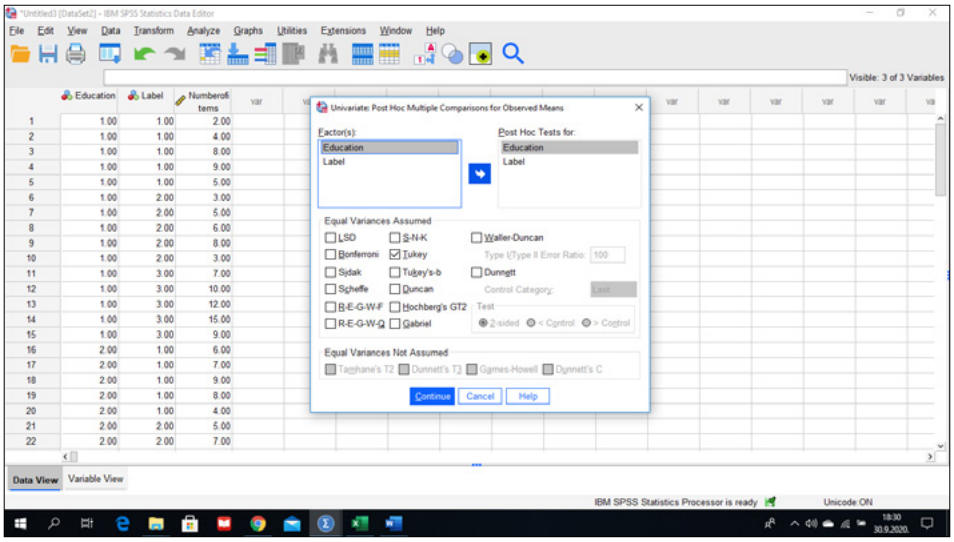


Figure 14. Setting post hoc analysis options

Source: The authors' own elaboration.

When analysis is finished, we will get results separated in several segments. Firstly, we will have two-way ANOVA general results (see Figure 15). On the basis of this table we will be able not to reject or reject our starting null hypotheses.

Tests of Between-Subjects Effects

Dependent Variable: Numberofitems

Source	Type III Sum of Squares	df	Mean Square	F	Sig
Corrected Model	611.783[a]	11	55.617	9.715	.000
Intercept	4950.417	1	4950.417	864.702	.000
Education	144.850	3	48.283	8.434	.000
Label	39.033	2	19.517	3.409	.041
Education * Label	427.900	6	71.317	12.457	.000
Error	274.800	48	5.725		
Total	5837.000	60			
Corrected Total	886.583	59			

a. R Squared = .690 (Adjusted R Squared = .619)

Figure 15. Two-way ANOVA with replication results

Source: The authors' own elaboration.

We will pay attention to column "Sig." where *p*-values are shown. In this column we will observe which values are less than .05. In our case all significance values are lower than .05 (Education $p < .001$, Label $p = .041$, Education * Label $p < .001$). Therefore, we can reject all three H0 hypotheses. Therefore, at significance level .05 we can conclude that:

- there is a difference between means of number of items in shopping cart grouped by number of educations taken;
- there is a difference between means of means of number of items in shopping cart grouped by label of product;
- there is interaction between factors.

In Figure 16 the plot diagram is shown. We can observe that Education and Label curve intersect which means that those variables are in interaction and that we have to take this fact into account when interpret our post hoc data.

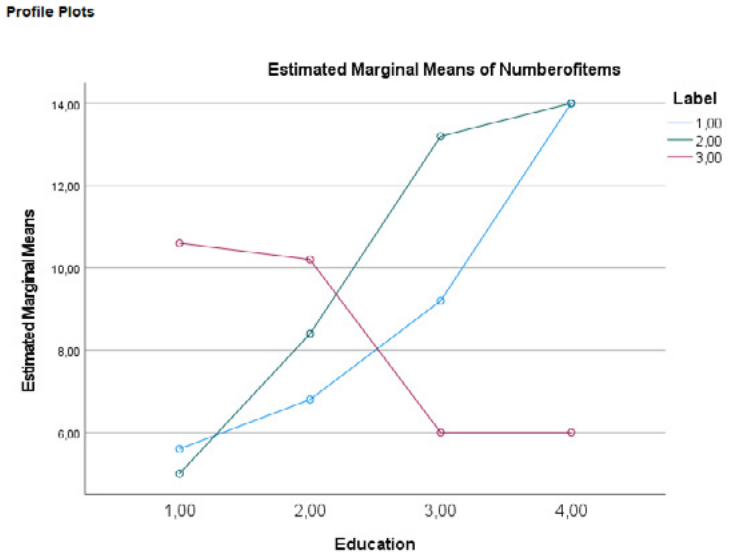


Figure 16. Plot diagram

Source: The authors' own elaboration.

As the two-way ANOVA results have shown that there are differences between groups (factors), it is recommended to do post hoc analysis to get better insight into the results. Results of the post hoc analysis help us to observe between which groups are strongest differences and are there some groups which do not differ from each other.

In Figure 17 Post hoc analysis results are shown based on Label. We can observe that between those who did not attend any education (1) and those who attended

two educations (3), then between those who did not attend any education (1) and those who attended three educations (4), and between those who attended one education (2) and those who attended three educations (4) at level of .05, there is statistically significant difference between means of number of items in shopping cart. While other pairs of groups did not show statistically significant difference (*p*-values in column "Sig." are higher than .05).

Multiple Comparisons

Dependent Variable: Numberofitems
Tukey HSD

(I) Education	(J) Education	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1.00	2.00	-1.4000	.87369	.387	-3.7252	.9252
	3.00	-2.4000*	.87369	.041	-4.7252	-.0748
	4.00	-4.2667*	.87369	.000	-6.5919	-1.9415
2.00	1.00	1.4000	.87369	.387	-.9252	3.7252
	3.00	-1.0000	.87369	.664	-3.3252	1.3252
	4.00	-2.8667*	.87369	.010	-5.1919	-.5415
3.00	1.00	2.4000*	.87369	.041	.0748	4.7252
	2.00	1.0000	.87369	.664	-1.3252	3.3252
	4.00	-1.8667	.87369	.156	-4.1919	.4585
4.00	1.00	4.2667*	.87369	.000	1.9415	6.5919
	2.00	2.8667*	.87369	.010	.5415	5.1919
	3.00	1.8667	.87369	.156	-.4585	4.1919

Based on observed means.
The error term is Mean Square(Error) = 5.725.
*. The mean difference is significant at the 0.05 level.

Figure 17. Post hoc analysis results—Education

Source: The authors' own elaboration.

Multiple Comparisons

Dependent Variable: Numberofitems
Tukey HSD

(I) Label	(J) Label	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1.00	2.00	-1.2500	.75664	.234	-3.0799	.5799
	3.00	.7000	.75664	.627	-1.1299	2.5299
2.00	1.00	1.2500	.75664	.234	-.5799	3.0799
	3.00	1.9500*	.75664	.034	.1201	3.7799
3.00	1.00	-.7000	.75664	.627	-2.5299	1.1299
	2.00	-1.9500*	.75664	.034	-3.7799	-.1201

Based on observed means.
The error term is Mean Square(Error) = 5.725.
*. The mean difference is significant at the 0.05 level.

Figure 18. Post hoc analysis results—Label

Source: The authors' own elaboration.

In Figure 18 Post hoc analysis results are shown based on Education. We can observe that between products with label Fair trade (2) and products with no label (3) at level of .05, there is a statistically significant difference between means of number of items in shopping cart. While other pairs of groups did not show statistically significant difference (p-values in column "Sig." are higher than .05).

Testing the hypotheses in Excel

To perform two-way ANOVA with replication in Excel, it is extremely important to have the same number of observations for one factor, in our case—education level. Therefore, we made and entered 5 observations for each level of education (see rows in Figure 19).

	A	B	C	D	E	F
1	Number of items in shopping cart					
2						
3			Eco-friendly	Fair trade	No label	
4	Number of educations	No education	2	3	7	
5			4	5	10	
6			8	6	12	
7			9	8	15	
8			5	3	9	
9		One education	6	5	7	
10			7	7	12	
11			9	10	13	
12			8	11	9	
13			4	9	10	
14		Two educations	8	15	7	
15			9	12	4	
16			10	13	8	
17			12	16	6	
18			7	10	5	
19		Three education	12	16	4	
20			17	14	5	
21			19	15	6	
22			12	13	8	
23			10	12	7	
24						

Figure 19. Dataset for two-way ANOVA (with replication) analysis in Excel

Source: The authors' own elaboration.

Data is entered into Excel in format suitable for data analysis grouped by number of educations as we have same number of respondents for each level of education, then in columns we enter another grouping variable, i.e. product labels (see Figure 19). In each row we enter data on one survey participant. For instance, we enter data for participants which did not attend any seminar (no education) in five rows, each row for one participant. First participant without education put 2 eco-friendly items, 3 fair-trade items and 7 no label items to his/her shopping cart, while second participant without education put 4 eco-friendly items, 5 fair trade items and 10 items without labels into his/her shopping cart.

To perform two-way ANOVA with replication, we have to click Data tab and we have to choose Data Analysis tab (within Analysis group of commands). Within the list of methods, we choose ANOVA: two factor with replication (see Figure 20).

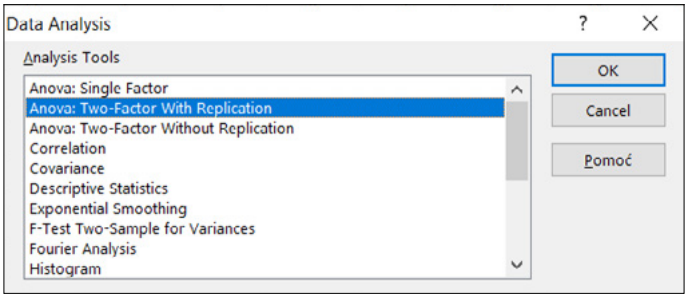


Figure 20. Data Analysis tab in Excel—selection of the method ANOVA: two-factor with replication

Source: The authors' own elaboration.

In the dialog box of ANOVA: two-factor with replication we have to configure as follows (see Figure 21):

- input range of the dataset including labels, in our example it is B3:E23;
- input number of rows per sample (i.e. number of observations), in our case 5;
- output range, we can choose to show data at some position at the active Worksheet, then we have to specify exact cell from which our results are going to be presented (such as F3); but in our case we rather specified New Worksheet as the location of our results, also we can specify a name for our output (in our example ANOVA);
- the last thing is the level of significance, i.e. alpha value. There we can use default value as it is already set to .05.

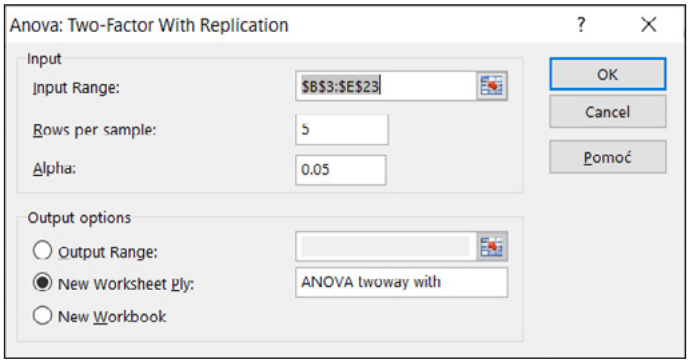


Figure 21. Dialog box ANOVA: two-factor with replication

Source: The authors' own elaboration.

	A	B	C	D	E	F
1	Anova: Two-Factor With Replication					
2						
3	SUMMARY	Eco-friend	Fair trade	No label	Total	
4	*No education*					
5	Count	5	5	5	15	
6	Sum	28	25	53	106	
7	Average	5.6	5	10.6	7.066667	
8	Variance	8.3	4.5	9.3	13.06667	
9						
10	*One education*					
11	Count	5	5	5	15	
12	Sum	34	42	51	127	
13	Average	6.8	8.4	10.2	8.466667	
14	Variance	3.7	5.8	5.7	6.409524	
15						
16	*Two educations*					
17	Count	5	5	5	15	
18	Sum	46	66	30	142	
19	Average	9.2	13.2	6	9.466667	
20	Variance	3.7	5.7	2.5	12.69524	
21						
22	*Three educations*					
23	Count	5	5	5	15	
24	Sum	70	70	30	170	
25	Average	14	14	6	11.33333	
26	Variance	14.5	2.5	2.5	20.80952	
27						
28	*Total*					

	A	B	C	D	E	F	G
34							
35	ANOVA						
36	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
37	Sample	144.85	3	48.28333	8.43377	0.000132	2.798061
38	Columns	39.03333	2	19.51667	3.409025	0.041269	3.190727
39	Interaction	427.9	6	71.31667	12.45706	2.08E-08	2.294601
40	Within	274.8	48	5.725			
41							
42	Total	886.5833	59				
43							

Figure 22. Two-way ANOVA with replication results

Source: The authors' own elaboration.

In Figure 22 there are results of two-way ANOVA with replication analysis. First of all, in Excel we have basic descriptive statistical data grouped by factor in rows (number of educations) systemized by factor in columns (in our case product labels). From this part we can read how many observations we had in which combination of factors (see rows Count in each sub table), then we can see what the average number of items for each combination of factors, together with data

on variance. For instance, when we observe sub table "No education", we will see that there are on average 5.5 items labeled as Eco-friendly in the shopping cart, 5 items labeled as Free trade, and 10.6 items with no label. On the other hand, in sub table "Three educations" there are on average 14 items labeled as Eco-friendly, 14 labeled as Fair-trade and 6 with no label.

In addition, ANOVA results are shown. In this table, the most important reading is *p*-value because by it we can decide not to reject or to reject the null hypothesis. In our case the *p*-value for Sample (in our case this data refers to number of education) is .000132 and it is less than significance level of .05 which means that we can reject the null hypothesis H0(1), and we can conclude that there is a difference between means of number of items in shopping cart grouped by number of educations taken. So, this difference is statistically significant at the level of .05.

In addition, we can observe that in our case the *p*-value for columns (in this case, product labels) is .041269 which is lower than .05. Therefore, we can reject the null hypothesis H0(2), and we can conclude that there is difference between means of means of number of items in shopping cart grouped by label of product. In other words, this difference is statistically significant at the level of .05.

Moreover, the *p*-value for testing interaction between our two factors (number of education and label of products) is 2.08E-08 which is lower than .05 and means that we can reject H0(3) and consequently conclude that there is an interaction between factors and we have to take it into account when we interpret our data. Because if an interaction effect is present, the impact of one factor depends on the level of the other factor.

Summary of the example

Dataset: The number of products in a shopping cart at the checkout in a local supermarket according to product labels is observed. Three product labels have been defined. Overall, 20 respondents have been selected to participate in the study according to their number of taken educations. Four levels of educations taken are recognized. In the study participated equal number of respondents according to the number of educations taken. Consequently, at each level of educations taken we had 5 respondents for which we measured the number of products with different product labels. On that way, three measurement for each respondent have been conducted.

Data info:

- variable 1: number of educations taken—nominal (1—No education, 2—One education, 3—Two educations, 4—Three educations);
- variable 2: label of product—nominal (1—Eco-friendly, 2—Fair trade, 3—No label);
- variable 3: number of items (products) in the shopping cart—numerical.

The two-way ANOVA with replication approach was used to inspect whether our education is effective in changing consumer habits to purchase sustainable products in everyday life. The results have shown that there was a statistically significant interaction between the number of educations taken and label of product ($F(6, 48) = 12.457$, $p < .001$).

More info about two-way ANOVA with replication

All previously mentioned additional information about two-way ANOVA without replication apply to two-way ANOVA with replication.

The main difference between two-way ANOVA without replication and two-way ANOVA with replication is the sample structure. In the ANOVA without replication we have only a single observation for each combination of nominal variables while in two-way ANOVA with replication we have more than one observation. In other words, in two-way ANOVA without replication design there is only 1 experimental unit for each combination of the factors. While in two-way ANOVA with replication "there are more than one experimental unit per combination of the factors. In such design we have enough degrees of freedom and the interaction between factors can be estimated" (Field, 2013; Fraser, 2016). It is recommended that in two-way ANOVA with replication we should use balanced data or sample with uniform size to get results in efficient manner".

References

Balakirshnan, N., Render, B., & Stair, R. M. (2007). *Managerial decision modeling with spreadsheets*. Pearson Prentice Hall.

Barrow, M. (2017). *Statistics for economics, accounting and business studies*. Pearson

Dean, S., & Illowsky, B. (2013). *Introductory statistics*. OpenStax College.

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics* (5th ed.). Sage edge.

Fraser, C. (2016). *Business statistics for competitive advantage with Excel 2016*. Springer.

Randolph, K. A., & Myers, L. L. (2013). *Basic statistics in multivariate analysis*. Oxford University Press.

Winston, W. L. (2016). *Microsoft Excel 2016: Data analysis and business modelling*. Microsoft Press.

Experimental design and biometric research. Toward innovations, pp. 153-163
https://doi.org/10.18559/978-83-8211-079-1/III3

3.

DEPENDENT SAMPLES— SINGLE HYPOTHESIS TESTING

Sylwester Białowąs
Poznań University of Economics and Business

Adrianna Szyszka
Poznań University of Economics and Business

Abstract: This chapter deals with the approach of "within subjects" and focuses on single hypothesis testing. Both parametrical and non-parametrical versions are described. Every test is introduced, and the full step-by-step SPSS guidance is presented. The sections about effect size and about writing the report are included as well.

Keywords: paired sample *t*-test, Wilcoxon test.

3.1. The paired samples *t*-test

General information
Paired *t*-test is used to compare two related means mostly coming from a repeated measures design. In other words, data are collected by two measures from each observation, e.g. before and after a process or a phenomenon. For example, a researcher wants to test if the changes in the weight before and after a diet are significantly different from zero.

Hypotheses
H0: There is no difference between the paired mean scores.
H1: There is a difference between the paired mean scores.

Assumptions
There are the following assumptions associated with the paired samples *t*-test:
- the level of measurement should be interval or ratio (what in SPSS is indicated as scale level of measurement);
- the sample should be randomly selected which means that the data constitute a representative portion of the total population and every individual has the same chance to be selected into the sample (Verma & Abdel-Salam, 2019; Waters, 2011);
- the difference scores (not the raw scores) should follow the normal distribution.

Example
The community managing the apartment blocks has chosen a random group consisting of 58 families living in middle-size flats. The group got the instructions about electricity savings and recommendation to use the tools of controlling the electricity expenses. We have recorded two electricity bills of every family—one from the period of before, and the other one—after the recommendations.

Data info:
- variable 1: pretest—expenses before the recommendation—measurement level: scale (values: recorded electricity expenses per flat per month in EUR);
- variable 2: posttest—expenses after the recommendation—measurement level: scale (values: recorded electricity expenses per flat per month in EUR).

Testing the assumptions
Normality of distribution of differences

The first step in testing the normality of differences between scores is to calculate a new variable that is the difference between pretest and posttest values.

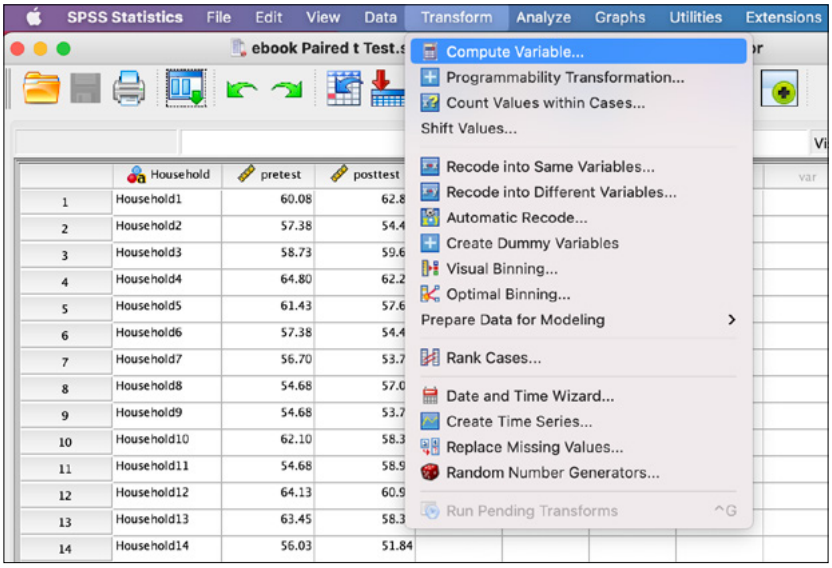


Figure 1. Calculating the difference between pretest and posttest values—path (1)

Source: The authors' own elaboration, IBM SPSS screenshot.

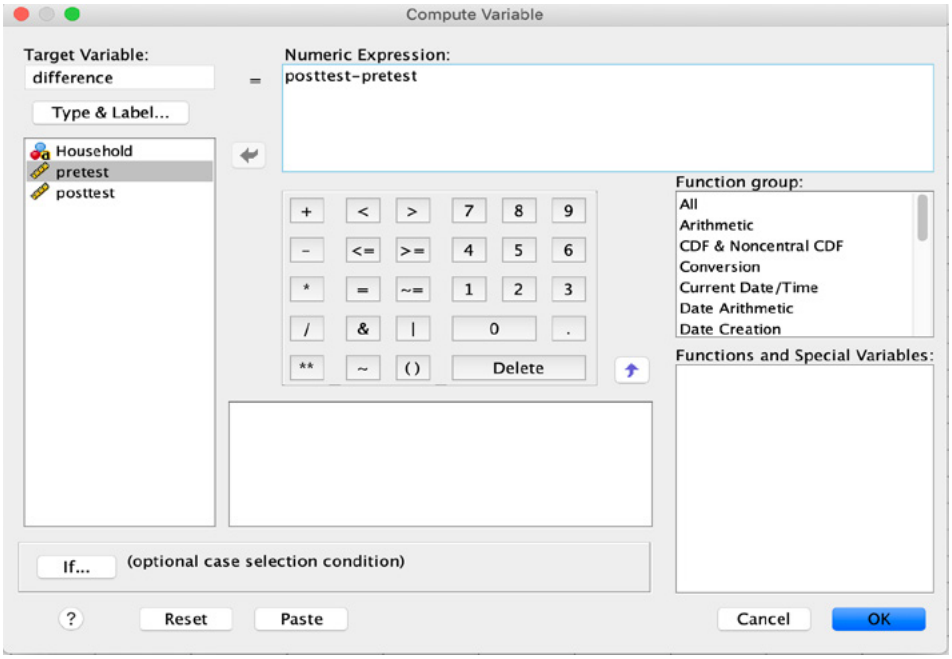


Figure 2. Calculating the difference between pretest and posttest values—path (2)

Source: The authors' own elaboration, IBM SPSS screenshot.

The commonly used test for testing the normality is the Kolmogorov-Smirnov test. This test compares the set of scores obtained in the study to the normally distributed scores.

The procedure of running the Kolmogorov-Smirnov test is shown in part 3, chapter 1. Of course, in the paired samples *t*-test we don't split the file and we measure only one variable—difference.

One-Sample Kolmogorov-Smirnov Test

		difference
N		58
Normal Parameters[a,b]	Mean	-3.5766
	Std. Deviation	3.46691
Most Extreme Differences	Absolute	.096
	Positive	.077
	Negative	-.096
Test Statistic		.096
Asymp. Sig. (2-tailed)		.200[c,d]

a. Test distribution is Normal.

b. Calculated from data.

c. Lilliefors Significance Correction.

d. This is a lower bound of the true significance.

Figure 3. Kolmogorov-Smirnov test—results

Source: The authors' own elaboration, IBM SPSS screenshot.

We decide about the hypothesis by interpreting the *p*-value. If the test is significant ($p < .05$) it means that the data do not follow normal distribution. If the test is non-significant ($p > .05$) the distribution of the obtained scores is normal (Field, 2013; Verma & Abdel-Salam, 2019). In this case, $p \geq .200$ which means that the assumption of normality is fulfilled.

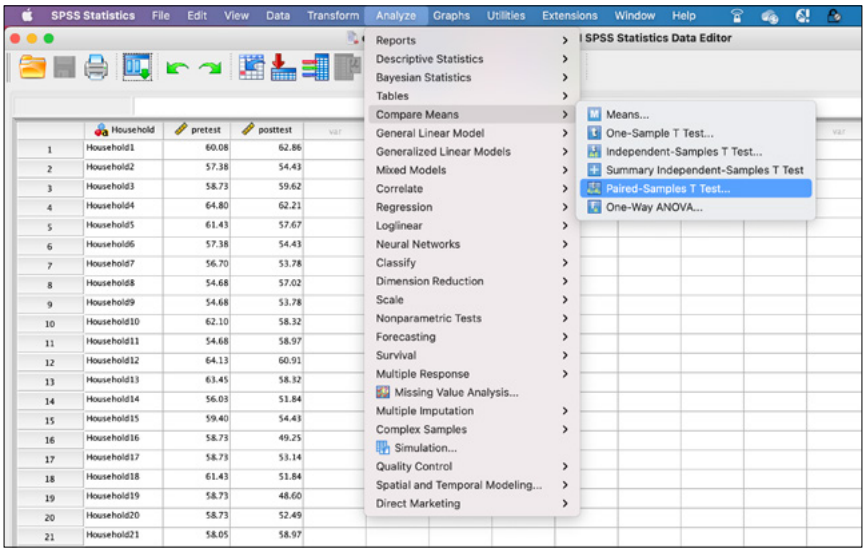


Figure 4. Paired samples *t*-test—path

Source: The authors' own elaboration, IBM SPSS screenshot.

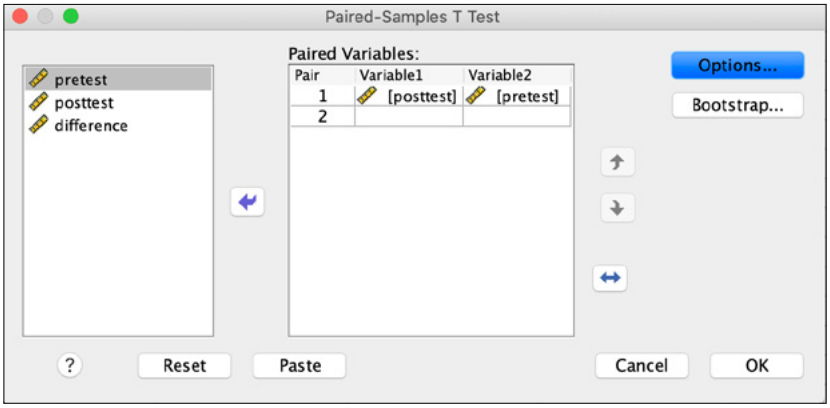
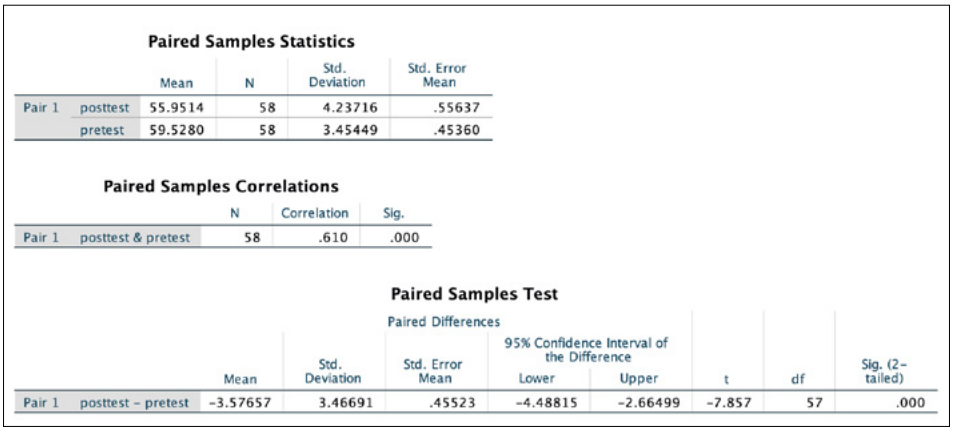


Figure 5. Paired samples *t*-test—dialog box

Source: The authors' own elaboration, IBM SPSS screenshot.



Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	posttest	55.9514	58	4.23716	.55637
	pretest	59.5280	58	3.45449	.45360

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	posttest & pretest	58	.610	.000

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	posttest – pretest	-3.57657	3.46691	.45523	-4.48815	-2.66499	-7.857	57	.000

Figure 6. Paired samples *t*-test—results

Source: The authors' own elaboration, IBM SPSS screenshot.

Results

In the upper table of the outcome (Paired Samples Statistics) we can read that the mean for the pretest is 59.53 and for the posttest is 55.95. It means that the average electricity bill declined by 3.58 EUR.

In the lowest table we can check if the difference is statistically significant by interpreting the *p*-value from the last column (Sig. 2-tailed). This value equals $p < .001$ which is lower than the critical value $p = .05$. It means that we can reject the null hypothesis and interpret the results as the statistically significant difference between pretest and posttest.

Paired samples *t*-test hypotheses resolution:
$p < .05$—there is a significant difference between pretest and posttest; reject H0;
$p > .05$—there is no significant difference between pretest and posttest; do not reject H0.

Effect size

In order to examine whether the observed difference is important, we can calculate effect size. For paired samples *t*-test a popular measure is Cohen's *d*:

$$d = \frac{\left|\overline{x}_1 - \overline{x}_2\right|}{s_{pre}}$$

x_1, x_2 – means of both groups;
s_{pre} – standard deviation of the pretest group.

The Cohen's *d* has the following interpretation:
Below 0.2—no effect,
< 0.2 – 0.5)—small effect,
< 0.5 – 0.8)—medium effect,
0.8 and more—large effect.

$$d = \frac{|55.95 - 59.53|}{3.45} = 1.04$$

In our case, we can observe the large effect (d = 1.04).

Summary
The community managing the apartment blocks has chosen a random group consisting of 58 families living in middle-size flats. The group got the instructions about electricity savings and recommendation to use the tools of controlling the electricity expenses. We have recorded two electricity bills of every family—one from the period of before, and the other one—after the recommendations.

Data info:
- variable 1: pretest—expenses before the recommendation—measurement level: scale (values: recorded electricity expenses per flat per month in EUR);
- variable 2: posttest—expenses after the recommendation—measurement level: scale (values: recorded electricity expenses per flat per month in EUR).

The electricity expenses of the households changed significantly after recommendations $t(58) = -7.857$, $p < .001$, $d = 1.04$. The bills decreased on average from 59.53 EUR (SD = 3.45) to 55.95 EUR (SD = 4.24). A t-test revealed that the difference of 3.58 EUR is statistically significant ($p < .001$), suggesting that the informed groups spent less on electricity than the control group. Cohen's d statistic indicates the large effect.

More info about the paired samples *t*-test
In order to estimate the effect size, we used pretest standard deviation as a baseline. The proposed formula of calculating the denominator is used especially when standard deviation is expected to be increased remarkably by the treatment. Nevertheless, the formula of standard deviation in the denominator may be calculated in other ways. The highly recommended estimate of the baseline is s_{av} given by the following formula:

$$s_{av} = \sqrt{\frac{s_{pre}^2 + s_{post}^2}{2}}$$

It enables us to compare the results with effect size *d* estimations for one group or two independent groups. However, sometimes in the literature the effect size is calculated using standard deviation of differences between scores. This approach is not very advisable since it may give notably different estimations in comparison with different methods (e.g. when the standard deviation of differences is small, the *d* estimation is larger than the calculation with s_{av}) (Cumming, 2012).

References

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge Taylor & Francis Group.

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics* (5th ed.). Sage edge.

Verma, J. P., & Abdel-Salam, G. A.-S. (2019). *Testing statistical assumptions in research*. John Wiley & Sons, Inc.

Waters, D. (2011). *Quantitative methods for business* (5th ed.). Pearson Education Limited.

3.2. Wilcoxon signed-rank test

General information

The Wilcoxon signed rank test is a commonly used nonparametric alternative to the paired samples *t*-test (when the assumptions are violated). It applies to the related samples when we compare the scores in two different points or under two different conditions (e.g. before and after the treatment). It is also used when the dependent variable is measured at ordinal scale. Since the Wilcoxon signed rank test does not require the normality of distribution of the data, it does not compare means but ranks ranks (Pallant, 2011; Verma & Abdel-Salam, 2019).

Hypotheses:
H0: There is no difference between the scores.
H1: There is a difference between the scores.

Assumptions

There are the following assumptions associated with the Wilcoxon signed-rank test:
- the level of measurement of dependent variable must be at least ordinal;
- the score of both groups should be related.

Example

Dataset: The company managing sharing bicycles decided to check the impact of the station location on use of the bicycles. The station was set 200 m from the entrance of the high school. Random sample of the students has been selected. Students were asked about the frequency of using the bicycles. In the middle of the semester the

company set the station closer to the entrance. After one month, the same group of students were asked about the frequency of using bicycles again.

Data info:

- variable 1: pretest—ordinal (declared frequency of using the shared bicycles; 1—more than once a day; 2—every day; 3—2–4 times a week; 4—once a week; 5—once a month; 6—less than once a month; 7—never);
- variable 2: posttest—ordinal (declared frequency of using the shared bicycles; 1—more than once a day; 2—every day; 3—2–4 times a week; 4—once a week; 5—once a month; 6—less than once a month; 7—never).

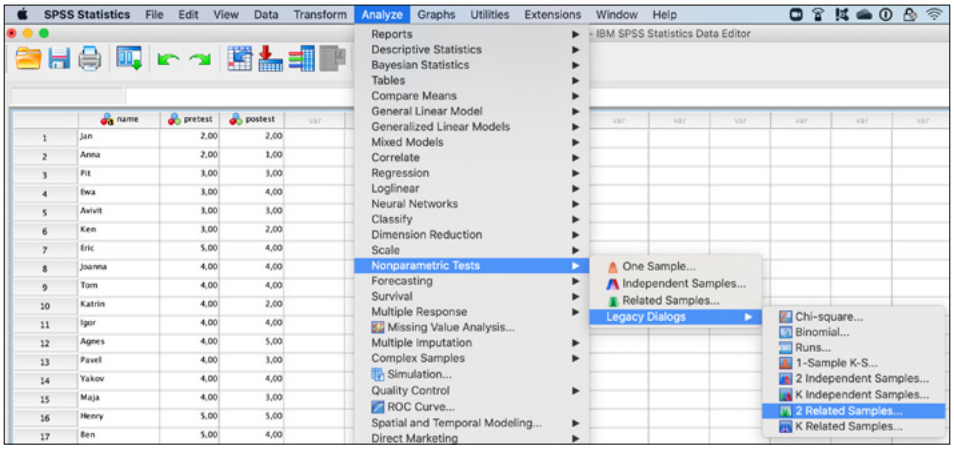


Figure 7. Wilcoxon signed-rank test—path

Source: The authors' own elaboration, IBM SPSS screenshot.

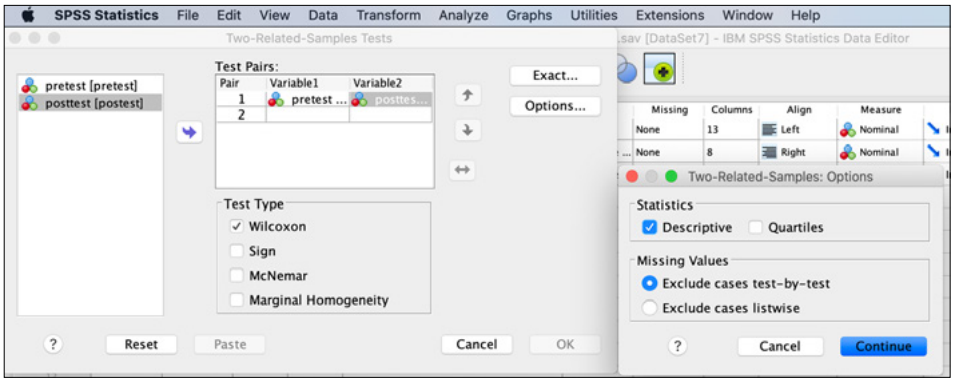


Figure 8. Wilcoxon signed-rank test—dialog box

Source: The authors' own elaboration, IBM SPSS screenshot.

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
pretest	35	4.9429	1.49397	2.00	7.00
posttest	35	4.6000	1.66627	1.00	7.00

Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
posttest – pretest	Negative Ranks	14[a]	9.79	137.00
	Positive Ranks	4[b]	8.50	34.00
	Ties	17[c]		
	Total	35		

a. posttest < pretest
b. posttest > pretest
c. posttest = pretest

Test Statistics[a]

	posttest – pretest
Z	-2.449[b]
Asymp. Sig. (2-tailed)	.014

a. Wilcoxon Signed Ranks Test
b. Based on positive ranks.

Figure 9. Wilcoxon signed-rank test—results

Source: The authors' own elaboration, IBM SPSS screenshot.

Results

In the lowest table we can check if the difference is statistically significant by interpreting the *p*-value from the last row (Asymp. Sig. (2-tailed)). This value equals $p = .014$ which is lower than the critical value $p = .05$. It means that we can reject the null hypothesis and interpret the results as the statistically significant difference between pretest and posttest.

Wilcoxon signed ranked test hypotheses resolution:
$p < .05$—there is a significant difference between pretest and posttest; reject H0;
$p > .05$—there is no significant difference between pretest and posttest; do not reject H0.

Effect size

The effect size measure for Wilcoxon signed ranked test is *r* that is calculated using the statistic *Z* value and *N* which is total number of observations in both groups (the sum of observations in two groups):

$$r = \frac{|Z|}{\sqrt{N}}$$

The *r* has the following interpretation:
Below .1—no effect,
< .1 – .3)—small effect,
< .3 – .5)—medium effect,
.5 and more—large effect (Field, 2013; Pallant, 2011).

$$r = \frac{|-2.449|}{\sqrt{70}} = .29$$

In our example, $r = .29$ which may be considered as a small effect.

Summary

Dataset: The company managing sharing bicycles decided to check the impact of the station location on use of the bicycles. The station was set 200 m from the entrance of the high school. Random sample of the students has been selected. Students were asked about the frequency of using the bicycles. In the middle of the semester the company set the station closer to the entrance. After one month, the same group of students were asked about the frequency of using bicycles again.

Data info:
- variable 1: pretest—ordinal (declared frequency of using the shared bicycles; 1—more than once a day; 2—every day; 3—2–4 times a week; 4—once a week; 5—once a month; 6—less than once a month; 7—never);
- variable 2: posttest—ordinal (declared frequency of using the shared bicycles; 1—more than once a day; 2—every day; 3—2–4 times a week; 4—once a week; 5—once a month; 6—less than once a month; 7—never).

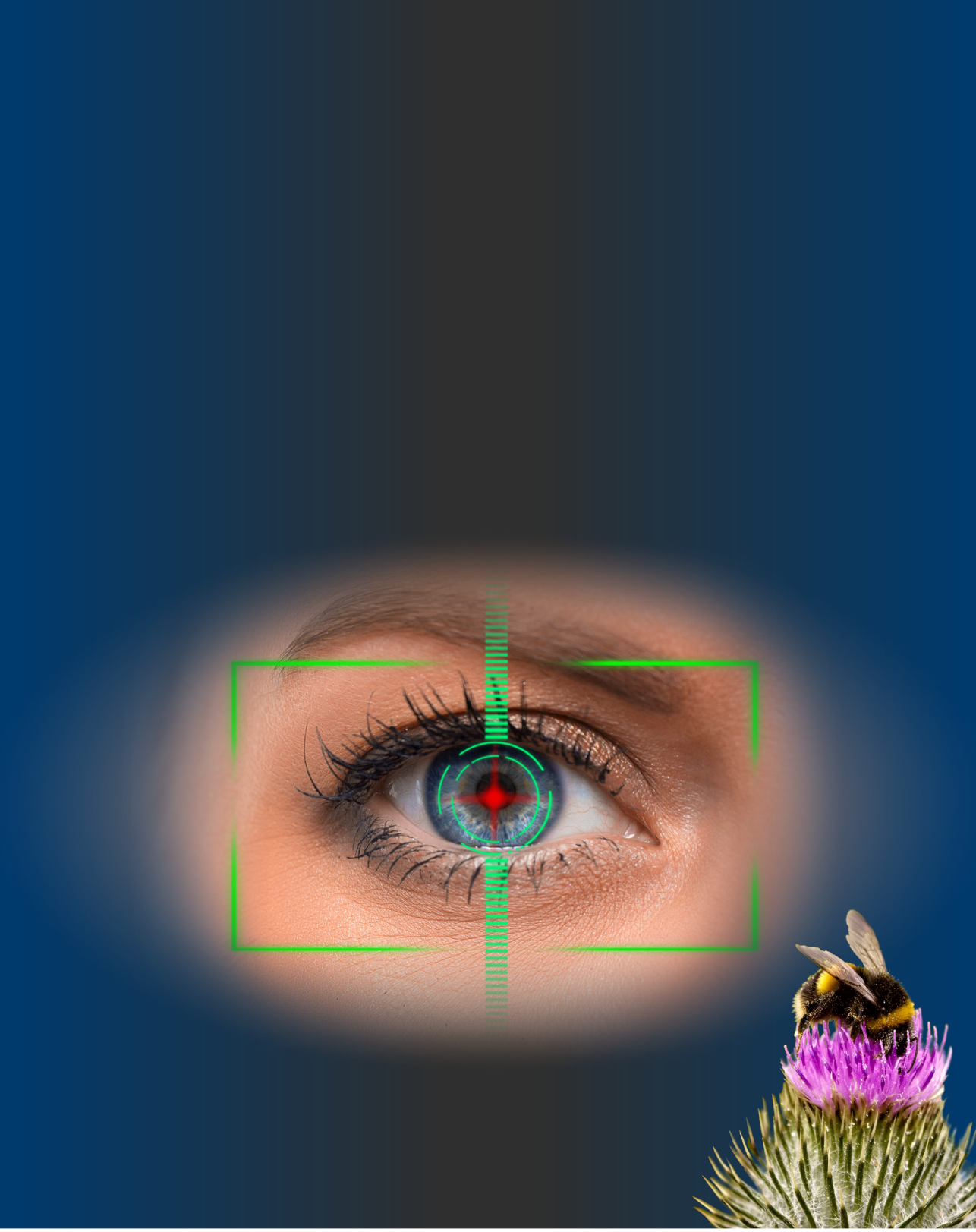
After relocation of the station, the frequency of using the shared bicycles changed significantly $Z(35) = -2.45$, $p = .014$. The students used the shared bicycles more frequent ($Mdn = 4$) compared to the initial location ($Mdn = 5$). However, effect size is rather small ($r = .29$).

References

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics* (5th ed.). Sage edge.

Pallant, J. (2011). *SPSS survival manual: a step by step guide to data analysis using SPSS* (4th ed.). Allen & Unwin.

Verma, J. P., & Abdel-Salam, G. A.-S. (2019). *Testing statistical assumptions in research*. John Wiley & Sons, Inc.



eISBN: 978-83-8211-079-1