

ON SELECTING HYPOTHESIS TESTS FOR GROUP DIFFERENCES

L. Grabinger, J. Mottok

OTH Regensburg (GERMANY)

Abstract

Deciding on the right method of analyzing empirical research data can be difficult – especially when it comes to choosing specific methods of inferential statistics, where there is often not one right way, but a variety of valid options (e.g., using an ANOVA or its non-parametric alternative for data that is not perfectly normally distributed). Novice researchers not only lack the experience to know when a particular hypothesis test is appropriate, but struggle to find suitable literature to familiarize themselves (i.e., literature that is not too superficial, yet comprehensible). With the present article we provide a remedy following the didactic method of *scaffolding*: We present a systematization of the most elementary inferential statistical methods, namely hypothesis tests for group differences. We start by explaining basic terms (e.g., independent variable) and then give step-by-step instructions for choosing a proper hypothesis test based on data properties, implementing it from scratch, and reporting or interpreting its results. With these practical cookbook-like guidelines, this article serves as a concise starting point for young researchers entering the field of empirical research, as a valuable resource for their instructors, and as a basis for automating statistical procedures in a software system.

Keywords: empirical research, data analysis, inferential statistics, guidelines

1 INTRODUCTION

Choosing a proper statistical procedure is a challenge for many researchers – especially for beginners. In this article, we use the didactic method of *scaffolding* [1] to remedy the situation: We present a 4-step-workflow for finding a correct hypothesis test for group differences in a given setting¹ – from the choice of method to the interpretation of results. For the purposes of this article, we restrict ourselves to the case with one independent variable and one (quasi-)metric dependent variable, as this is the most common setting following [2]². Furthermore, we assume typical controlled experiments: We consider two-sided tests for the case of multiple samples.

The article begins with a contextualization within existing work and a brief description of our research methodology. We then cover basic terminology and notation, detail our 4-step-workflow, and conclude by outlining the contribution of the work.

2 RELATED WORK

Various sources can be found that list formulas for statistical methods – either in the form of comprehensive books (e.g., [9]) or thematically tailored papers (e.g., [10]). While the former offer extensive insights, working through them is time-consuming and often difficult, especially for beginners; The latter, on the other hand, are mostly limited to a singular detail (e.g., the computation of effect sizes in [10]). Some sources also offer process diagrams (e.g., [11, p.958][12, p.373][13, p.274]), but these usually cover a broader setting (e.g., also correlation analyses or the case of multiple independent variables [11, p.958]). Accordingly, these diagrams are superficial, e.g., junctions are simply based on the fact whether or not assumptions are met without listing the specific assumptions [11, p.958]. Moreover, even if the assumptions to be tested are detailed, it is not indicated how they can be accessed [12, p.373][13, p.274].

To the best of our knowledge, the present article is the only one that covers the area of two-sided hypothesis tests for group differences with one independent variable and one (quasi-)metric dependent variable in a compact but comprehensive and comprehensible manner.

¹Note that in statistics there is usually not just one correct solution, but a multitude of possibilities.

²This statement has so far only been supported in the case of eye tracking research in the field of software engineering, the authors' main area of research – from modeling languages [3, 4, 5] to programming languages [6, 7, 8].

3 METHODS

All information in this article is based on the authors' expertise and unsystematic literature research using **Google Books** or **Google Scholar**. In order to keep the presentation compact, derivations and background information are omitted. Yet, where applicable, we provide references for further reading.

4 RESULTS

The present section starts with an explanation of some basic terms as well as the notation used. After that, we detail our proposed 4-step-workflow for hypothesis tests for group differences: selecting, conducting, reporting, and interpreting.

4.1 Vocabulary

Designing: In experimental research, two concepts play a central role: *dependent variables* and *independent variables*. The former is the quantity that is actually measured (i.e., the sample data), the latter is the condition that divides this data into groups [14, p.27]. Depending on the *experimental design*, these groups can be *paired* or *unpaired*. In the first case, the individual sample elements of both groups are related to each other [14, p.110], e.g., recorded on the same participant [14, p.138][15, p.307][12, p.264].

Testing: Hypothesis testing begins with the formulation of the (so-called *alternative*) hypothesis to be tested and an opposing *null hypothesis*. The general idea is to mathematically falsify the null hypothesis (i.e., to prove that it is extremely unlikely given the sample at hand), which in turn confirms the original hypothesis [15, p.205][9, p.99]. Hypothesis tests can be conducted without any, with two, or with multiple groups within the data. In the first case, the alternative hypothesis is formulated to compare the considered dependent variable against a fixed value (i.e., *one sample test*); For the latter two, the considered dependent variable is compared between the groups. For the first two cases, one can distinguish between *left-sided*, *right-sided* or *two-sided* tests – depending on whether the alternative hypothesis states that the considered dependent variable is smaller, larger, or simply different than the fixed value or the other group, respectively [15, p.219].

All hypothesis tests run in a similar way. The choice of procedure (e.g., Fisher's ANOVA) determines the *test statistic* formula and its general *probability distribution* for the case that the null hypothesis holds true (e.g., the F-distribution). First, the sample data is used to calculate the respective value of the test statistic and to parameterise its probability distribution by the so-called *degrees of freedom* [9, p.101][15, p.119]. This allows the *p-value* to be calculated as the probability for observing the test statistic found or an even more extreme value in the probability distribution [15, p.228][9, p.107]. With other words, the p-value is the probability that the observation given by the alternative hypothesis is only due to chance – if this probability is small enough (i.e., lower than the so-called *significance level*, usually 0.05), the test is *significant* (i.e., the null hypothesis is rejected) [15, p.230][14, p.77][9, p.107][12, p.176]. However, the test statistic increases with increasing sample size, so that even small effects can lead to significance [14, p.118]. For this reason, we also calculate and report *effect sizes* – measures for the magnitude of the observed effect [14, p.118][12, p.206][15, p.233][9, p.108].

In the case of test procedures for multiple groups, the result provides information on whether the data differs between the groups or not, but not between which groups [14, p.168][15, p.330]. This can be achieved by pairwise tests between two groups each, i.e., *post-hoc tests* [12, p.598][15, p.330][14, p.168]. However, it is necessary to compensate for the fact that the same data is used multiple times (e.g., using data from group A to test against group B and also to test against group C). The simplest solution to this problem is the so-called *Bonferroni correction*, where the resulting p-values are multiplied by the number of tests [15, p.331][12, p.501][9, p.232].

Deciding: Apart from the experimental design (e.g., number or pairing of groups), the decision between particular hypothesis tests is usually based on whether or not the assumptions of normality, homogeneity of variances, or sphericity are met. The first assumption, i.e., *normality*, states that the data roughly follows a normal distribution. *Homogeneity of variances* is given when the variances of the sample values are practically equal between the different groups [15, p.299][12, p.303]. *Sphericity* holds if the variances of the differences between the sample values of each two groups are practically equal [12, p.506][16, p.71].

Table 1. Main notation elements and references for further reading.

Symbol	Explanation	References
x_i	an element in the sample	[19, p.467] ³
$x_{(i)}$	the i -th element in the ascending ordered sample	[9, p.27][17, p.65]
K	the number of groups given by the independent variable	[14, p.166][17, p.69][11, p.410] ⁴
n	the number of elements in a sample	[19, p.467]
n_i	the number of elements in a group	[20, p.66]
m	joint sample size for paired data (i.e., number of participants)	
m_0	the number of pairings with non-zero difference for paired data	
J	number of ties when ranking ⁵	
R_i	rank sum for elements in a group	[15, p.272]

Ranking: For some procedures, we need the concept of *ranking*, i.e., the numbering of sample elements according to their size. If there are identical elements in the data to be ranked (i.e., *ties*), the mean value of the corresponding ranks is assigned to all of them [15, p.268][16, p.27][17, p.162]. The final ranks are then further processed in so-called *rank sums* – the summation of ranks of certain elements that, e.g, elements that belong to the same group [14, p.132][15, p.269][18, p.95][16, p.140].

4.2 Notation

Notation is far from standardized in statistics literature. In the present article, we use the terminology given in Table 1. Additionally, for determining p-values, we make use of some continuous probability distributions, namely the z- or normal distribution [9, p.70][19, p.368][17, p.402], the χ^2 -distribution [9, p.74][17, p.441], the t-distribution [9, p.75][17, p.442], and the F-distribution [9, p.76]. We also consider some discrete distributions, i.e., the distributions of the test statistics of certain hypothesis tests: the U-value of the Mann-Whitney-U test, the W-value of the Wilcoxon signed-rank test, the H-value of the Kruskal-Wallis test, and the S-value of the Friedman test.

4.3 Workflow

We propose a four-step process for performing hypothesis tests for group differences. First, a suitable test procedure is to be chosen (step 1). Then, the test quantities are to be computed (step 2) and reported (step 3). In a final step (step 4), the results must be interpreted properly. On the following pages, we describe these steps in more detail and provide reference material for practical use.

Step 1: Choice of procedures

The process for choosing a suitable hypothesis test is shown in Fig. 1⁶. There, for each decision, the basis for the decision is also listed as one of *data structure*, *Shapiro-Wilk* test, or *Levene's* test. The former states that the decision can be read directly from the data (i.e., the number of groups or the experimental design). The Shapiro-Wilk test checks whether the data follows a normal distribution⁷ (i.e., significance rejects the normality assumption) [21, p.119][14, p.95]; Levene's test checks whether the assumption of homogeneity of variances is violated – here, too, significance rejects the assumption [22, p.289][15, p.299][12, p.303]. Note that the figure distinguishes between *parametric* and *non-parametric* test – the former fulfill certain assumptions (here normality), the latter do not require these assumptions and also work with ordinal data, as they are based on ranked data [14, p.79][9, p.586][16, p.56].

Some of the tests resulting from the figure are available in different variants, i.e., the non-parametric ones and the repeated-measures ANOVA. For the latter, the assumption of sphericity becomes relevant. This is usually checked using Mauchly's test (again, significance rejects sphericity) [16, p.72]; If sphericity does not hold, the degrees of freedom need to be corrected by some computational factor. However, the expressiveness of Mauchly's test is controversial, especially for small sample sizes [16, p.72][23, p.8][24, p.777][12, p.508], which is why corrections are recommended regardless of the result [12, p.508][16, p.72]. For the Mann-Whitney U test, there is the option of standardizing the test statistic and using the normality distribution to extract the p-value – this approximation becomes valid as the sample size

⁶Note that all the procedures presented have a common prerequisite: All elements of the sample must be independent of each other (apart from pairing), e.g., there must not be multiple measurements per participant within one group. With a suitable experimental design, however, this prerequisite is automatically met; Otherwise, it helps to summarize the related data values (e.g., averaging over all measurements for one participant within the group).

⁷Normality can also be accessed visually with histograms [12, p.115][14, p.98] or Q-Q plots [14, p.94].

increases; The exact threshold varies between samples size of around 20 to 40 (i.e., $n_A + n_B = 20$ [15, p.272][12, p.328], $n_A \approx n_B = 20$ [18, p.99], and $n_A > 20$ or $n_B > 20$ [16, p.141]). For strongly differing sample sizes, a *continuity correction* can be applied [16, p.142]. Likewise, the Wilcoxon signed-rank test enables such an approximation, whereby the threshold value ranges between a joint sample size (i.e., m) of 20 [14, p.139][12, p.318] to 50 [16, p.193]; A continuity correction is recommended for a joint sample size between 50 and 60 [16, p.193]. For the Kruskal-Wallis test, there is also the possibility for an approximation – this time, however, it is not the test statistic that is changed, but the probability distribution (i.e., to the χ^2 -distribution). This approximation becomes more accurate with increasing sample size [25, p.205] and is already viable from a sample size of 5 per group (i.e., $n_i = 5$) [12, p.432f][18, p.107][16, p.159]. There is a similar option for the Friedman test; For very small samples (i.e., $K = 3$ and $m \leq 9$ or $K = 4$ and $m \leq 4$ [16, p.204]), one should restrain from the approximation.

Step 2: Realization of procedures

Table 2 provides the formulas for computing all relevant quantities for the test procedures presented⁸; In each case, the formulas are selected that provide the same result as common implementations in the programming language R⁹ Quantities that are defined for several tests, such as degrees of freedom, are given unique indices. All formulas are broken down to such extent that they are based on only a few fundamental variables, i.e., *mean* (see formula 1 following [19, p.490][14, p.32][17, p.59][13, p.118][12, p.54][9, p.25]), *median* (see formula 2 following [15, p.43][17, p.63][12, p.52f][9, p.27]), *standard deviation* (see formula 3 following [19, p.493]), *pooled standard deviation* (see formula 4 following [20, p.44][20, p.67][26, p.8][14, p.115]), and the entries in the *double centered covariance matrix*. The latter is a $K \times K$ symmetric matrix, which is derived from the entries of the covariance matrix (see formula 5 following [19, p.615][14, p.185]) by averaging and summarizing (see formula 6 following [23, p.4]); The non-zero eigenvalues of the matrix are referenced below as λ_i , while the zero eigenvalue is omitted from the calculations.

$$\bar{x} = \hat{\mu} = \frac{\sum_{i=1}^n x_i}{n} \quad (1)$$

$$\tilde{x} = \begin{cases} \frac{1}{2} \cdot (x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}) & n \text{ is an even number} \\ x_{(\frac{n+1}{2})} & n \text{ is an uneven number} \end{cases} \quad (2)$$

$$s = \hat{\sigma} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}} \quad (3)$$

$$s_p = \begin{cases} \sqrt{\frac{(n_A - 1) \cdot s_A^2 + (n_B - 1) \cdot s_B^2}{n_A + n_B - 2}} & K = 2 \\ \sqrt{\frac{\sum_{k=1}^K ((n_k - 1) \cdot s_k^2)}{\sum_{k=1}^K n_k - K}} & K > 2 \end{cases} \quad (4)$$

$$s_{A,B} = \frac{\sum_{i=1}^m (x_{A,i} - \bar{x}_A) \cdot (x_{B,i} - \bar{x}_B)}{m - 1} \quad (5)$$

$$c_{A,B} = s_{A,B} - \frac{\sum_{k=1}^K s_{A,k}}{K} - \frac{\sum_{k=1}^K s_{B,k}}{K} + \frac{\sum_{k=1}^K \sum_{l=1}^K s_{k,l}}{K \cdot K} \quad (6)$$

Since the table conveys information concisely, a few comments should be made beforehand. First of all, most effect sizes are simply referred to by their assigned symbols (e.g., η^2 as "eta squared"); However, certain effect sizes have special names: For t-tests it is *Cohen's d* and *Hedges' g*, for ANOVAs it is *Cohen's f*, for Friedman's ANOVA it is *Kendall's W*, for Wilcoxon signed-rank and Mann-Whitney-U tests it is the *correlation coefficient r*, and for the repeated-measures ANOVA the index p is referred to as *partial*. Similarly, the sphericity corrections of the repeated-measures ANOVA go by *Greenhouse-Geisser* ϵ or *Huynh-Feldt* ϵ , respectively.

⁸For each procedure, if applicable, we present test statistic, degrees of freedom, p-value, effect sizes, and additional elements (e.g., auxiliary variables). If several test statistics are specified, the first is the exact and the second is the approximated.

⁹Some sources deviate from the procedure presented here. For reasons of space, these deviations are largely omitted.

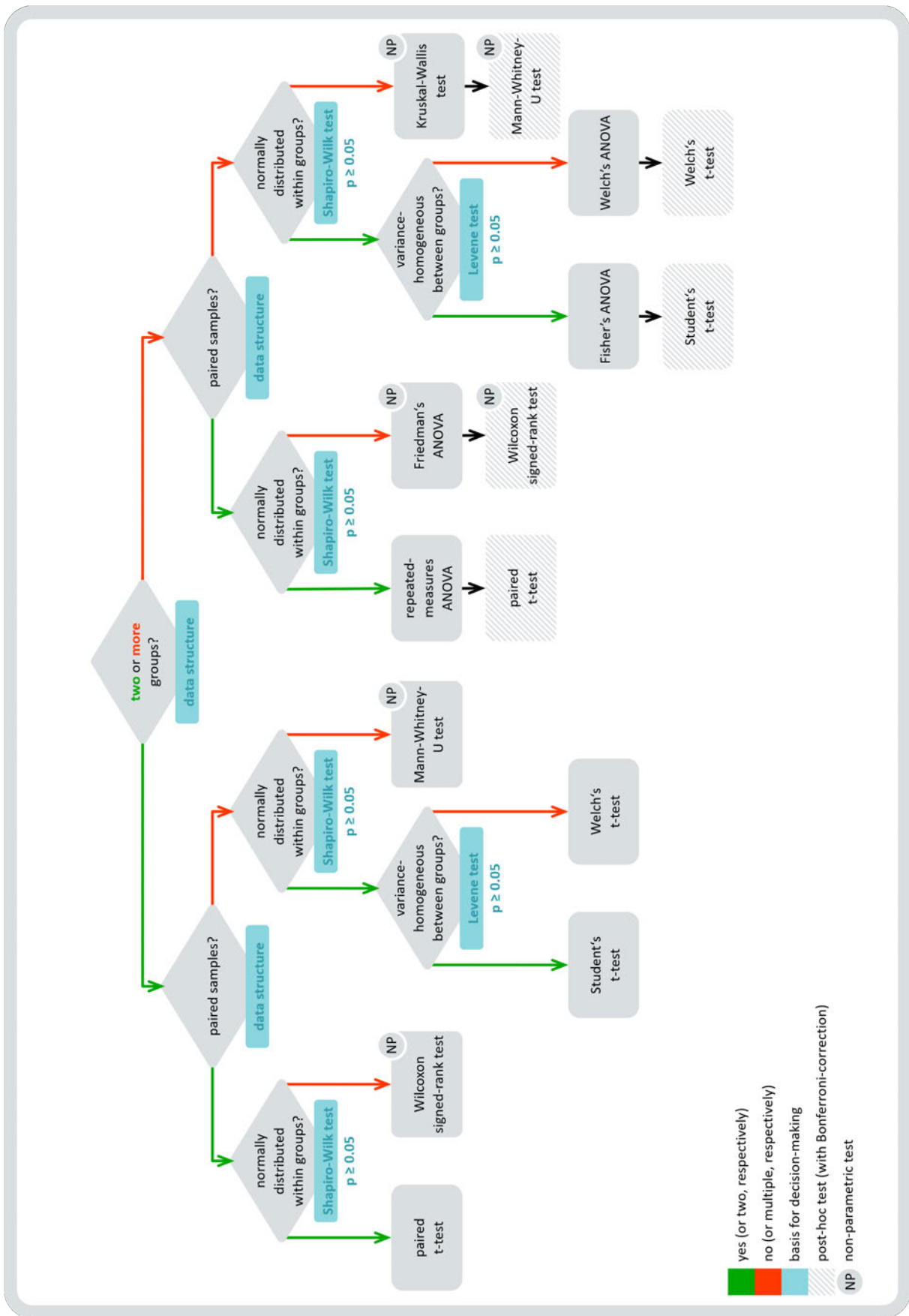


Figure 1. Choice of procedures.

Yet, there are not only comments on the naming, but also on the formulas themselves. For one, the formula given for Hedges' g is an approximation of the actual relationship [20, p.66][26, p.3][27, p.9], which is based on the computationally intensive gamma function [27, p.9]. Secondly, for the Mann-Whitney-U test and the Wilcoxon signed-rank test, the calculation of the effect size is based on the z-value without continuity correction. Third, note that rank sums referenced in the table are calculated differently for the different tests. For Mann-Whitney-U tests, ranking is done jointly for both groups [14, p.132][25, p.116][12, p.326][18, p.95][16, p.140]; Rank sums are formed for each group [14, p.132][15, p.269][18, p.95][16, p.140]; Ties are given by equal values across both groups. For Wilcoxon signed-rank tests, ranking is done for the absolute values of the non-zero differences between the paired elements of both groups [14, p.139][25, p.40][12, p.316][18, p.105][16, p.192]; Rank sums are formed once for positive and once for negative differences [14, p.139][25, p.46][12, p.316][18, p.105][16, p.192]; Ties are given by equal absolute values of differences between paired elements of both groups. For Kruskal-Wallis tests, ranking is done jointly for all groups [14, p.176][25, p.204][12, p.427][18, p.106]; Rank sums are formed for each group [14, p.177][25, p.204][12, p.428][18, p.106]; Ties are given by equal values across all groups. For Friedman's ANOVA, ranking is done separately for each participant [11, p.688][25, p.292][12, p.520][16, p.204]; Rank sums are formed for each group across all participants [11, p.688][25, p.292][12, p.520][16, p.204]; Ties are given by equal values within one participant.

One might notice that the table shows different effect sizes for parametric tests, but only one for each non-parametric test. This is because for the latter, the effect sizes are not so crucial as they can be calculated directly from the test statistic and the experiment characteristics. For the former, it is recommended to use Hedges' g for t-tests, ϵ_p^2 for repeated-measures ANOVA, and ϵ^2 for the remaining.

Table 2. Formulas for particular hypothesis tests and references for further reading.

Element	Formula	References
Student's t-test		
test statistic	$t_{ST} = \frac{\bar{x}_A - \bar{x}_B}{s_p \cdot \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$	[14, p.115][9, p.121][12, p.297]
degrees of freedom	$df_{ST} = n_A + n_B - 2$	[14, p.115][9, p.121]
p-value	$p_{ST} = 2 \cdot P(t(df_{ST}) \leq - t_{ST})$	
effect sizes	$d_{ST} = \frac{ \bar{x}_A - \bar{x}_B }{s_p}$ $g_{ST} = d_{ST} \cdot \frac{4 \cdot df_{ST} - 4}{4 \cdot df_{ST} - 1}$	[26, p.8][27, p.7][20, p.66][12, p.298] [20, p.66][26, p.3][27, p.9]
Welch's t-test		
test statistic	$t_{WT} = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$	[14, p.116][9, p.123][12, p.297]
degrees of freedom	$df_{WT} = \frac{(\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B})^2}{\frac{1}{n_A - 1} \cdot (\frac{s_A^2}{n_A})^2 + \frac{1}{n_B - 1} \cdot (\frac{s_B^2}{n_B})^2}$	[14, p.116][27, p.15][9, p.123][12, p.303]
p-value	$p_{WT} = 2 \cdot P(t(df_{WT}) \leq - t_{WT})$	
effect sizes	$d_{WT} = \frac{ \bar{x}_A - \bar{x}_B }{\sqrt{\frac{s_A^2 + s_B^2}{2}}}$ $g_{WT} = d_{WT} \cdot \frac{4 \cdot df_{WT} - 4}{4 \cdot df_{WT} - 1}$	[27, p.16] [27, p.17]
Paired t-test		
test statistic	$t_{PT} = \frac{\frac{1}{m} \cdot \sum_{i=1}^m (x_{A,i} - x_{B,i})}{\sqrt{\frac{1}{(m-1) \cdot m} \cdot \sum_{i=1}^m ((x_{A,i} - x_{B,i}) - \frac{1}{m} \cdot \sum_{i=1}^m (x_{A,i} - x_{B,i}))^2}}$	[9, p.125][14, p.127][12, p.279]
degrees of freedom	$df_{PT} = m - 1$	[14, p.127][9, p.125][12, p.279]
p-value	$p_{PT} = 2 \cdot P(t(df_{PT}) \leq - t_{PT})$	
effect sizes	$d_{PT} = \frac{ \frac{1}{m} \cdot \sum_{i=1}^m (x_{A,i} - x_{B,i}) }{\sqrt{\frac{1}{m-1} \cdot \sum_{i=1}^m ((x_{A,i} - x_{B,i}) - \frac{1}{m} \cdot \sum_{i=1}^m (x_{A,i} - x_{B,i}))^2}}$ $g_{PT} = d_{PT} \cdot \frac{4 \cdot df_{PT} - 4}{4 \cdot df_{PT} - 1}$	[26, p.8][27, p.7][20, p.66][12, p.298]
Fisher's ANOVA		
sums of squares	$SS_{between,FA} = \sum_{k=1}^K n_k \cdot (\bar{x}_k - \bar{x})^2$ $SS_{within,FA} = \sum_{k=1}^K \sum_{i=1}^{n_k} (x_{i,k} - \bar{x}_k)^2$ $SS_{total,FA} = \sum_{k=1}^K \sum_{i=1}^{n_k} (x_{i,k} - \bar{x})^2$ $SS_{total,FA} = SS_{between,FA} + SS_{within,FA}$	[15, p.325][9, p.207][12, p.401] [15, p.326][9, p.208][12, p.401] [15, p.324][9, p.207][12, p.401] [15, p.327][9, p.209][12, p.401]
mean squares	$MS_{between,FA} = \frac{SS_{between,FA}}{df_{between,FA}}$ $MS_{within,FA} = \frac{SS_{within,FA}}{df_{within,FA}}$	
test statistic	$F_{FA} = \frac{MS_{between,FA}}{MS_{within,FA}}$	[15, p.327][9, p.209]
degrees of freedom	$df_{between,FA} = K - 1$ $df_{within,FA} = n - K$	[15, p.325][9, p.208][12, p.404ff] [15, p.325][9, p.208][12, p.404ff]
p-value	$p_{FA} = P(F(df_{between,FA}, df_{within,FA}) \geq F_{FA})$	[12, p.408f]
effect sizes	$\eta_{FA}^2 = \frac{SS_{between,FA}}{SS_{total,FA}}$	[28, p.5][26, p.5][20, p.281]

$$f = \sqrt{\frac{\eta_{FA}^2}{1 - \eta_{FA}^2}} \quad [28, p.5][20, p.281][12, p.411]$$

$$\omega_{FA}^2 = \frac{SS_{between,FA} - df_{between,FA} \cdot MS_{within,FA}}{SS_{total,FA} + MS_{within,FA}} \quad [28, p.5][26, p.6][29, p.130][12, p.412]$$

$$\epsilon_{FA}^2 = \frac{SS_{between,FA} - df_{between,FA} \cdot MS_{within,FA}}{SS_{total,FA}} \quad [28, p.5][29, p.130][12, p.410]$$

Welch's ANOVA

auxiliary variable	$w_k = \frac{n_k}{s_k^2}$	[30, p.217]
sums of squares	see Fisher's ANOVA	
mean squares	see Fisher's ANOVA	
test statistic	$F_{WA} = \frac{\frac{1}{K-1} \cdot \sum_{k=1}^K w_k \cdot (\bar{x}_k - \frac{\sum_{l=1}^K w_l \cdot \bar{x}_l}{\sum_{l=1}^K w_l})^2}{1 + \frac{2 \cdot (K-2)}{3 \cdot df_{WA}}}$	[30, p.217]
degrees of freedom	$df_{WA} = \frac{K^2 - 1}{3 \cdot \sum_{k=1}^K (1 - \frac{w_k}{\sum_{k=1}^K w_k})^2 \cdot \frac{1}{n_k - 1}}$	[30, p.217]
effect sizes	$p_{WA} = P(F(df_{between,FA}, df_{WA})) \geq F_{WA}$ see Fisher's ANOVA	[12, p.408f]

Repeated-measures ANOVA

sphericity corrections	$\epsilon_{GG} = \frac{(\sum_{k=1}^K c_{k,k})^2}{(K-1) \cdot \sum_{k=1}^K \sum_{l=1}^K c_{k,l}^2}$ $\epsilon_{HF} = \frac{\epsilon_{GG}}{(K-1) \cdot (m-1 - (K-1) \cdot \epsilon_{GG})}$	[23, p.4] [23, p.6]
sums of squares	$SS_{between,RA} = K \cdot \sum_{i=1}^m (\bar{x}_i - \bar{x})^2$ $SS_{within,RA} = \sum_{k=1}^K \sum_{i=1}^m (x_{i,k} - \bar{x}_i)^2$ $SS_{repeated} = K \cdot \sum_{i=1}^m (\bar{x}_i - \bar{x})^2$ $SS_{within,RA} = SS_{repeated} + SS_{residual}$	[12, p.493][9, p.286] [12, p.495][9, p.286] [12, p.496][9, p.286] [9, p.286][12, p.492]
mean squares	$MS_{repeated} = \frac{SS_{repeated}}{df_{repeated}}$ $MS_{residual} = \frac{SS_{residual}}{df_{residual}}$	
test statistic	$F_{RA} = \frac{MS_{repeated}}{MS_{residual}}$	[15, p.327][9, p.209]
degrees of freedom	$df_{repeated} = \begin{cases} K-1 & \text{for given sphericity} \\ (K-1) \cdot \epsilon_{GG} & \text{for } \epsilon_{GG} < 0.75 \\ (K-1) \cdot \epsilon_{HF} & \text{for } \epsilon_{GG} \geq 0.75 \end{cases}$ $df_{residual} = \begin{cases} (K-1) \cdot (m-1) & \text{for given sphericity} \\ (K-1) \cdot (m-1) \cdot \epsilon_{GG} & \text{for } \epsilon_{GG} < 0.75 \\ (K-1) \cdot (m-1) \cdot \epsilon_{HF} & \text{for } \epsilon_{GG} \geq 0.75 \end{cases}$	[9, p.301] [9, p.301]
p-value	$p_{RA} = P(F(df_{repeated}, df_{residual})) \geq F_{RA}$	[12, p.408f]
effect sizes	$\eta_{RA}^2 = \frac{SS_{repeated}}{SS_{total,RA}}$ $f_{RA} = \sqrt{\frac{\eta_{RA}^2}{1 - \eta_{RA}^2}}$ $\omega_{RA}^2 = \frac{SS_{repeated} - df_{repeated} \cdot MS_{residual}}{SS_{total,RA} + MS_{between,RA}}$ $\epsilon_{RA}^2 = \frac{SS_{repeated} - df_{repeated} \cdot MS_{residual}}{SS_{total,RA}}$ $\eta_p^2 = \frac{SS_{repeated}}{SS_{repeated} + SS_{residual}}$ $\omega_p^2 = \frac{SS_{repeated} - df_{repeated} \cdot MS_{residual}}{SS_{repeated} + (n - df_{repeated}) \cdot MS_{residual}}$ $\epsilon_p^2 = \frac{SS_{repeated} - df_{repeated} \cdot MS_{residual}}{SS_{repeated} + SS_{residual}}$	[31, p.276][26, p.6] [31, p.276] [31, p.276] [31, p.276] [31, p.276] [31, p.276]

Mann-Whitney-U test

test statistic	$U_{MW} = n_A \cdot n_B + \frac{n_B \cdot (n_B + 1)}{2} - R_B$ $z_{MW} = \begin{cases} \frac{ U_{MW} - \frac{n_A \cdot n_B}{2} - 0.5}{\sqrt{\frac{n_A \cdot n_B \cdot (n+1)}{12}}} & \text{without ties} \\ \frac{ U_{MW} - \frac{n_A \cdot n_B}{2} - 0.5}{\sqrt{\frac{n_A \cdot n_B \cdot (n^3 - n - \frac{1}{12} \cdot \sum_{j=1}^J n_j \cdot (n_j^2 - 1))}{n \cdot (n-1)}}} & \text{with ties} \end{cases}$	continuity correction continuity correction
p-value	$p_{MW} = \begin{cases} 2 \cdot P(U \leq \min(U_{MW}, n_A \cdot n_B - U_{MW})) & \text{exact} \\ 2 \cdot P(z \leq - z_{MW}) & \text{approximated} \end{cases}$	[14, p.133][15, p.272][12, p.329][18, p.99][16, p.142] [16, p.142][25, p.117]
effect sizes	$r_{MW} = \frac{z_{MW}}{\sqrt{n}}$	[13, p.238][14, p.137][10, p.23]

Wilcoxon signed-rank test

test statistic	$W_{WS} = R_+$ $z_{WS} = \begin{cases} \frac{ W_{WS} - \frac{m_0 \cdot (m_0 + 1)}{4} - 0.5}{\sqrt{\frac{m_0 \cdot (m_0 + 1) \cdot (2 \cdot m_0 + 1)}{24}}} & \text{without ties} \\ \frac{ W_{WS} - \frac{m_0 \cdot (m_0 + 1)}{4} - 0.5}{\sqrt{\frac{m_0 \cdot (m_0 + 1) \cdot (2 \cdot m_0 + 1) - \frac{1}{2} \cdot \sum_{j=1}^J n_j \cdot (n_j^2 - 1)}{24}}} & \text{with ties} \end{cases}$	continuity correction continuity correction
p-value	$p_{WS} = \begin{cases} 2 \cdot P(W \leq \min(W_{WS}, \frac{m_0 \cdot (m_0 + 1)}{2} - W_{WS})) & \text{exact} \\ 2 \cdot P(z \leq - z_{WS}) & \text{approximated} \end{cases}$	[12, p.319][16, p.196] [25, p.41]
effect sizes	$r_{WS} = \frac{z_{WS}}{\sqrt{m}}$	[10, p.23][12, p.321]

Kruskal-Wallis test

test statistic	$H_{KW} = \begin{cases} \frac{12}{n \cdot (n+1)} \cdot \sum_{k=1}^K \frac{R_k^2}{n_k} - 3 \cdot (n+1) & \text{without ties} \\ \frac{12}{n \cdot (n+1)} \cdot \sum_{k=1}^K \frac{R_k^2}{n_k} - 3 \cdot (n+1) & \text{with ties} \end{cases}$	[25, p.204][12, p.430][18, p.107][16, p.158]
degrees of freedom	$df_{KW} = K - 1$	[25, p.204][12, p.433][18, p.107][16, p.158]
p-value	$p_{KW} = \begin{cases} P(H \geq H_{KW}) & \text{exact} \\ P(\chi^2(df_{KW}) \geq H_{KW}) & \text{approximated} \end{cases}$	[25, p.204]
effect sizes	$\eta_{KW}^2 = \frac{H - K + 1}{n - K}$	[10, p.24]

Friedman's ANOVA

test statistic	$\chi_{FA}^2 = \begin{cases} \frac{12}{m \cdot K \cdot (K+1)} \cdot \sum_{k=1}^K R_k^2 - 3 \cdot m \cdot (K+1) & \text{without ties} \\ \frac{12}{m \cdot K \cdot (K+1)} \cdot \sum_{k=1}^K R_k^2 - 3 \cdot m \cdot (K+1) & \text{with ties} \end{cases}$	[11, p.689][25, p.292][12, p.521][16, p.204]
degrees of freedom	$df_{FA} = K - 1$	[11, p.689][25, p.293][12, p.522][16, p.205]
p-value	$p_{FA} = \begin{cases} P(S \geq \chi_{FA}^2) & \text{exact} \\ P(\chi^2(df_{FA}) \geq \chi_{FA}^2) & \text{approximated} \end{cases}$	[25, p.292]
effect sizes	$W_{FA} = \frac{\chi^2}{m \cdot (K-1)}$	[10, p.24]

Shapiro-Wilk test

auxiliary variables	$a_i = \begin{cases} a_n & i = n \\ a_{n-1} & i = n - 1 \text{ and } n > 5 \\ -a_n & i = 1 \\ -a_{n-1} & i = 2 \text{ and } n > 5 \\ \sqrt{\frac{1 - 2 \cdot a_n^2}{\sum_{i=1}^n m_i \cdot m_i - 2 \cdot m_n^2}} \cdot m_i & i \in [2, n - 1] \text{ and } n \leq 5 \\ \sqrt{\frac{1 - 2 \cdot a_n^2 - 2 \cdot a_{n-1}^2}{\sum_{i=1}^n m_i \cdot m_i - 2 \cdot m_n^2 - 2 \cdot m_{n-1}^2}} \cdot m_i & i \in [3, n - 2] \text{ bei } n > 5 \end{cases}$	[21, p.117]
	$a_n = c_n + 0,221157 \cdot x - 0,147981 \cdot x^2 - 2,071190 \cdot x^3 + 4,434685 \cdot x^4 - 2,706056 \cdot x^5$	[21, p.117]
	$a_{n-1} = c_{n-1} + 0,042981 \cdot x - 0,293762 \cdot x^2 - 1,752461 \cdot x^3 + 5,682633 \cdot x^4 - 3,582663 \cdot x^5$	[21, p.117]
	$c_i = \frac{1}{\sqrt{\sum_{i=1}^n m_i \cdot m_i}} \cdot m_i$	[21, p.117]
	$m_i = \Phi^{-1}\left(\frac{i - 0,375}{n + 0,25}\right)$	[21, p.117]
	$x = \frac{1}{\sqrt{n}}$	[21, p.117]
	$w_1 = \frac{6}{\pi} \cdot (\sin^{-1}(\sqrt{W_{SW}}) - \sin^{-1}(\sqrt{0,75}))$	[21, p.118]
	$w_2 = -\ln(-2,273 + 0,459 \cdot n - \ln(1 - W_{SW}))$	[21, p.118]
	$\mu_2 = 0,5440 - 0,39978 \cdot n + 0,025054 \cdot n^2 - 0,0006714 \cdot n^3$	[21, p.118]
	$\sigma_2 = e^{1,3822 - 0,77857 \cdot n + 0,062767 \cdot n^2 - 0,0020322 \cdot n^3}$	[21, p.118]
	$w_3 = \ln(1 - W_{SW})$	[21, p.118]
	$\mu_3 = -1,5861 - 0,31082 \cdot y - 0,083751 \cdot y^2 + 0,0038915 \cdot y^3$	[21, p.118]
	$\sigma_3 = e^{-0,4803 - 0,082676 \cdot y + 0,0030302 \cdot y^2}$	[21, p.118]
	$y = \ln(n)$	[21, p.118]
test statistic	$W_{SW} = \frac{(\sum_{i=1}^n a_i \cdot x(i))^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$	[21, p.117]
p-value	$p_{SW} = \begin{cases} w_1 & n = 3 \\ P(z \geq \frac{w_2 - \mu_2}{\sigma_2}) & 4 \leq n \leq 12 \\ P(z \geq \frac{w_3 - \mu_3}{\sigma_3}) & 12 \leq n \leq 2000 \end{cases}$	[21, p.118]

Levene's test

auxiliary variable	$z_{k,i} = x_{k,i} - \bar{x}_k $	[22, p.289]
test statistic	$L_{LE} = \frac{(n-K) \cdot \sum_{k=1}^K n_k \cdot (\bar{z}_k - \bar{z})^2}{(K-1) \cdot \sum_{k=1}^K \sum_{i=1}^{n_k} (z_{k,i} - \bar{z}_k)^2}$	[22, p.289]
degrees of freedom	$df_{LE,1} = K - 1$	[22, p.289]
	$df_{LE,2} = n - K$	[22, p.289]
p-value	$p_{LE} = P(F(df_{LE,1}, df_{LE,2}) \geq L_{LE})$	[22, p.289]

Mauchly's test

auxiliary variable	$d = (1 - \frac{2 \cdot (K-1)^2 + (K-1) + 2}{6 \cdot (K-1) \cdot (m-1)})$	[24, p.777][32, p.237]
test statistic	$W_{MA} = \frac{\prod_{\lambda_j \neq 0} \lambda_j}{(\frac{1}{k-1} \cdot \sum_{j=1}^k \lambda_j)^{k-1}}$	[23, p.8]
	$\chi_{MA}^2 = -(m-1) \cdot d \cdot \ln(W_{MA})$	[23, p.8]
degrees of freedom	$df_{MA} = \frac{(K-1) \cdot K}{2} - 1$	[24, p.777][32, p.237]
p-value	$p = P(\chi^2(df_{MA}) \geq \chi_{MA}^2)$	[23, p.8][24, p.777][32, p.237]

Step 3: Reporting of results

The results computed are typically reported as depicted in Fig. 2: The test statistic with degrees of freedom, followed by the p-value and an effect size. In accordance with the APA standard [33], the p-value is reported with three decimal places and without leading zero (e.g., $p = .003$); All other quantities are given with two decimal places (e.g., $\eta^2 = .45$) with a leading zero whenever the absolute value of the

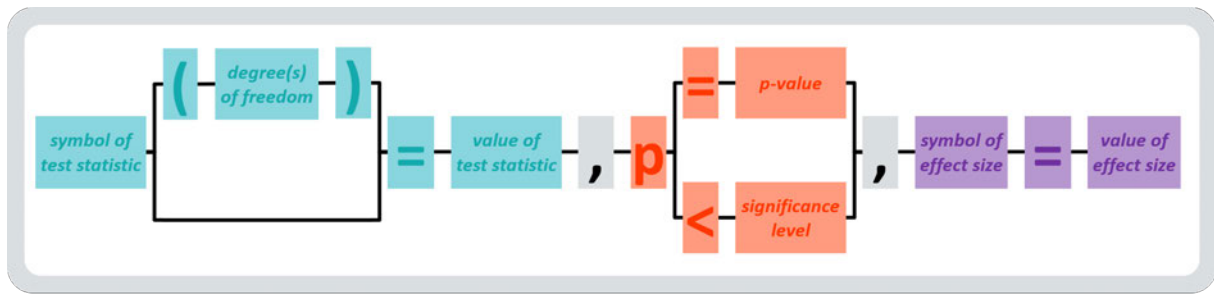


Figure 2. Reporting of results.

Table 3. Interpretation of effect size values and references for further reading.

Quantity	No effect	Small effect	Moderate effect	Large effect	References
d	under 0.20	between 0.20 and 0.50	between 0.50 and 0.80	over 0.80	[20, p.25][26, p.3][15, p.234]
ϵ^2	under 0.01	between 0.01 and 0.06	between 0.06 and 0.14	over 0.14	[26, p.7][20, p.285][12, p.411]
f	under 0.10	between 0.10 and 0.25	between 0.25 and 0.37	over 0.37	[20, p.285]
g	under 0.20	between 0.20 and 0.50	between 0.50 and 0.80	over 0.80	[20, p.25][26, p.3][15, p.234]
η^2	under 0.01	between 0.01 and 0.06	between 0.06 and 0.14	over 0.14	[26, p.7][20, p.285][12, p.411][10, p.24]
r	under 0.10	between 0.10 and 0.30	between 0.30 and 0.50	over 0.50	[20, p.79][13, p.152]
W	under 0.10	between 0.10 and 0.30	between 0.30 and 0.50	over 0.50	[10, p.24]
ω^2	under 0.01	between 0.01 and 0.06	between 0.06 and 0.14	over 0.14	[26, p.7][20, p.285][12, p.411]

quantity can exceed the value one (i.e., the leading zero might be omitted from the effect sizes ϵ^2 , η^2 , r, W, and ω^2). Note that in the figure, there are two junctions. First, if given, the degrees of freedom are reported comma-separated and enclosed in parentheses directly after the symbol of the test statistic – in the same order as in Table 2 for the p-value. Second, if the p-value is below the significance level or below 0.000, it is not stated exactly, but as $p < \text{significance level}$ or $p < .000$, respectively.

Step 4: Interpretation of results

Although the test quantities (i.e., test statistics and degrees of freedom) are normally reported, they are irrelevant for the interpretation of the result. The interpretation is based on the p-value and the value of the effect size. The former indicates whether the result is significant or not, i.e., whether the assumption could be confirmed or not. In other words, if the p-value of a hypothesis test for group differences is less than the significance level (usually 0.05), there is a difference between the groups; if the p-value of Shapiro-Wilk, Levene's, or Mauchly's test is less than 0.05, the assumptions of normality, homogeneity of variances, or sphericity are rejected. In contrast, as the name suggests, the effect size quantifies the extent of the effect. The respective reference points for categorization as a small, medium, or large effect for the individual effect sizes are compiled in Table 3. Note that the interpretation of Kendall's W is far from consistent throughout the sources. Some (e.g., R's **effectsize** package) refer to thresholds presented in [34, p.165], yet even the authors call them arbitrary; Others (e.g., R's **rstatix** package) interpret it like the correlation coefficient as it also ranges between zero and one [10, p.24].

5 CONCLUSION

This article provides practical, cookbook-style instructions for selecting a proper hypothesis test for group differences. For young researchers, it is primarily an introduction to statistics, but it can also serve to take away their fear of this discipline by showing that it can be broken down into simple steps. For instructors or researchers, it offers as a concise source of reference as well as a model for dealing with complex topics in teaching. For us, the work is the basis for the automation of statistical procedures in a software system (see [35][36]).

ACKNOWLEDGEMENTS

The paper is supported by the 'German Federal Ministry of Education and Research' (BMBF) within the funding project HASKI (FKZ: 16DHBK1035). We would like to thank Prof. Dr. Sven Hilbert from the University of Regensburg for his insights and expertise as well as our colleagues from the OTH Regensburg for many fruitful discussions on the topic.

REFERENCES

- [1] J. Arnold, K. Kremer, and J. Mayer, "Scaffolding beim Forschenden Lernen: Eine empirische Untersuchung zur Wirkung von Lernunterstützungen," *Zeitschrift für Didaktik der Naturwissenschaften*, vol. 23, pp. 21–37, 11 2017.
- [2] L. Grabinger, F. Hauser, C. Wolff, and J. Mottok, "On eye tracking in software engineering," *SN COMPUT. SCI.*, vol. 5, pp. 1–20, 7 2024.
- [3] L. Grabinger, F. Hauser, and J. Mottok, "Assessing the presentation of causal graphs and an application of gestalt principles with eye tracking," in *Proceedings of the 29th IEEE International Conference on Software Analysis, Evolution and Reengineering*, SANER '22, (New York, NY, USA), pp. 1267–1274, IEEE, 2022.
- [4] L. Grabinger, F. Hauser, and J. Mottok, "On the perception of graph layouts," *J. Softw.*, vol. 36, pp. 1–18, 5 2024.
- [5] L. Grabinger, F. Hauser, and J. Mottok, "Evaluating graph-based modeling languages," in *5th European Conference on Software Engineering Education (ECSEE 2023)*, pp. 120–129, ACM, 6 2023.
- [6] F. Hauser, L. Grabinger, J. Mottok, and H. Gruber, "Visual expertise in code reviews: Using holistic models of image perception to analyze and interpret eye movements," in *2023 Symposium on Eye Tracking Research and Applications (ETRA 2023)*, pp. 1–7, ACM, 5 2023.
- [7] A. Homann, L. Grabinger, F. Hauser, and J. Mottok, "An eye tracking study on misra c coding guidelines," in *5th European Conference on Software Engineering Education (ECSEE 2023)*, pp. 130–137, ACM, 6 2023.
- [8] F. Hauser, L. Grabinger, T. Ezer, J. Mottok, and H. Gruber, "Analyzing and interpreting eye movements in c++: Using holistic models of image perception," in *Proceedings of the 2024 Symposium on Eye Tracking Research and Applications (ETRA 2024)*, pp. 1–7, ACM, 6 2024.
- [9] J. Bortz and C. Schuster, *Statistik für Human- und Sozialwissenschaftler mit 163 Tabellen*. Springer, 7 ed., 2010.
- [10] M. Tomczak and E. Tomczak, "The need to report effect size estimates revisited. an overview of some recommended measures of effect size," *Trends in Sport Sciences (Trends Sport. Sci.)*, vol. 1, pp. 19–25, 3 2014.
- [11] A. Field, J. Miles, and Z. Field, *Discovering Statistics Using R*. SAGE Publications, 2012.
- [12] M. Bühner and M. Ziegler, *Statistik für Psychologen und Sozialwissenschaftler*. Pearson, 2 ed., 2017.
- [13] A. Field and G. Hole, *How to Design and Report Experiments*. SAGE Publications, 1 ed., 2002.
- [14] M. Jesussek and H. Volk-Jesussek, *Statistik leichtgemacht: Eine verständliche Einführung*. DATAtab, 3 ed., 2023.
- [15] P. Planing, *Statistik Grundlagen*. Planing Publishing, 2022.
- [16] J. Bortz and G. Lienert, *Kurzgefasste Statistik für die Klinische Forschung*. Springer Medizin Verlag, 3 ed., 2008.
- [17] T. Stocker and I. Steinke, *Statistik: Grundlagen und Methodik*. De Gruyter, 2 ed., 2022.
- [18] B. Rasch, M. Friese, W. Hofmann, and E. Naumann, *Quantitative Methoden 2: Einführung in die Statistik für Psychologen und Sozialwissenschaftler*. Springer, 4 ed., 2014.
- [19] L. Papula, *Mathematik für Ingenieure und Naturwissenschaftler: Band 3*. Vieweg, 2 ed., 1997.
- [20] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 2 ed., 1988.

- [21] P. Royston, "Approximating the shapiro-wilk w-test for non-normality," *Statistics and Computing (Stat. Comput.)*, vol. 2, pp. 117–119, 9 1992.
- [22] T.-S. Lim and W.-Y. Loh, "A comparison of tests of equality of variances," *Computational Statistics & Data Analysis (Comput. Stat. Data Anal.)*, vol. 22, pp. 287–301, 7 1996.
- [23] H. Abdi, "The greenhouse-geisser correction," in *Encyclopedia of Research Design* (N. Salkind, ed.), pp. 544–548, SAGE Publications, 2010.
- [24] S. Moulton, "Mauchly test," in *Encyclopedia of Research Design* (N. Salkind, ed.), pp. 776–777, SAGE Publications, 2010.
- [25] M. Hollander, D. Wolfe, and E. Chicken, *Nonparametric Statistical Methods*. John Wiley & Sons, 3 ed., 2015.
- [26] D. Lakens, "Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and anovas," *Frontiers in Psychology (Front. Psychol.)*, vol. 4, pp. 1–12, 11 2013.
- [27] M. Delacre, D. Lakens, C. Ley, L. Liu, and C. Leys, "Why hedges' g*s based on the non-pooled standard deviation should be reported with welch's t-test," *PsyArXiv*, pp. 1–46, 5 2021.
- [28] C. Albers and D. Lakens, "When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias," *PsyArXiv*, pp. 1–37, 7 2017.
- [29] K. Okada, "Is omega squared less biased? a comparison of three major effect size indices in one-way anova," *Behaviormetrika*, vol. 40, pp. 129–147, 7 2013.
- [30] M. Mendes and E. Akkartal, "Comparison of anova f and welch tests with their respective permutation versions in terms of type i error rates and test power," *Kafkas Universitesi Veteriner Fakultesi Dergisi*, vol. 16, pp. 711–716, 1 2010.
- [31] S. Olejnik and J. Algina, "Measures of effect size for comparative studies: Applications, interpretations, and limitations," *Contemporary Educational Psychology (Contemp. Educ. Psychol.)*, vol. 25, pp. 241–286, 7 2000.
- [32] J. E. Cornell, A. L. M. Memorial, V. Hospital, D. M. Young, S. L. Seaman, and R. E. Kirk, "Power comparisons of eight tests for sphericity in repeated measures designs," *Journal of Educational Statistics (J. Educ. Stat.)*, vol. 17, pp. 233–249, 9 1992.
- [33] A. P. Association, *Publication Manual of the American Psychological Association: The Official Guide to Apa Style*. American Psychological Association, 2020.
- [34] R. Landis and G. Koch, "The measurement of observer agreement for categorical data," *Source: Biometrics*, vol. 33, pp. 159–174, 3 1977.
- [35] L. Grabinger and J. Mottok, "Statistical analysis of eye movement data for beginners," in *Proceedings of Mensch und Computer 2024*, MuC '24, (New York, NY, USA), ACM, 2024.
- [36] L. Grabinger, T. Ezer, F. Hauser, and J. Mottok, "The impact of eyenalyzer," in *Proceedings of the 17th annual International Conference of Education, Research and Innovation (ICERI '24)*, (Valencia, Spain), pp. 1–7, IATED, 2024.