FRANKFURT UNIVERSITY OF APPLIED SCIENCES

How to Evaluate Results

Human-Computer Interaction Exercise

Prof. Dr. Valentin Schwind

Results and the Research Question

- **All results need a research question**
  - Research questions are often related to a paradigm or theoretical framework
    - e.g., “Is there an Uncanny Valley of animals?” [1]
  - A hypothesis asks for:
    - e.g., „There is a significant dip in perceived familiarity at higher levels fo realism of dogs, cats, …“
- **A research question is a question** (no theory, no answer, no assumption,…)
  - In **quantitative studies, you have a hypothesis is your answer on that research question**
    - „Explorative“ user studies can have multiple hypotheses
  - In **qualitative studies** you have **no hypothesis (yet)**
  - In **mixed methods studies** (quantitative and qualitative methods), you can combine them!

[1] V. Schwind, K. Leicht, S. Jäger, K. Wolf, N. Henze, Is there an uncanny valley of virtual animals? A quantitative and qualitative investigation, International Journal of Human-Computer Studies,

Research Question

- Recap: **A RQ must ask for new knowledge**
  - You write the results because you are the only one with this knowledge!
- It can be
  - answered **in whole**
  - answered **in part** or under certain circumstances
  - rejected as **unanswerable**
  - only an **apparent** problem
  - question to **a theory**

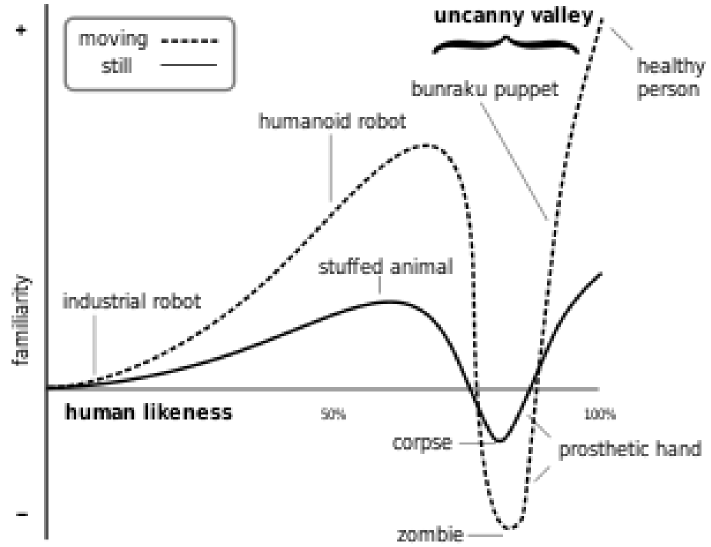


Image from https://en.wikipedia.org/wiki/Uncanny_valley
Smurrayinchester - self-made, based on image by Masahiro Mori and Karl MacDorman at http://www.androidscience.com/theuncannyvalley/proceedings2005/uncannyvalley.html

[1] V. Schwind, K. Leicht, S. Jäger, K. Wolf, N. Henze, Is there an uncanny valley of virtual animals? A quantitative and qualitative investigation, International Journal of Human-Computer Studies,

Hypothesis vs Theory

- A **theory**…
  - is an abstract and generalized thinking about a phenomenon
  - contains a group of logical explanations based on empirical data
- A **hypothesis**…
  - is a proposed explanation (for a phenomenon)
  - is a logical consequence („if… then“…)
  - can be tested and falsified
- A ***working* hypothesis**…
  - is your hypothesis that is *provisionally* accepted as a basis for further research
- What is a hypothesis?

Internal & External Validity

- **Internal Validity**
  - **Quality criteria**: **objectivity**, **reproducibility**, **documentation**, and **elimination of confounds**
  - **High**, when there are **no alternative explanations for your results**
    - The variation of your dependent variable is caused by the variation of your independent variable
  - **Low**, when there when experimental effects can be explained otherwise
    - The variation of your dependent variable can by explained by the variation of confounds
- **External Validity** / Ecological validity
  - The extent to which **results can be generalized**
  - **High**, when results of the study can be transferred to the real world
    - e.g., does the sample represent the general population?
  - **Low** when the results cannot be applied to the population or real-life situations outside of the research setting

Internal vs. External Validity

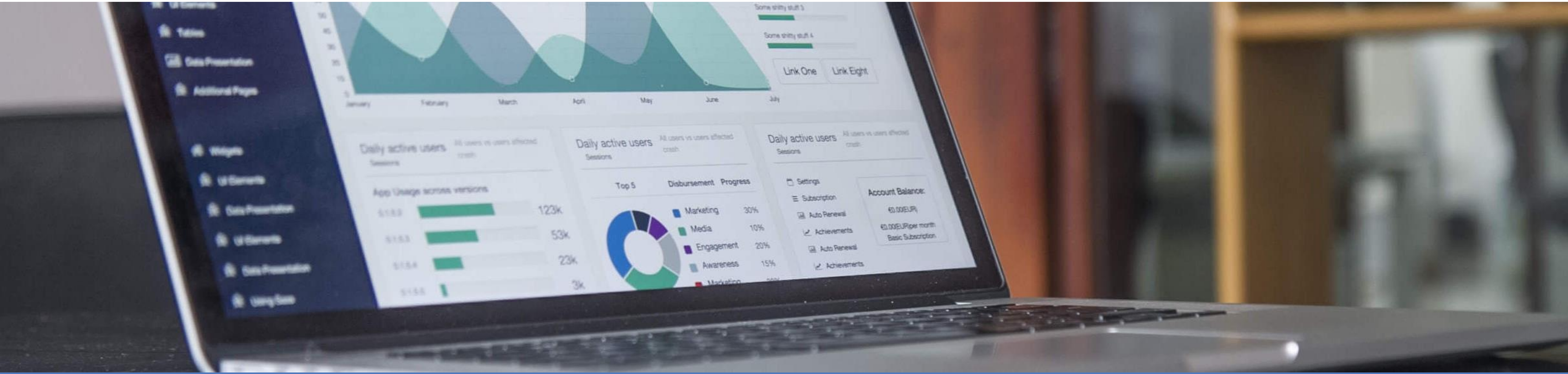
- Do **internal and external validity contradict** each other?
  - **Internal validity**: “You have to control all interfering variables”
  - **External validity**: “You establish an artificial, experimental setting”
- **Theories are being tested deductively**, not inductively
  - Theories are always based on the **assumption of falsification**
  - Does the observation of an experiment with high internal validity contradicts the theory?
    - If yes: irrelevant if the results are “representative”

      → The theory must be discarded or refined
    - If no: the experiment supports the theory

      → The theory must be further tested

What‘s next?

- **Evaluating Quantitative Data**: objectively report the
  - Descriptive **and** Inferential Statistics
    - **Descriptive statistics are easy**: text, plot, or table
    - **Inferential statistics are horror for students**: There are thousands of inferential statistical tests and the challenge is to find the correct one for your study
      - Running this one test is super easy!
- **Evaluating Qualitative Data**: subjectively analyze the feedback and observations
  - Thematic Analysis **or** Grounded Theory
    - Both need time but grounded theory needs more time than thematic analysis
    - If you have only qualitative feedback in your study it is highly recommended to perform grounded theory to extent the contribution
    - In mixed-method designs: thematic analysis is okay



Evaluating Quantitative Data

Human-Computer Interaction Exercise

Let‘s Test Our New Swipe Feature

Swipe

No Swipe

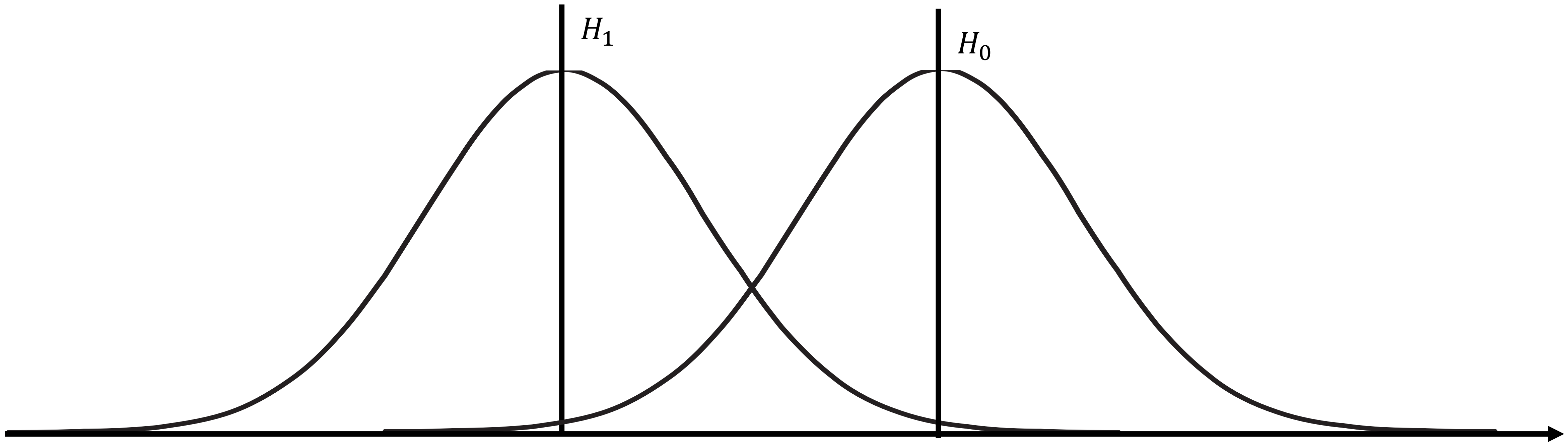

SAMSUNG

Example

- ~~Swipe users type very fast~~
- Typing with Swipe increases the typing performance (H1) and decreases workload (H2) compared to typing without Swipe.
  - Typing performance can operationalized by
    - words per minute (WPM)
    - characters per minute (CPM)
    - error rate
      - number of wrong / number of total words
      - number of backspace presses / number of characters
  - Workload can be operationalized by
    - NASA Taskload Index TLX, Quantitative Workload Inventory QWI, CarMen-Q Questionnaire,…
  - Do you assume an effect on…. all measures?

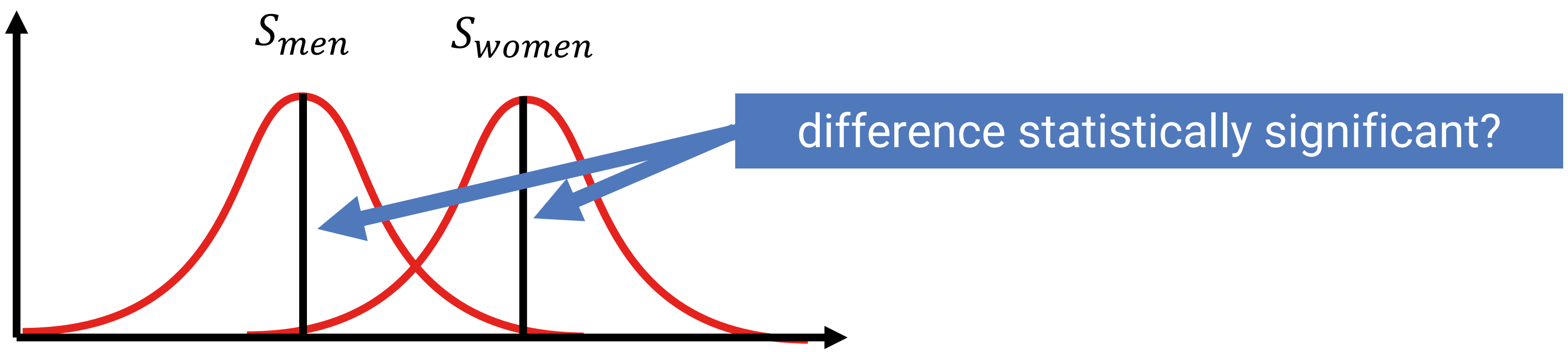
The Hypothesis

- **You** consider the hypothesis as **an explanation**
- **Statisticans** consider the hypothesis as **a signal** (or its probability distribution)
  - Thus, to assess the meaning of a signal (and the hypothesis) **we need a comparison**



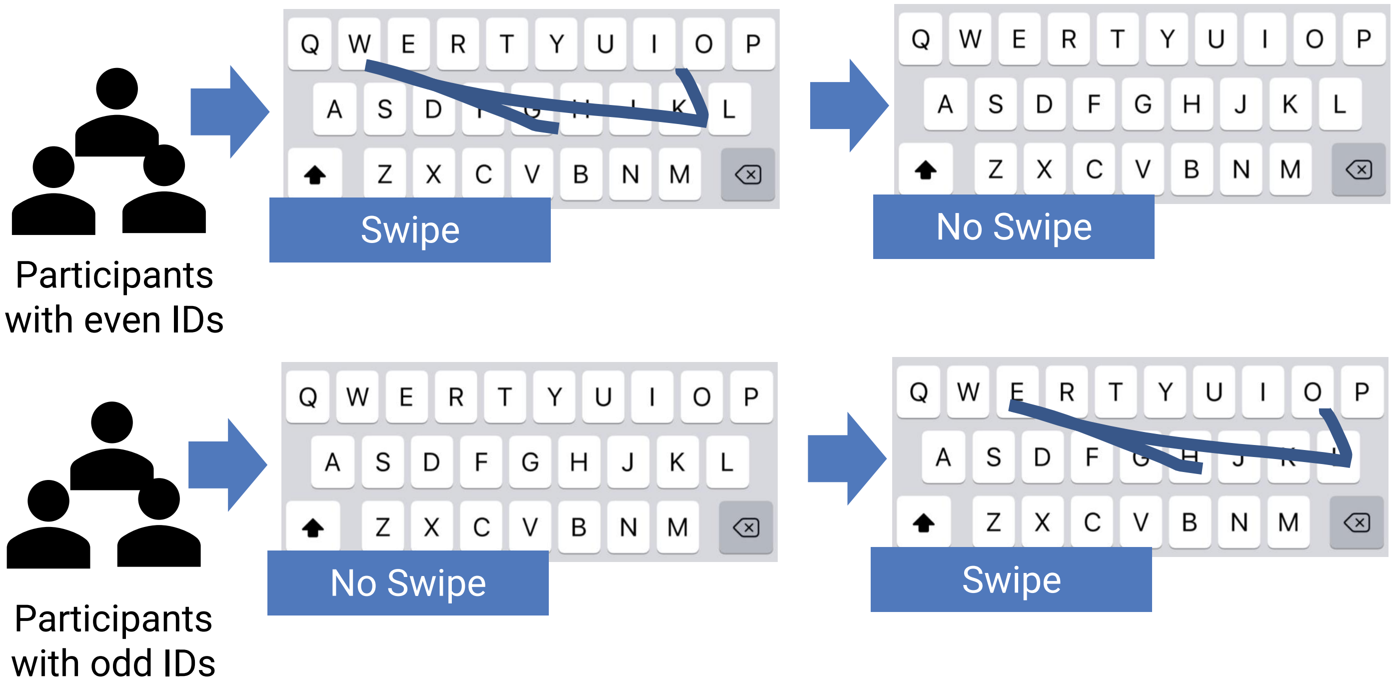
Comparing Hypotheses

- **Alternative Hypothesis** (“H1”, “H2”…)
  - e.g., *“There is a difference in typing speed between males and females“*
  - Directional Hypothesis („H1a”):
    - e.g., *“Males have a lower typing speed than females“*
- **Null hypothesis** (“H0”)
  - e.g., *“There is no difference in typing speed between males and females“*



Comparing Two Prototypes

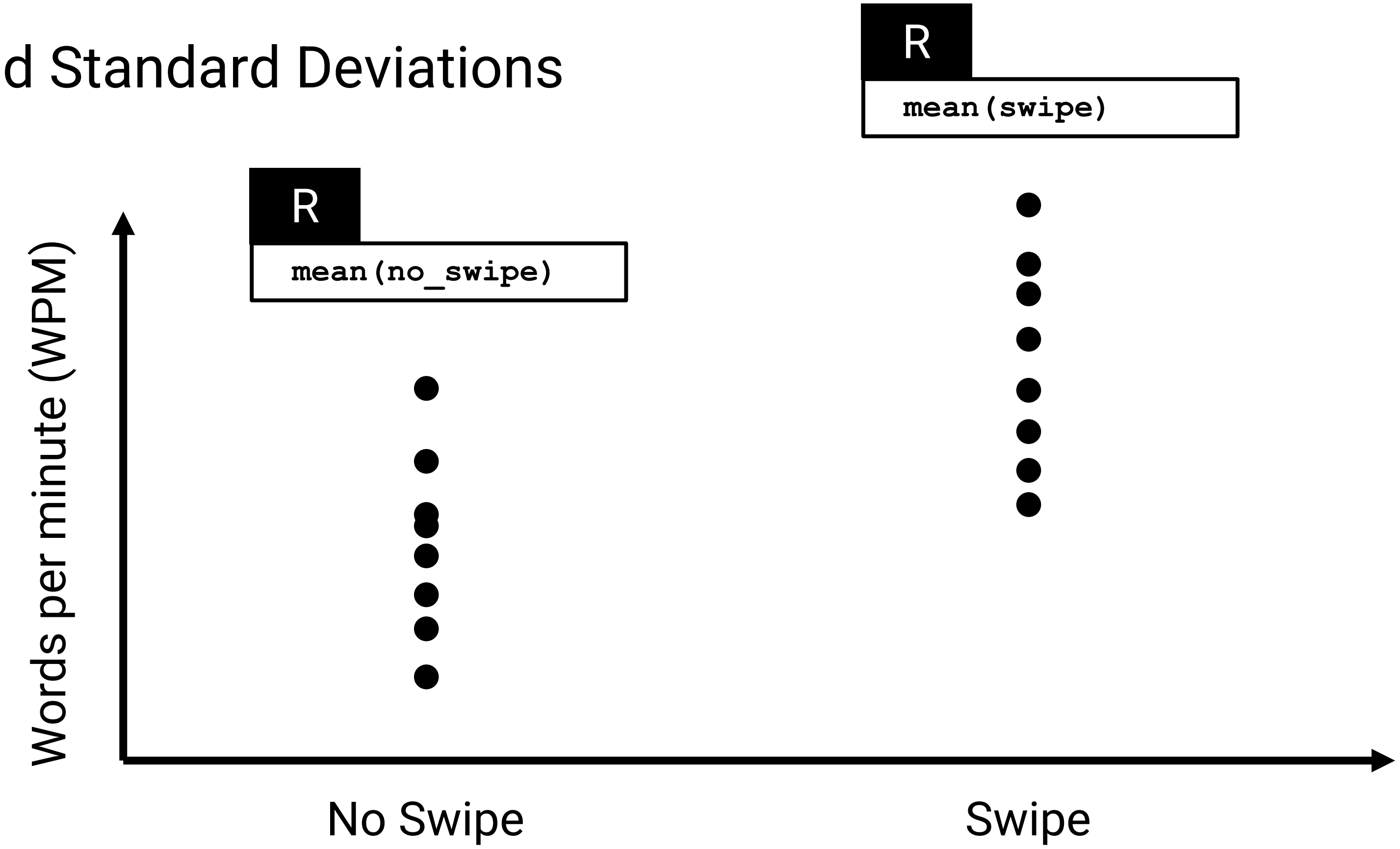
- Within-subject: Counter-balanced order
  - Swipe → No Swipe
  - No Swipe → Swipe




SAMSUNG

Comparing Two Prototypes

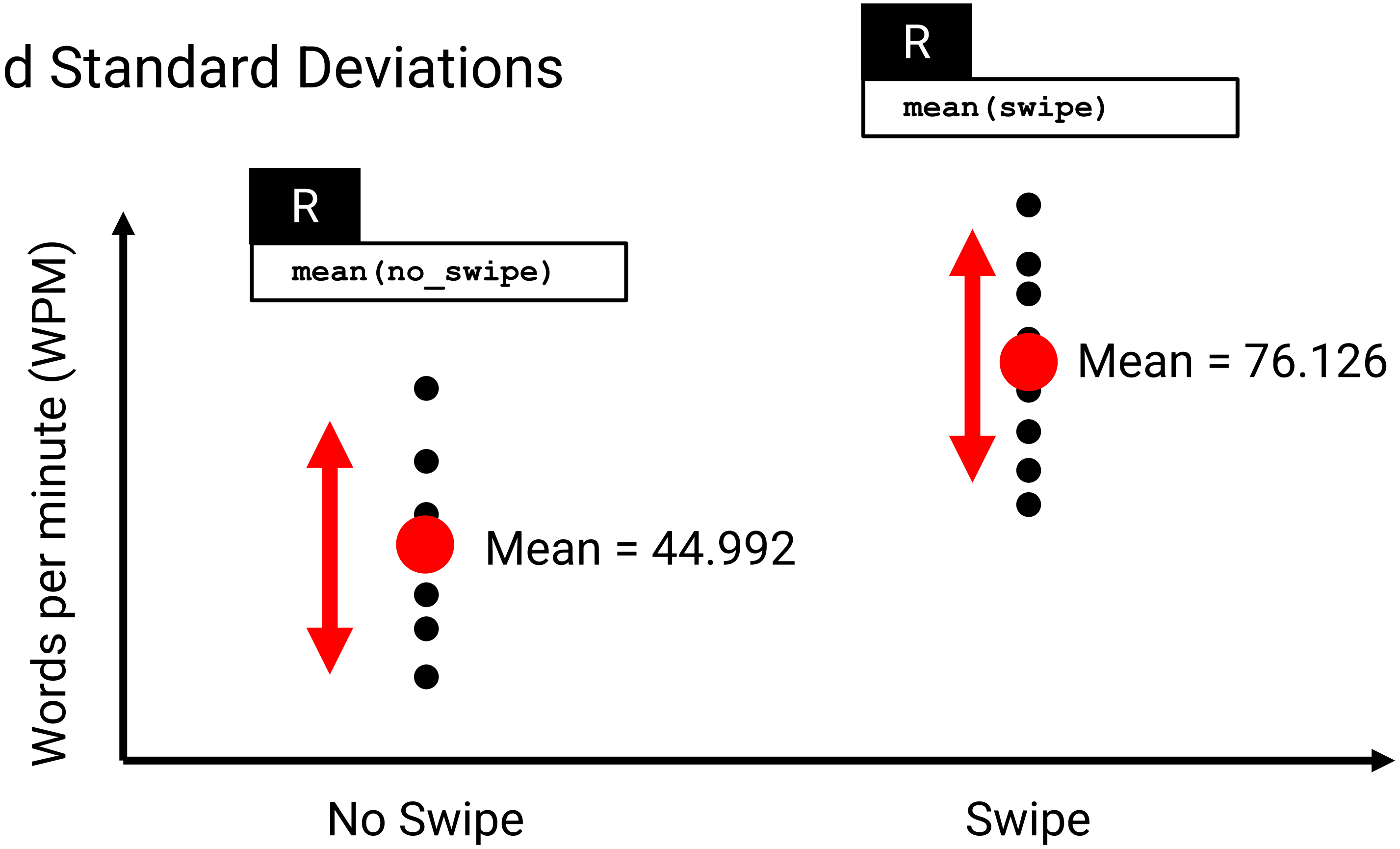
- Means and Standard Deviations



#	No Swipe	Swipe
1	44.559	75.381
2	42.951	76.255
3	44.398	93.795
4	25.026	82.015
5	54.82	55.238
6	44.034	70.151
7	50.782	75.997
8	55.549	88.35
9	56.425	63.869
10	43.983	68.029
11	30.747	77.1
12	46.634	87.327

Comparing Two Prototypes

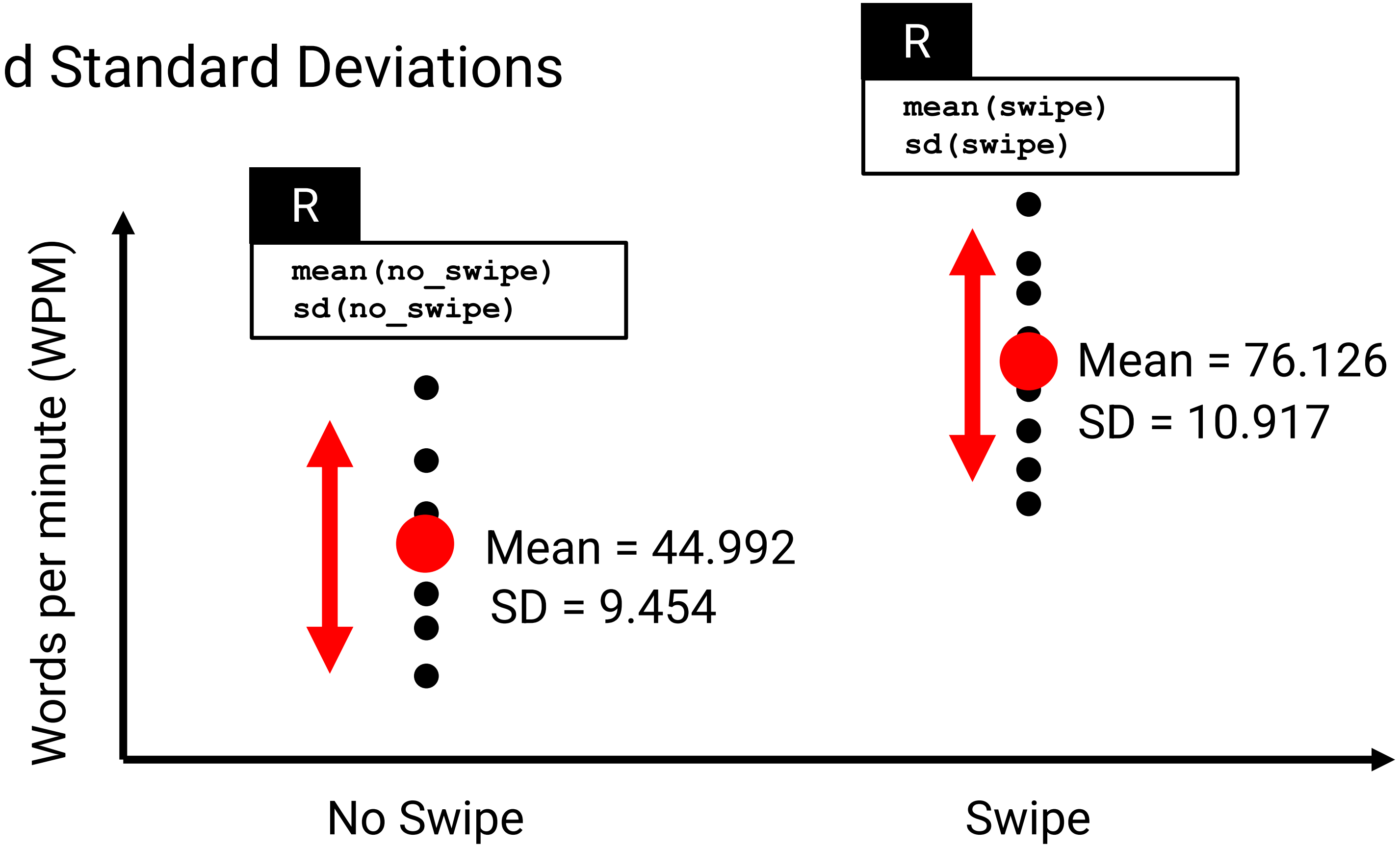
- Means and Standard Deviations



#	No Swipe	Swipe
1	44.559	75.381
2	42.951	76.255
3	44.398	93.795
4	25.026	82.015
5	54.82	55.238
6	44.034	70.151
7	50.782	75.997
8	55.549	88.35
9	56.425	63.869
10	43.983	68.029
11	30.747	77.1
12	46.634	87.327

Comparing Two Prototypes

- Means and Standard Deviations



#	No Swipe	Swipe
1	44.559	75.381
2	42.951	76.255
3	44.398	93.795
4	25.026	82.015
5	54.82	55.238
6	44.034	70.151
7	50.782	75.997
8	55.549	88.35
9	56.425	63.869
10	43.983	68.029
11	30.747	77.1
12	46.634	87.327

Descriptive Statistics: Text, Plot, or Table

- *„The average WPM without Swipe was M = 44.992 (SD = 9.454) while the average WPM using Swipe was M = 76.126 (SD = 10.917).“*

Text

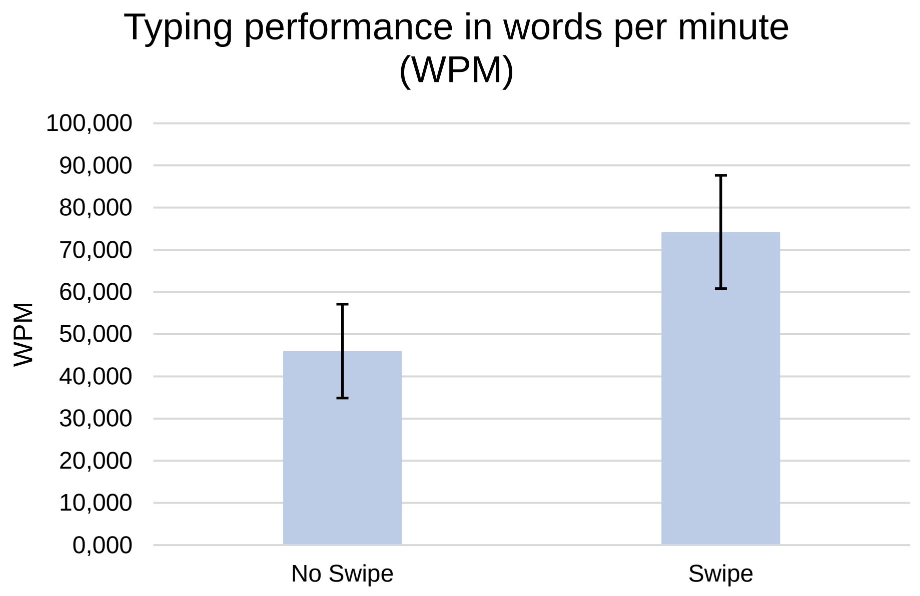


Figure 3: Average typing speeds of Swipe and typing without. Error bars show standard deviation.

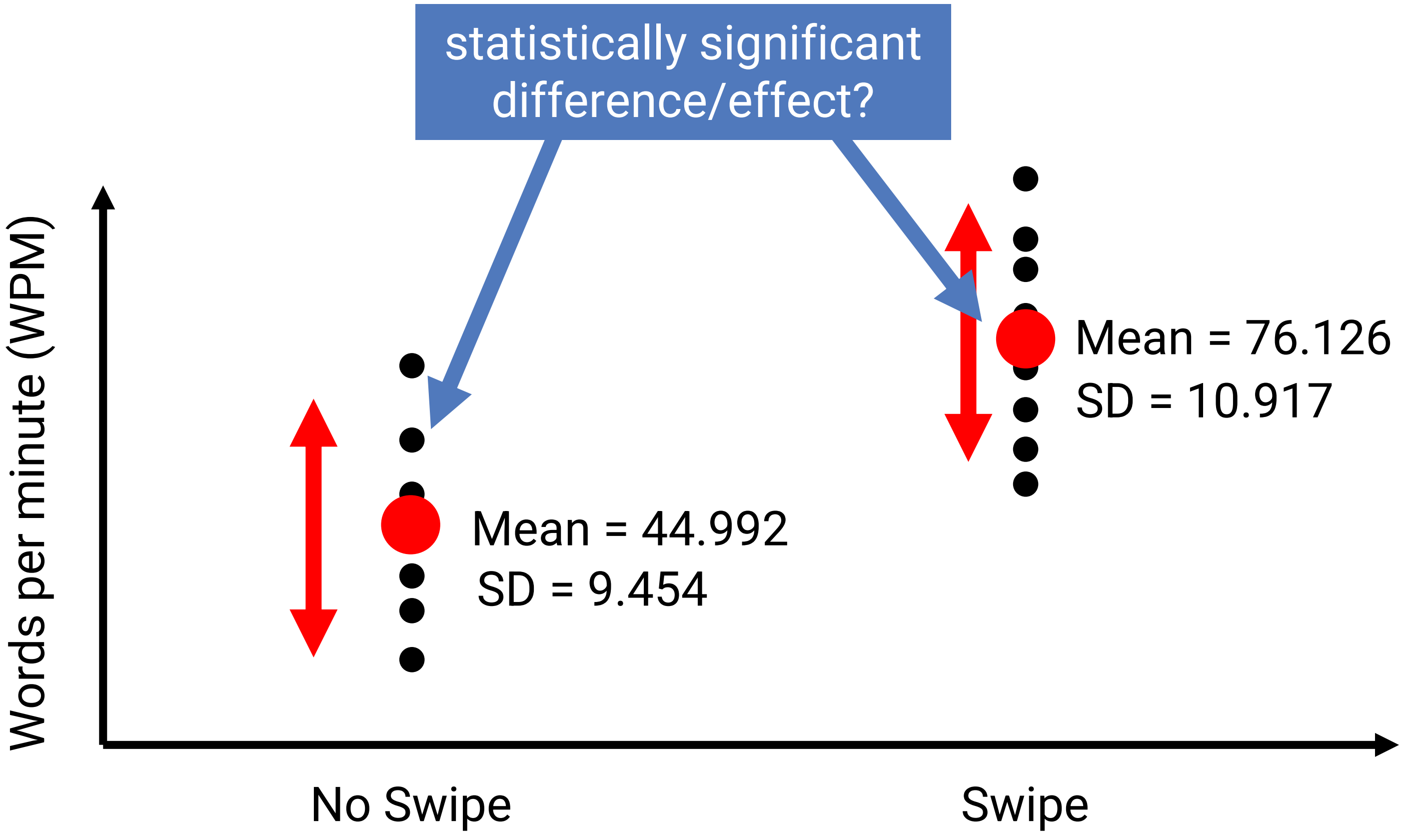
Plot

Keyboard	Mean WPM	SD WPM
No Swipe	44.992	9.454
Swipe	76.126	10.917

Table

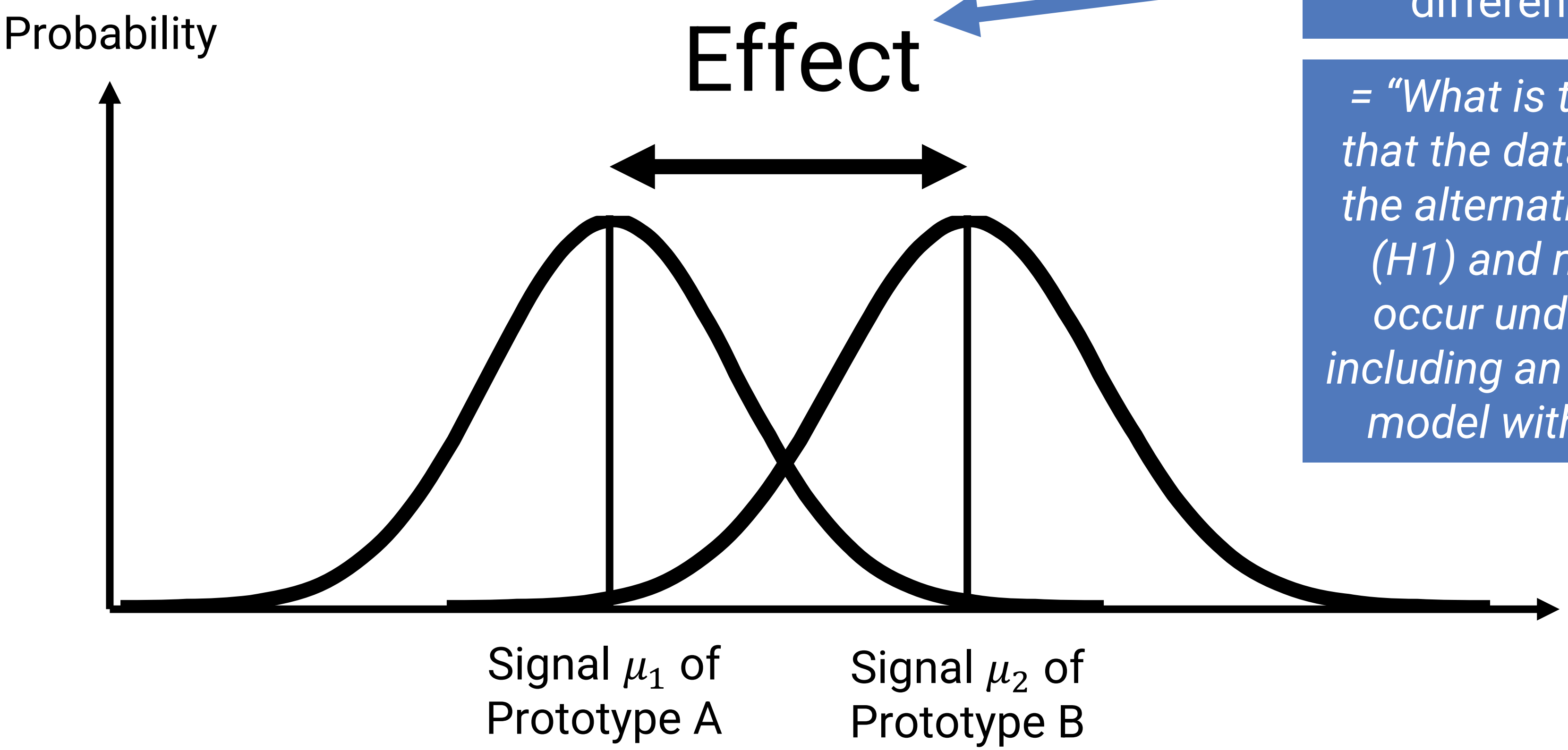
Combine visual with text or table

Comparing Two Prototypes



#	No Swipe	Swipe
1	44.559	75.381
2	42.951	76.255
3	44.398	93.795
4	25.026	82.015
5	54.82	55.238
6	44.034	70.151
7	50.782	75.997
8	55.549	88.35
9	56.425	63.869
10	43.983	68.029
11	30.747	77.1
12	46.634	87.327

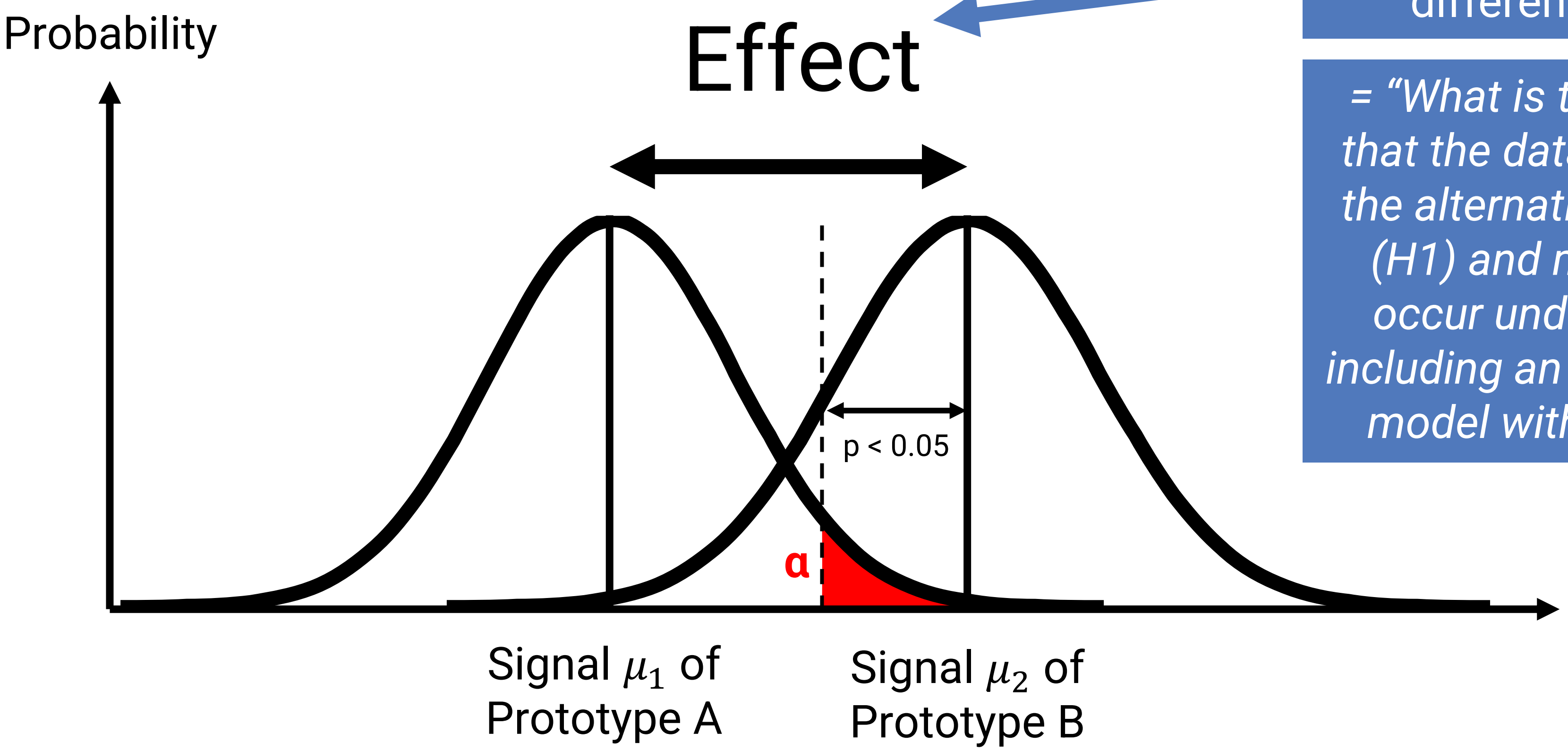
Statistical Significant Difference/Effect



statistically significant difference/effect?

= "What is the probability that the data is in favor of the alternative hypothesis (H1) and more likely to occur under the model including an effect than the model without it (H0)"?

Statistical Significant Difference/Effect



statistically significant difference/effect?

= “What is the probability that the data is in favor of the alternative hypothesis (H1) and more likely to occur under the model including an effect than the model without it (H0)”?

Statistical Significance

- A **statistically significant** effect (or difference) exists if the probability that the difference occurred is **below a certain significance level**
- Significance level (α)
  - **Lower significance level means higher evidence**
  - Arbitrary, but typical significance level: **$\alpha = 0.05$**
- Significant results (**$p < \alpha$**)
  - **Null hypothesis can be rejected**
    - “There is a statistically significant effect (or difference)”
- Non-Significant results (**$p >= \alpha$**)
  - **Null hypothesis cannot be rejected**
    - “We cannot conclude anything!”

Type I & Type II Errors

- Let‘s say we have p = .028

Effect found	Effect exists: false	Effect exists: true
true	**Type I error** (False Positive) non-existing effect found 2.8%	
false		

Type I & Type II Errors

- Let‘s say we have p = .028

	Effect exists: false	**Effect exists: true**
Effect found: true	**Type I error** (False Positive) non-existing effect found 2.8%	**Correct** (True Positive) existing effect was found 97.2%
Effect found: false		

Type I & Type II Errors

- Let‘s say we have p = .028

Type I error (False Positive) non-existing effect found 2.8%	**Correct** (True Positive) existing effect was found 97.2%	**true**	**Effect found**
Correct (True Negative) no effect exists, no effect found		**false**	
false	**true**		
Effect exists			

Type I & Type II Errors

- Let‘s say we have p = .028

	Effect exists: false	**Effect exists: true**
Effect found: true	**Type I error** (False Positive) non-existing effect found 2.8%	**Correct** (True Positive) existing effect was found 97.2%
Effect found: false	**Correct** (True Negative) no effect exists, no effect found	**Type II error** (False Negative) effect exists, but is not found

Type III and Type IV Errors

- **Type III: “Wrong hypothesis, right answer”**
  - Researcher is either focusing on theory or on evaluation but not on the reasoning chain
  - Incorrect operationalization of variables
  - Poor theory (e.g., ad hoc explanations of findings)
  - Mis-identifying causal architecture
    - e.g.: focusing on inter-individual factors (gender- or age-related differences) rather than structural factors
- **Type IV: “Right hypothesis, wrong answer”**
  - Collinearity among predictors
  - Aggregation bias
  - Wrong test

Statistical Power

- Statistical power of a binary hypothesis test is **the probability that a statistical test correctly rejects the null hypothesis** (in %) when a specific alternative hypothesis is true.
- Aspects that increase the statistical power
  - **increase the statistical significance criterion** (α = 0.05)

    → You need a justification why you increased α. Almost impossible because it is consensus.
  - **more conditions**

    → Statisticians have a trick to get around this
  - **more measures**

    → You need a justification why a measure is part of the research question
  - **higher sample size**

    → Invite more participants
  - **higher effect size**

    → Make something impactful (we like that)
  - **reducing noise in your data**

    → Decrease the variance to get statistically significant results

Statistical Power

- Statistical power of a binary hypothesis test is **the probability that a statistical test correctly rejects the null hypothesis** (in %) when a specific alternative hypothesis is true.
- Aspects that increase the statistical power
  - **increase the statistical significance criterion** (α = 0.05)

    → You need a justification why you increased α. Almost impossible because it is consensus.
  - **more conditions**

    → Statisticians have a trick to get around this
  - **more measures**

    → You need a justification why a measure is part of the research question
  - **higher sample size**

    → Invite more participants
  - **higher effect size**

    → Make something impactful (we like that)
  - **reducing noise in your data**

    → Decrease the variance to get statistically significant results

be very careful!

optimize for that!

Estimate the Effect Size

- An effect size **measures the strength of the relationship between independent and dependent variable**
- Cohen's d [1] is a measure of the standardized difference between two samples

$$d = \frac{m_1 - m_2}{sd} \quad \text{with} \quad sd = \sqrt{\frac{sd_1^2 + sd_2^2}{2}}$$

M1: Mean of Group 1, M2: Mean of Group 2, SD: Pooled standard deviation

- **We can interpretate them**:
  - Negligible effect size: $|d| \approx 0.0$
  - Small effect size: $|d| \approx 0.2$
  - Medium effect size: $|d| \approx 0.5$
  - Large effect size: $|d| \approx 0.8$
- **There are many more measures of the effect size!**
  - They also depend on the statistical test (etc. t-test effect size != ANOVA effect size)

[1] J. Cohen: Statistical Power Analysis for the Behavioral Sciences. 2. Auflage. Lawrence Erlbaum Associates, Hillsdale 1988, ISBN 0-8058-0283-5.

Estimate the Sample Size

- Typically, you invite **the correct sample number** participants/samples
  - **You need the estimated effect size**:
    - $|d| < 0.20$ (negligible), $|d| > 0.20$ (small), $|d| > 0.50$ (medium), $|d| > 0.80$ (large)
- **Quick'n'dirty: Lehr's rule of thumb** [1] for sample sizes: $N = \frac{16}{d^2}$ with $d = \frac{m_1 - m_2}{sd}$
  - e.g., to detect a 10-point difference between two groups with a SD of 20:
    - $\mathrm{N} = \frac{16}{(100-90/20)^2} = 64$
    - You need 64 people
- **Correct: Power Analyses**
  - Compute the statistical power analyses for your test with G*Power [2]

[1] Robert Lehr (1992), "SixteenS-squared overD-squared: A relation for crude sample size estimates", Statistics in Medicine (in German), vol. 11, no. 8, pp. 1099–1102, doi:10.1002/sim.4780110811, ISSN 0277-6715
[2] https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower

Statistical Tests

- We need a **statistical tests** to find **statistically signficant effects**
- **Before you start with any experimental research:**
  1. **Ensure that there is a** statistical **test** for your design
  2. **Determine the** correct statistical **test** for your design
  3. **Experiment with placeholder data** (e.g., from earlier or hypothetical experiments) if the test can be performed correctly:
  4. **The kind of data** your dependent and independent variable

Kind of Data

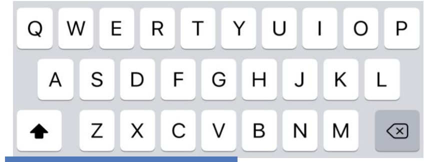
Kind	Continuous			Discrete
Data Type	Ratio Data	Interval Data	Ordinal Data	Nominal Data
Quantity	Devided numbers	Numbers with fixed intervals	Numbers you can sort	Things in categories
Examples	scores, bandwidth, speed, rates, throughput, **words per minute**…	distance in meters, time, temperature in C°, weight, …	marks in school (1,2,3,4,5,6), counts, ranks, orders, …	gender, countries, animals, groupsprototypes, scenarios…
Data Properties	Parametric		Nonparametric Data	
Variance	▪ Homogeneous (equal)		▪ Homo- and Hetereogeneous	
Relationships	▪ Independent		▪ Any	
Central Measure	▪ Means ▪ Standard Deviation ▪ Standard Error ▪ 95% Confidence Interval		▪ Median ▪ 25/50/75/95% Quantiles	
Distribution	▪ Normal		▪ Any	
Plot	▪ Line Chart, Bar chart, etc.		▪ Boxplot, Violinplot	
Benefits	▪ Generalizeable ▪ Can draw more conclusions		▪ Simpel ▪ Robust against outliers	

Better, more generalizable

The tricky part

Statistical Test Checklist

- **Which test is the correct for my design?** Depends on
  1. **The number** of dependent variables in my hypothesis: **?**
  2. **The kind** of dependent variables: **?**
  3. **The number** of independent variables: **?**
  4. **The kind** of independent variables: **?**
  5. **The levels** of the independent variable: : **?**
  6. Are the independent variable **between/within-subjects/both**: **?**
  7. **Is the DV really parametric data**: **?**



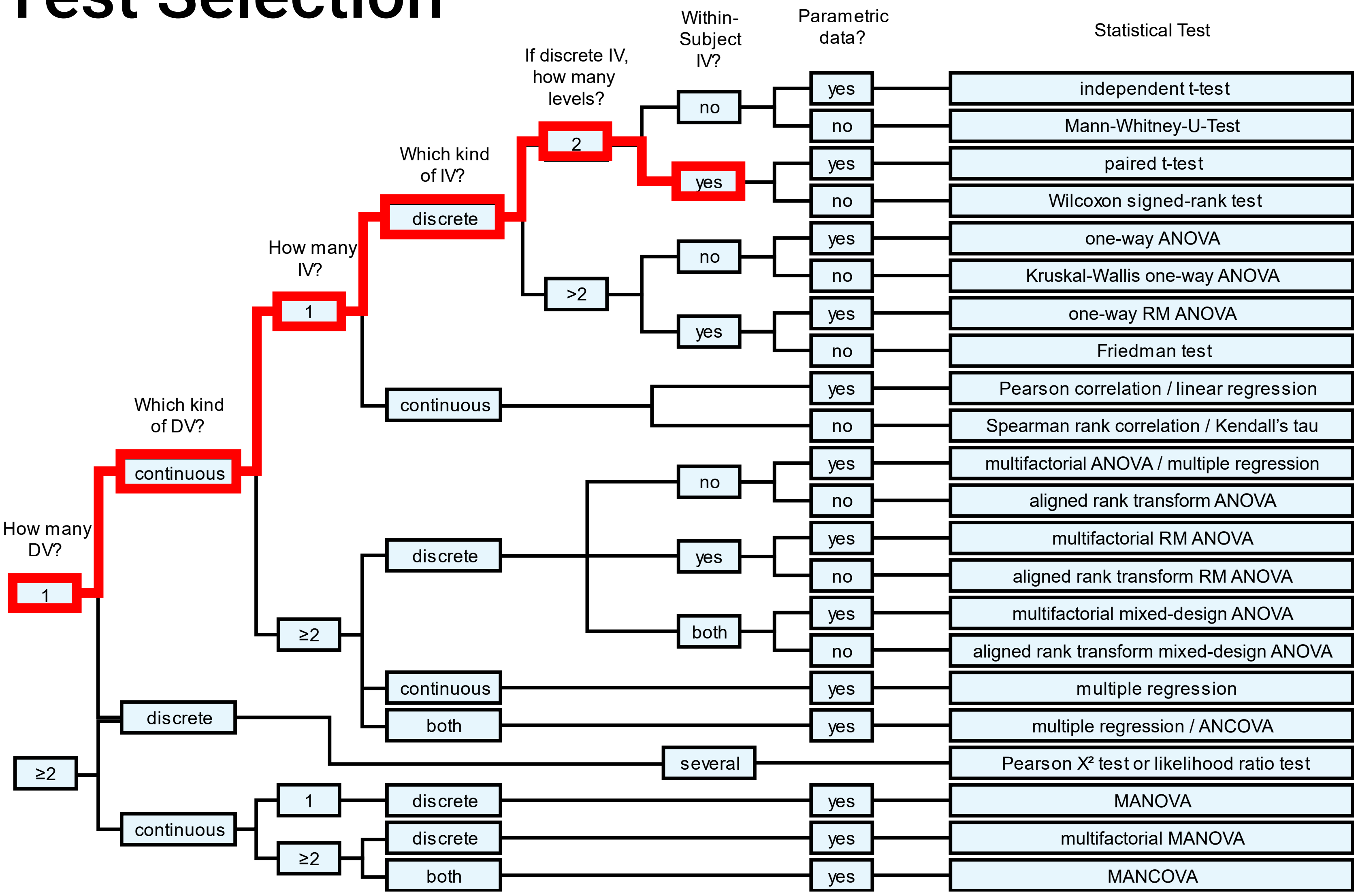
No Swipe

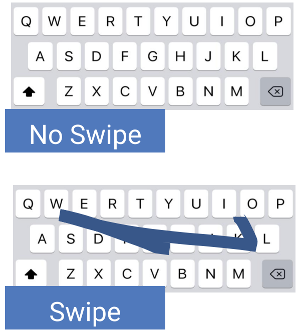
Swipe

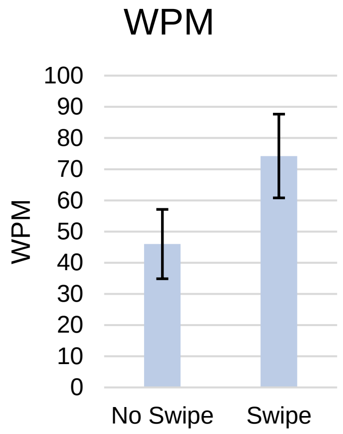
Statistical Test Checklist

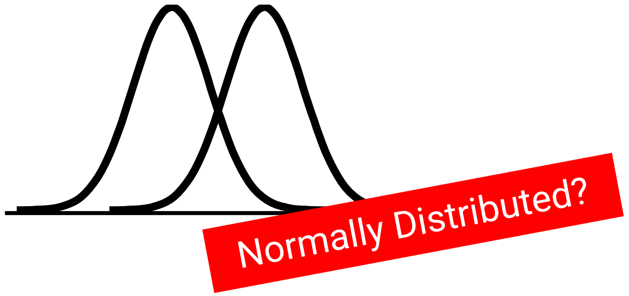
- **Which test is the correct for my design?** Depends on
  1. **The number** of dependent variables in my hypothesis: **WPM → 1**
  2. **The kind** of dependent variables: **words per minute → continous**
  3. **The number** of independent variables: **Keyboard → 1**
  4. **The kind** of independent variables: **Swipe, No Swipe → discrete**
  5. **The levels** of the independent variable: : **Swipe, No Swipe → 2**
  6. Are the independent variable **between/within-subjects/both**: **within**
  7. **Is the DV <u>really</u> parametric data**: **???**
  - **We have no idea unless we ran the study. There are different tests for parametric and non-parametric data!**

Test Selection

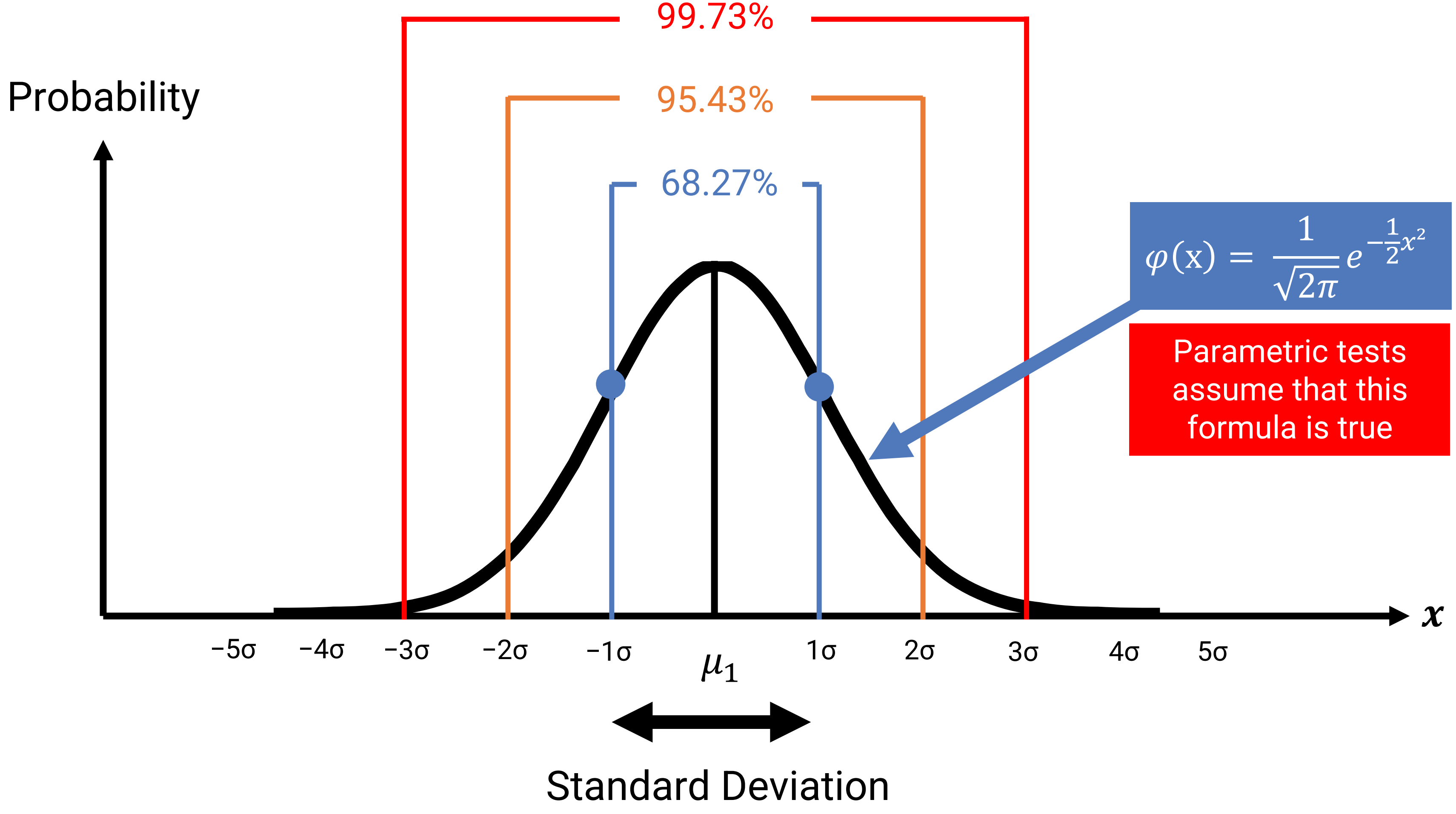




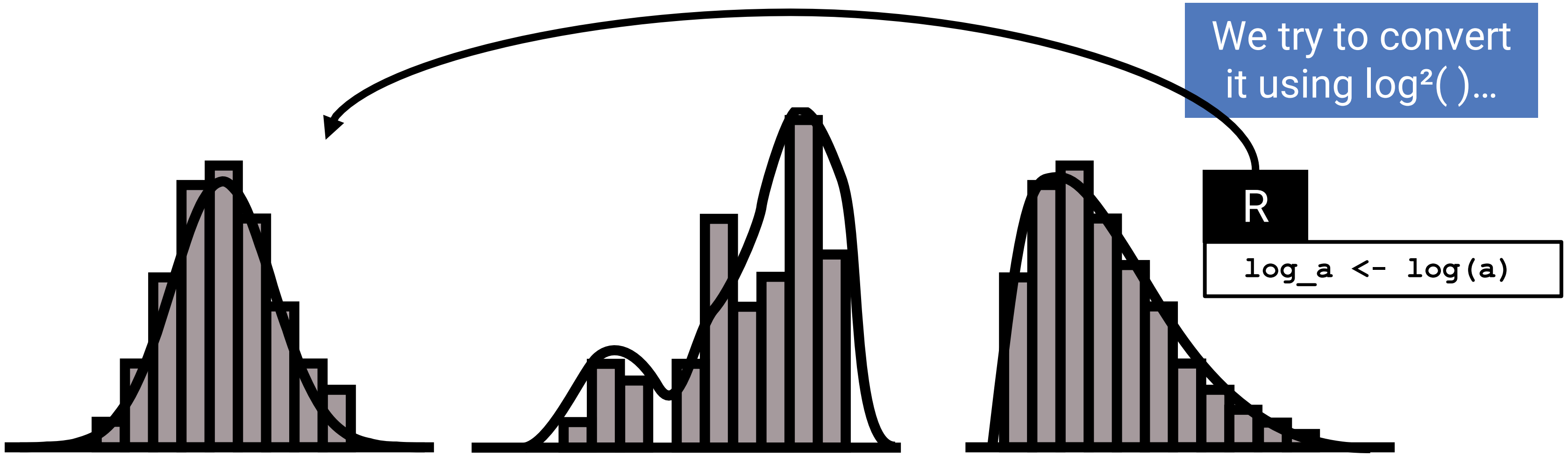




Parametric Data = Normal Distributed Data



Normality



normally distributed

not normally distributed

skewed (consider log-transform)

Parametric

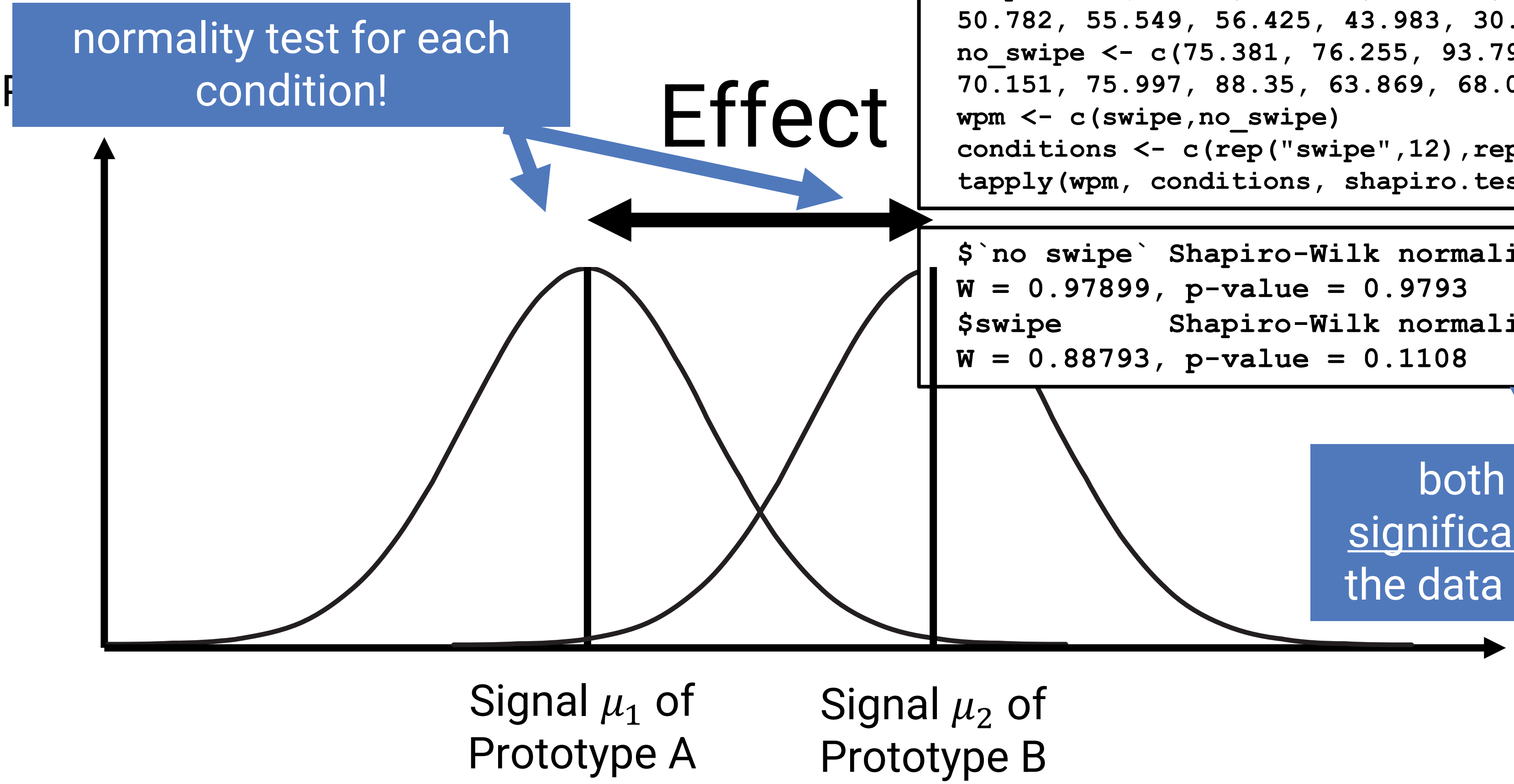
Non-Parametric

Non-Parametric

Normality Test

- **Many statistical tests** such as correlation, regression, t-test, and analysis of variance (ANOVA) **require** the data to follow a **normal distribution** or Gaussian distribution
  - These tests are called **parametric tests**, because their validity depends on the distribution of **parametric data**
- Before using a parametric test, **we must perform a test on normal distribution** to make sure that the test assumptions are met
  - **If not, non-paramatric tests are needed**
    - There are non-parametric test, but not for every study design
    - Parametric tests allow less conclusions (e.g., on the significance of absolute values)

Shapiro-Wilk Normality Test



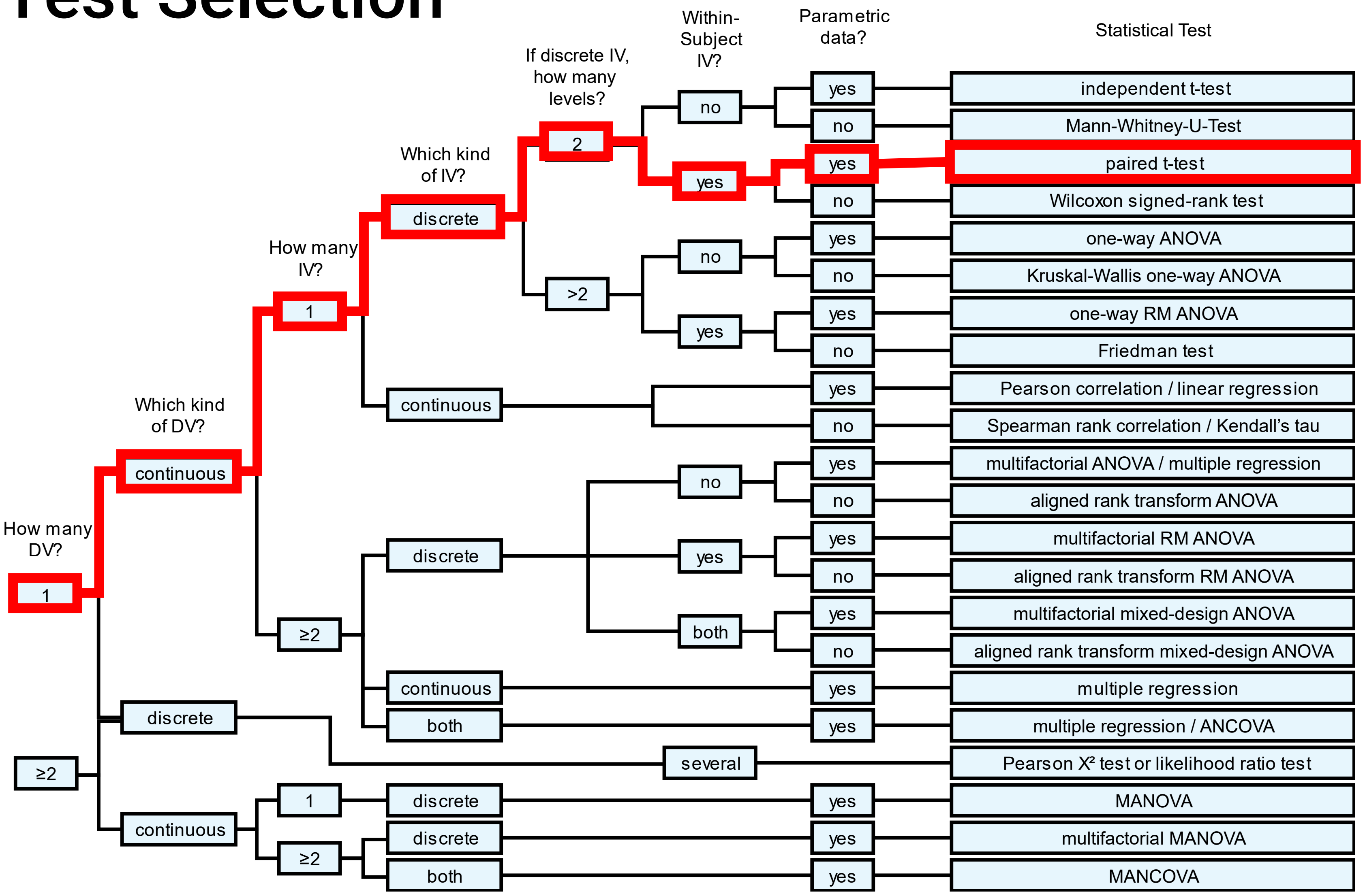
R

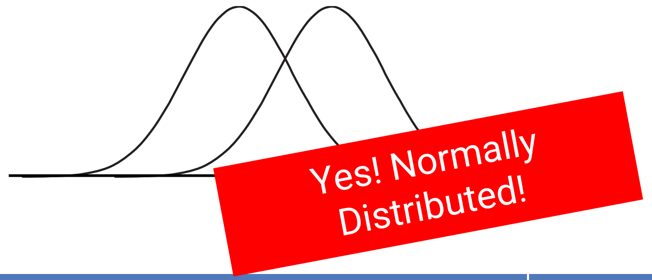
```
swipe <- c(44.559, 42.951, 44.398, 25.026, 54.82, 44.034,
50.782, 55.549, 56.425, 43.983, 30.747, 46.634)
no_swipe <- c(75.381, 76.255, 93.795, 82.015, 55.238,
70.151, 75.997, 88.35, 63.869, 68.029, 77.1, 87.327)
wpm <- c(swipe,no_swipe)
conditions <- c(rep("swipe",12),rep("no swipe",12))
tapply(wpm, conditions, shapiro.test)
```

```
$`no swipe` Shapiro-Wilk normality test
W = 0.97899, p-value = 0.9793
$swipe      Shapiro-Wilk normality test
W = 0.88793, p-value = 0.1108
```

both tests are not significant = normality of the data can be assumed

Test Selection





Paired t-Test

- also called „dependent t-test“

R

```
t.test(swipe, no_swipe, paired = TRUE)

        Paired t-test

data:  swipe and no_swipe
t = -6.5882, df = 11, p-value = 3.926e-05
alternative hypothesis: true difference in means is not equal to 0
```

Yes! Statistically significant!

- is: $p < \alpha$?
  - yes: $p = 3.926 \cdot 10^{-05} < 0.05 < 0.001$
  - The null hypothesis („that no statistical significant effect occured“) can be rejected (we like that)
- What is „t“? What is„df“?
  - **t = the test-statistic**: the difference presented in units of the standard errors (the higher, the better)
  - **df = N - 1**: the degree of freedom

#	No Swipe	Swipe
1	44.559	75.381
2	42.951	76.255
3	44.398	93.795
4	25.026	82.015
5	54.82	55.238
6	44.034	70.151
7	50.782	75.997
8	55.549	88.35
9	56.425	63.869
10	43.983	68.029
11	30.747	77.1
12	46.634	87.327

Descriptive and Inferential Statistics: Text, Plot, and Table

- *„The average WPM without Swipe was M = 44.992 (SD = 9.454) while the average WPM using Swipe was M = 76.126 (SD = 10.917). A paired t-test revealed that the difference between them was statistically significant, t(7) = -6.589, p < .001."*

Text

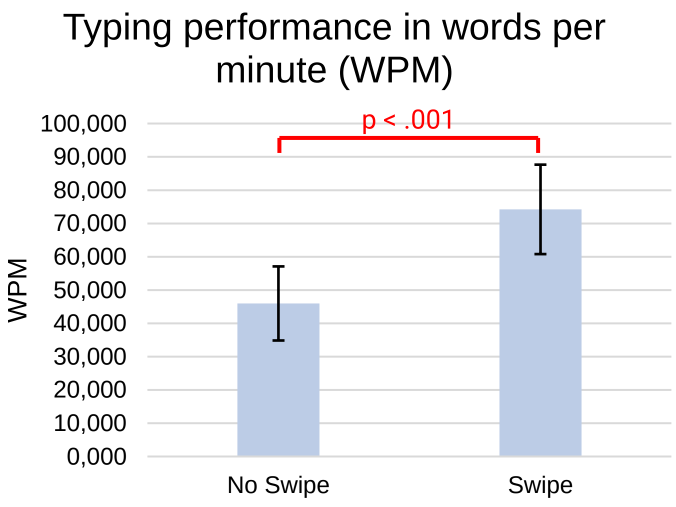


Figure 3: Average typing speeds of Swipe and typing without. Error bars show standard deviation.

Plot

Pairwise Comparison	t(13)	p
No Swipe - Swipe	-6.589	< .001

Table makes sense when you have more than one comparison…

Descriptive and Inferential Statistics: Text, Plot, and Table

- „*The average WPM without Swipe was M = 44.992 (SD = 9.454) while the average WPM using Swipe was M = 76.126 (SD = 10.917). A paired t-test revealed that the difference between them was statistically significant, t(11) = -6.589, p < .001.*"

Text

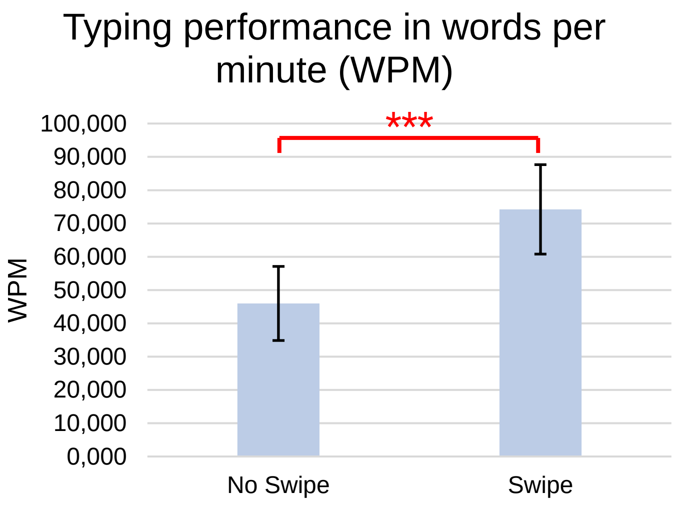


Figure 3: Average typing speeds of Swipe and typing without. Error bars show standard deviation.

Plot

Asterisk Interpretation

ns	→ not sig.
*	→ p < .05
**	→ p < .01
***	→ p < .001

Pairwise Comparison	**t(11)**	**p**
No Swipe - Swipe	-6.589	< .001

Table makes sense when you have more than one comparison…

But...

https://pxhere.com/de/photo/1582787

Statistically significant – but also important?

- **A significant difference does not tell much about how "large" the effect is**
- Thus, **we must also determine the effect size**
  - Tells us which extent of the difference is caused by our manipulation
    - In our case: ***Cohen's d***

$$d = \frac{M_1 - M_2}{SD} \qquad SD = \sqrt{\frac{SD_1^2 + {SD_2}^2}{2}}$$

    - Widely accepted interpretations are:

$|d| < 0.20$ (negligable), $|d| > 0.20$ (small), $|d| > 0.50$ (medium), $|d| > 0.80$ (large)

R

```
(mean(swipe) - mean(no_swipe)) / sqrt((sd(swipe)^2 + sd(no_swipe)^2) / 2)
```

    - Effect size in our example: $d = -3.049(large)$

Descriptive and Inferential Statistics: Text, Plot, and Table

- *„The average WPM without Swipe was M = 44.992 (SD = 9.454) while the average WPM using Swipe was M = 76.126 (SD = 10.917). A paired t-test revealed that the difference between them was statistically significant, t(11) = -6.589, p < .001, and had a large effect size (d=−3.049).“*

Text

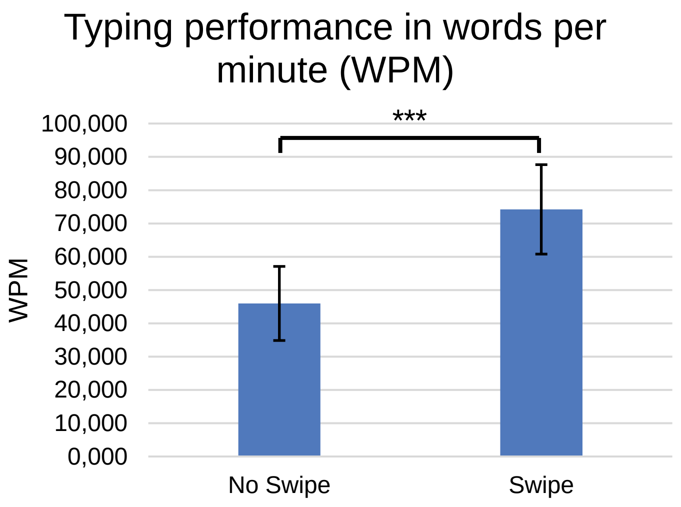


Figure 3: Average typing speeds of Swipe and typing without. Error bars show standard deviation.

Plot

Pairwise Comparison	t(13)	p	d
No Swipe - Swipe	-6.589	< .001	−3.049

Table makes sense when you have more than one comparison…

https://pxhere.com/de/photo/781480

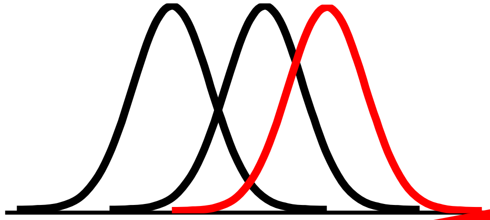
The Number of Tests

Adding a third condition…

R

```
swipe <- c(44.559, 42.951, 44.398, 25.026, 54.82, 44.034, 50.782, 55.549, 56.425,
43.983, 30.747, 46.634)
no_swipe <- c(75.381, 76.255, 93.795, 82.015, 55.238, 70.151, 75.997, 88.35, 63.869,
68.029, 77.1, 87.327)
vibro_swipe <- c(88.695, 95.79, 75.601, 95.237, 94.21, 58.038, 73.444, 53.036,
83.062, 71.641, 72.46, 75.809)
wpm <- c(swipe, no_swipe, vibro_swipe)
conditions <- c(rep("No Swipe",12), rep("Swipe",12), rep("Vibro Swipe",12))
tapply(wpm, conditions, shapiro.test)
```

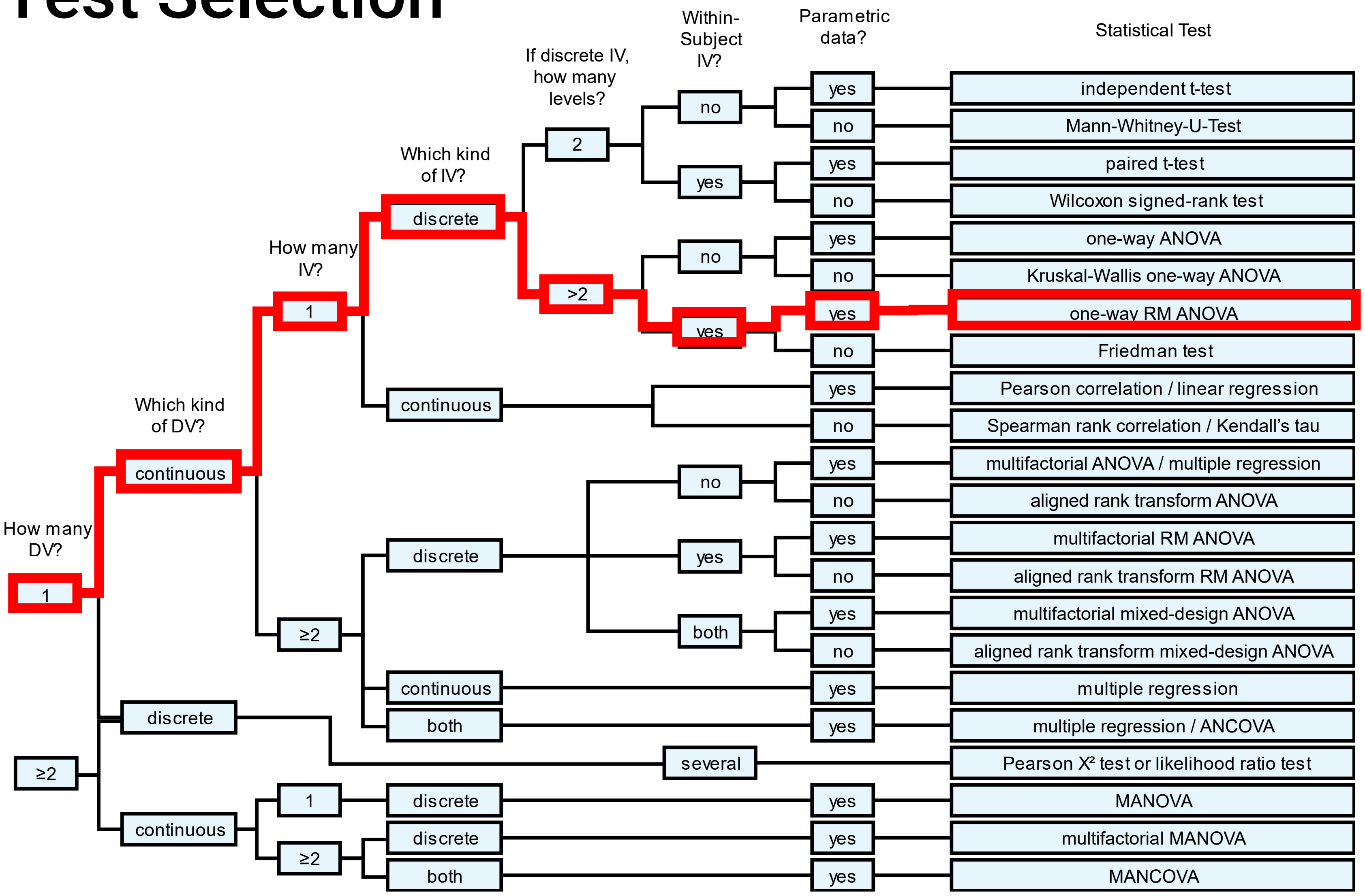
```
$`no swipe`    Shapiro-Wilk normality test
W = 0.88793, p-value = 0.1108
$Swipe         Shapiro-Wilk normality test
W = 0.97899, p-value = 0.9793
$`Vibro Swipe` Shapiro-Wilk normality test
W = 0.92701, p-value = 0.3495
```

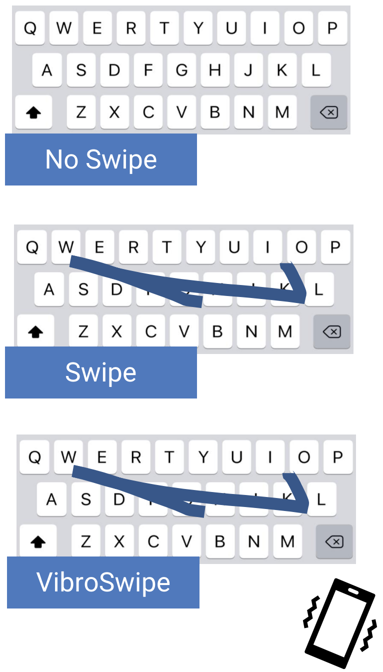


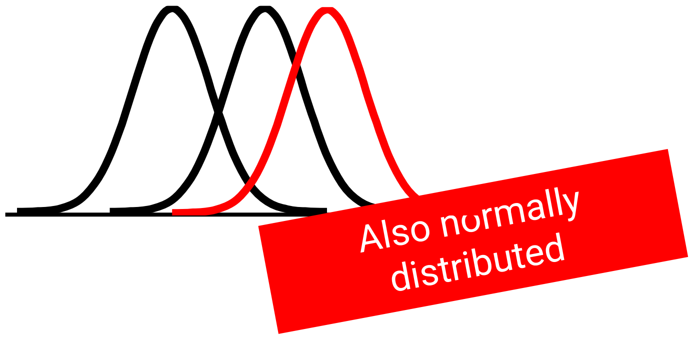
Normally Distributed

#	No Swipe	Swipe	VibroSwipe
1	44.559	75.381	88.695
2	42.951	76.255	95.79
3	44.398	93.795	75.601
4	25.026	82.015	95.237
5	54.82	55.238	94.21
6	44.034	70.151	58.038
7	50.782	75.997	73.444
8	55.549	88.35	53.036
9	56.425	63.869	83.062
10	43.983	68.029	71.641
11	30.747	77.1	72.46
12	46.634	87.327	75.809

Test Selection







One-Way RM-ANOVA

- **One-Way** = **One Factor** (your keyboard)
- **RM** = **Repeated Measures** (every participant was exposed to every condition)
- **ANOVA** = **Analysis of Variance** (the most used statistical test in the world)
- RM-ANOVAs musk first ask for: „Are the variances of the differences between the within-subject conditions are equal?"
  - If yes: „sphericity" can be assumed

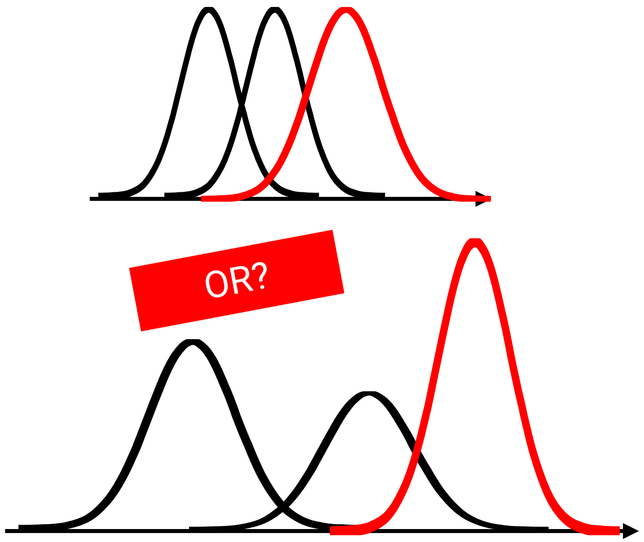
  → you can proceed

  - If not: <u>no</u> „sphericity" can be assumed

  → you need a kind of correction, to prevent an inflation of the F-ratio (the quality of how well the model fits the data)

WTF? Sphericity?

#	No Swipe	Swipe	VibroSwipe	A-B	A-C	B-C
1	44.559	75.381	88.695	44.559	75.381	88.695
2	42.951	76.255	95.79	42.951	76.255	95.79
3	44.398	93.795	75.601	44.398	93.795	75.601
4	25.026	82.015	95.237	25.026	82.015	95.237
5	54.82	55.238	94.21	54.82	55.238	94.21
6	44.034	70.151	58.038	44.034	70.151	58.038
7	50.782	75.997	73.444	50.782	75.997	73.444
8	55.549	88.35	53.036	55.549	88.35	53.036
9	56.425	63.869	83.062	56.425	63.869	83.062
10	43.983	68.029	71.641	43.983	68.029	71.641
11	30.747	77.1	72.46	30.747	77.1	72.46
12	46.634	87.327	75.809	46.634	87.327	75.809
			Variances	267,975	348,607	406,672



Comparable?

Creating a Table („data frame“) with Subject IDs…

R

```
swipe <- c(44.559, 42.951, 44.398, 25.026, 54.82, 44.034, 50.782, 55.549, 56.425, 43.983, 30.747, 46.634)
no_swipe <- c(75.381, 76.255, 93.795, 82.015, 55.238, 70.151, 75.997, 88.35, 63.869, 68.029, 77.1, 87.327)
vibro_swipe <- c(88.695, 95.79, 75.601, 95.237, 94.21, 58.038, 73.444, 53.036, 83.062, 71.641, 72.46, 75.809)
wpm <- c(swipe, no_swipe, vibro_swipe)
keyboard <- c(rep("No Swipe",12), rep("Swipe",12), rep("Vibro Swipe",12))

# normality check
tapply(wpm, keyboard, shapiro.test)

# to perform an RM-ANOVA, we need subject IDs (typically we have already a CSV with subject IDs
subjectID <- rep(seq(12, length=12),times=3)

 # create the table / data frame with subject IDs and conditions as fixed factors
df <- data.frame(subjectID = as.factor(subjectID), conditions= as.factor(conditions), wpm)

# if not installed add: install.packages("rstatix")
library(rstatix) # load the library

# run the ANOVA
anova <- anova_test(data = df, dv = wpm, wid = subjectID, within = keyboard, effect.size = "pes")

# automatically does the sphericity correction if necessary (we like that)
get_anova_table(anova, correction = "auto")
```

Please note: the data frame (df) has now this strange „long“ format

subjectID	keyboard	wpm
<fct>	<fct>	<dbl>
1	swipe	75.381
1	noswipe	86.255
2	noswipe	93.795
2	swipe	61.435

Reporting the RM-ANOVA Output

- RM-ANOVA Output
  - Sphericity can be assumed, because otherwise the output would look a bit different…

	Effect	DFn	DFd	F	p	p<.05	pes
1	keyboard	2	22	24.225	2.73e-06	*	0.688

  - *„A one-way RM-ANOVA revealed a significant effect of the three* K*EYBOARDS on the WPM measure, F(,) = , p , η_p^2 = .“*

Reporting the RM-ANOVA Output

- RM-ANOVA Output
  - Sphericity can be assumed, because otherwise the output would look a bit different…

	Effect	DFn	DFd	F	p	p<.05	pes
1	conditions	2	22	24.225	2.73e-06	*	0.688

  - *„A one-way RM-ANOVA revealed a significant effect of the three KEYBOARDS on the WPM measure, $F(2, 22) = 24.225$, $p < .001$, $\eta_p^2 = 0.688$."*

Reporting the RM-ANOVA Output

- RM-ANOVA Output
  - Sphericity can be assumed, because otherwise the output would look a bit different…

	Effect	DFn	DFd	F	p	p<.05	pes
1	keyboard	2	22	24.225	2.73e-06	*	0.688

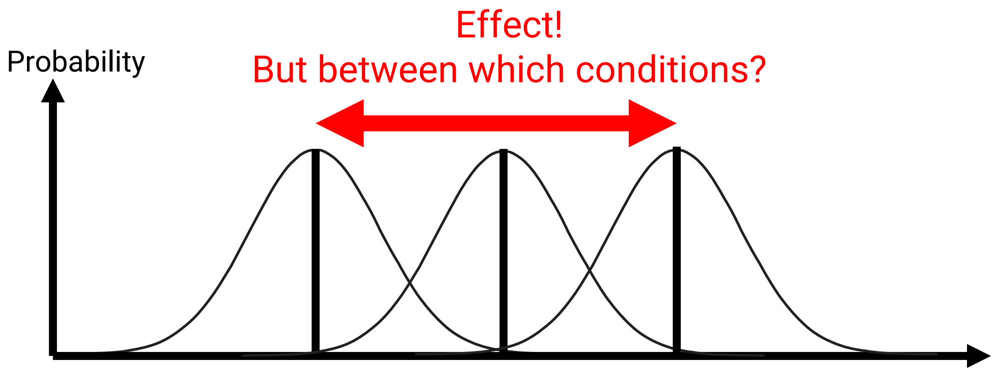
  - „*A one-way RM-ANOVA revealed a significant effect of the three* KEYBOARDS *on the WPM measure, F(2, 22) = 24.225, p < .001,* η_p^2 *= 0.688.*“

Reporting the RM-ANOVA Output

- RM-ANOVA Output
  - Sphericity can be assumed, because otherwise the output would look a bit different…

	Effect	DFn	DFd	F	p	p<.05	pes
1	conditions	2	22	24.225	2.73e-06	*	0.688

  - *„A one-way RM-ANOVA revealed a significant effect of the three* KEYBOARDS *on the WPM measure, F(2, 22) = 24.225, p < .001,* η_p^2 *= 0.688."*



Post Hoc Analysis

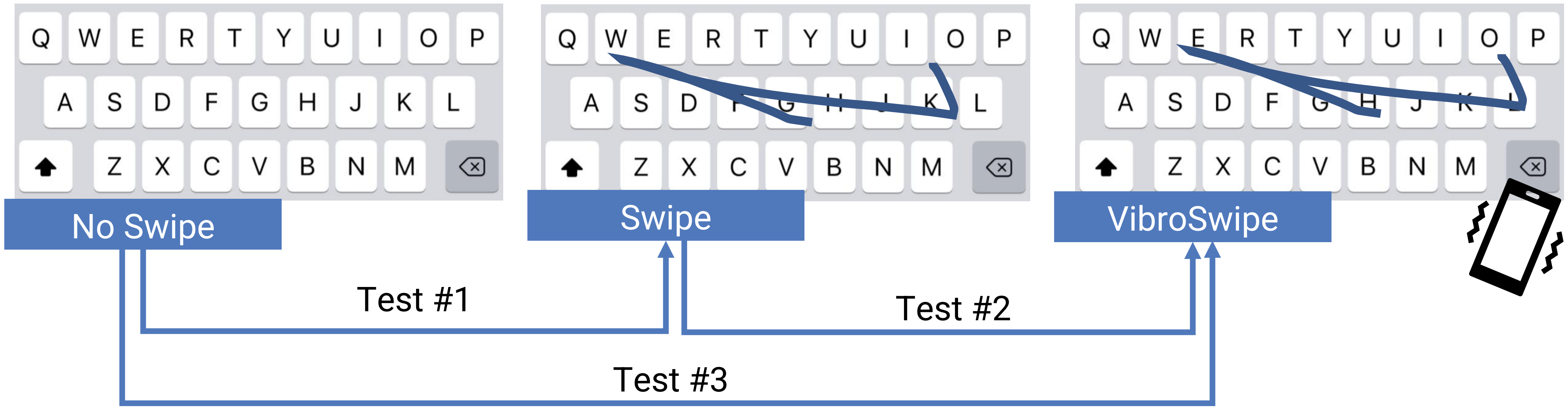
- **When the ANOVA revealed an effect, we can run post hoc analysis** to show between which conditions significant differences exists
  - (We must still consider if we need parametrical or nonparametrical tests)
  - Parametrical data: paired t-tests
  - Nonparametrical data: paired Wilcoxon signed-rank tests

R

```r
pairwise.t.test(wpm, conditions, paired = TRUE)
```

BIG PROBLEM!!!

The Number of Tests



= 3 TESTS!

Familywise Error Rate (FWER)

- **Too many tests** (because of too many conditions) **increase the probability of Type I errors**. An estimation of the FWER is:

$$F \leq 1 - (1 - \alpha)^c$$

  - α = alpha level for an individual test (e.g., 0.05), c = number of tests
- For example:
  - with an alpha level of 5% and a series of **3 tests**, the FWER is:

$$F = 1 - (1 - 0.05)^3 = .142 = 14\%$$

  - with an alpha level of 5% and a series of **10 tests**, the FWER is:

$$F = 1 - (1 - 0.05)^{10} = .401 = 40\%$$

- very high

Statisticians must correct this….

- **Bonferroni correction**: *„Divide the alpha level by the number of tests you're running and apply that alpha level to each individual test."*
  - e.g., if your $\alpha = .05$ and you are running e.g., 3 tests (because of three conditions), then each test will have an alpha level of $\frac{0.05}{3} = 0.017$
  - Statistical tests auto-apply the new alpha level to each test for finding p-values
    - In this example, the p-value would have to be 0.017 and decreased statistical significance

R

```
pairwise.t.test(wpm, conditions, p.adj = "bonf", paired = TRUE)
```

multiple tests means: we need Bonferroni correct p-value adjustment

Reporting the RM-ANOVA and Post hoc test

- Post hoc Output

	No Swipe	Swipe
Swipe	0.00012	-
Vibro Swipe	0.00022	1.00000

  - *„A one-way RM-ANOVA revealed a significant effect of the three KEYBOARDS on the WPM measure, F(2, 26) = 25.124, p < .001, η_p^2 = 0.659. Pairwise post hoc comparisons using Bonferroni-corrected t-tests revealed significant differences between No Swipe and Swipe (p), No Swipe and Vibro Swipe (p), however, not between Swipe and Vibro Swipe (p).“*

Reporting the RM-ANOVA and Post hoc test

- Post hoc Output

	No Swipe	Swipe
Swipe	0.00012	-
Vibro Swipe	0.00022	1.00000

› *„A one-way RM-ANOVA revealed a significant effect of the three* K*EYBOARDS on the WPM measure, F(2, 26) = 25.124, p < .001,* η_p^2 *= 0.659. Pairwise post hoc comparisons using Bonferroni-corrected t-tests revealed significant differences between No Swipe and Swipe (p < .001), No Swipe and Vibro Swipe (p < .001), however, not between Swipe and Vibro Swipe (p = 1.000).“*

Reporting the RM-ANOVA and Post hoc test

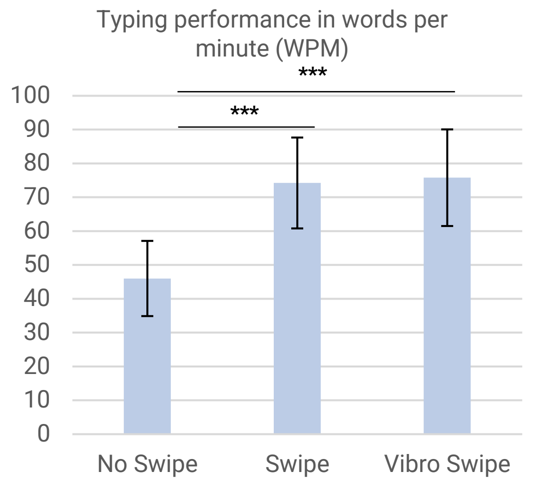


Figure 3: Average typing speeds without Swipe, with Swipe, and with VibroSwipe. Error bars show standard deviation.

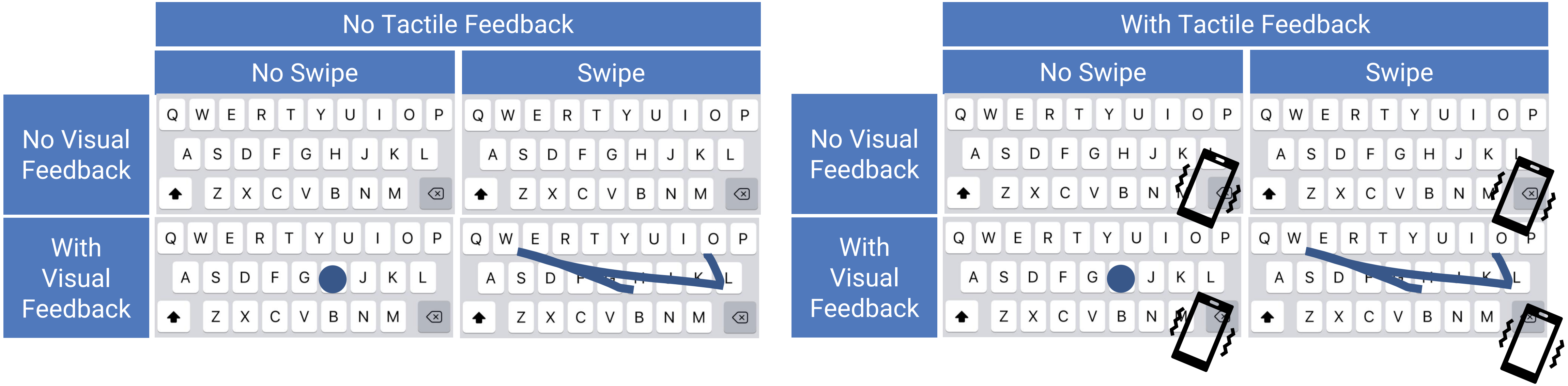
- Post hoc Output

	No Swipe	Swipe
Swipe	0.00012	-
Vibro Swipe	0.00022	1.00000

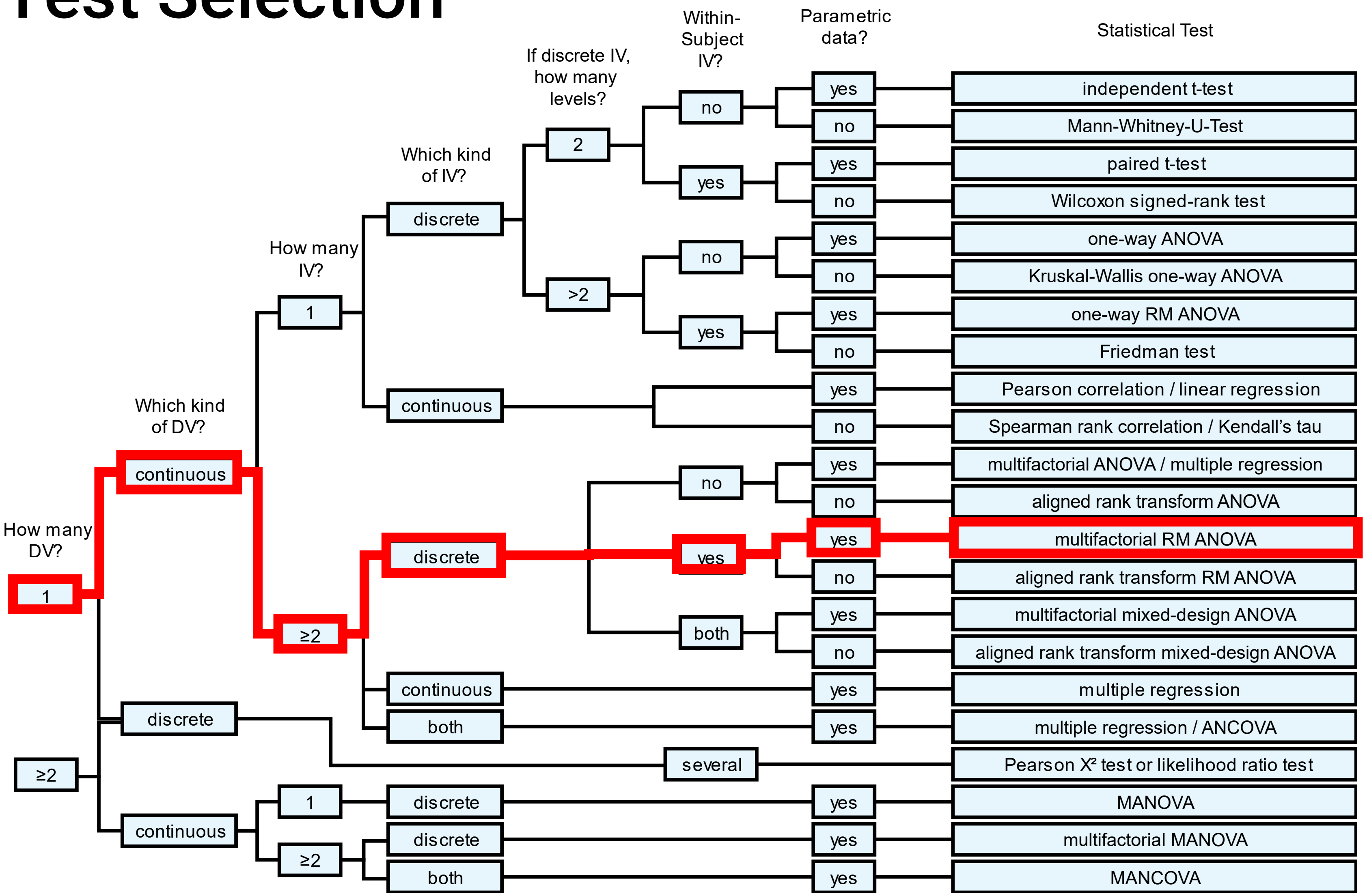
  - *„A one-way RM-ANOVA revealed a significant effect of the three KEYBOARDS on the WPM measure, F(2, 26) = 25.124, p < .001, η_p^2 = 0.659. Pairwise post hoc comparisons using Bonferroni-corrected t-tests revealed significant differences between No Swipe and Swipe (p < .001), No Swipe and Vibro Swipe (p < .001), however, not between Swipe and Vibro Swipe (p = 1.000). Means and standard deviations are shown in Figure 3. Thus, the analysis indicates that typing with Swipe increases the performance, however, adding vibration feedback does not further support the users' typing speed with Swipe.“*

Multi-Factorial Study Design

- Next example: **SWIPE × VISUAL FEEDBACK × TACTILE FEEDBACK**
- 16 participants in a balanced Latin square
- What is the effect on typing performance? Can we evaluate this?



Test Selection

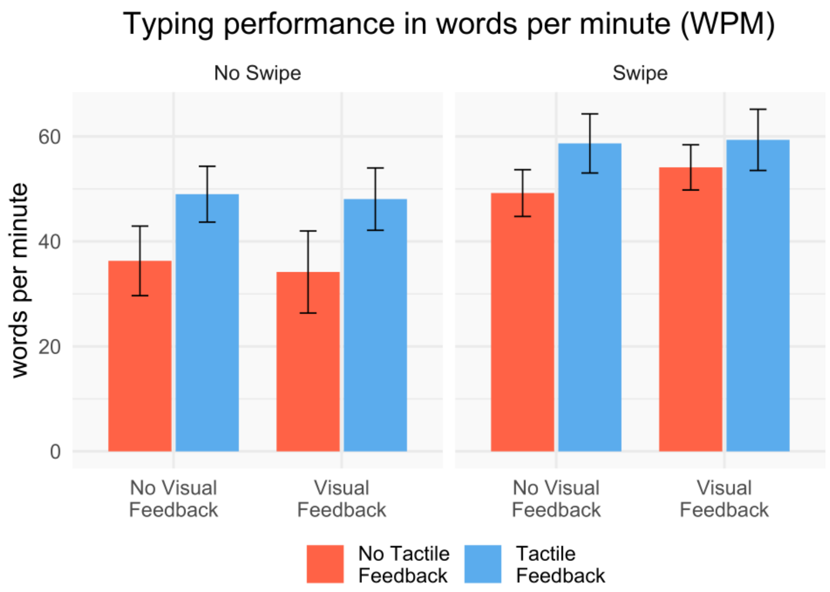


Three-Way RM ANOVA

R

```
library(rstatix)
anova <- anova_test(data = df, dv = wpm, wid = subjectID, within = c(swipe, visual, tactile), effect.size = "pes")
get_anova_table(anova, correction = "auto")
```

Effect	DFn	DFd	F	p	p<.05		pes
1	swipe	1	15	63.590	8.96e-07	*	0.809
2	visual	1	15	0.145	7.09e-01		0.010
3	tactile	1	15	21.494	3.23e-04	*	0.589
4	swipe:visual	1	15	0.952	3.45e-01		0.060
5	**swipe:tactile**	**1**	**15**	**7.227**	**1.70e-02**	*	**0.325**
6	visual:tactile	1	15	0.132	7.22e-01		0.009
7	swipe:visual:tactile	1	15	0.308	5.87e-01		0.020



- **Main Effect of Swipe**. **Main Effect of Tactile**. **No main effect of Visual**.
- **Interaction Effect: Swipe and Tactile combined increases the WPM!** Higher effect size: Swipe

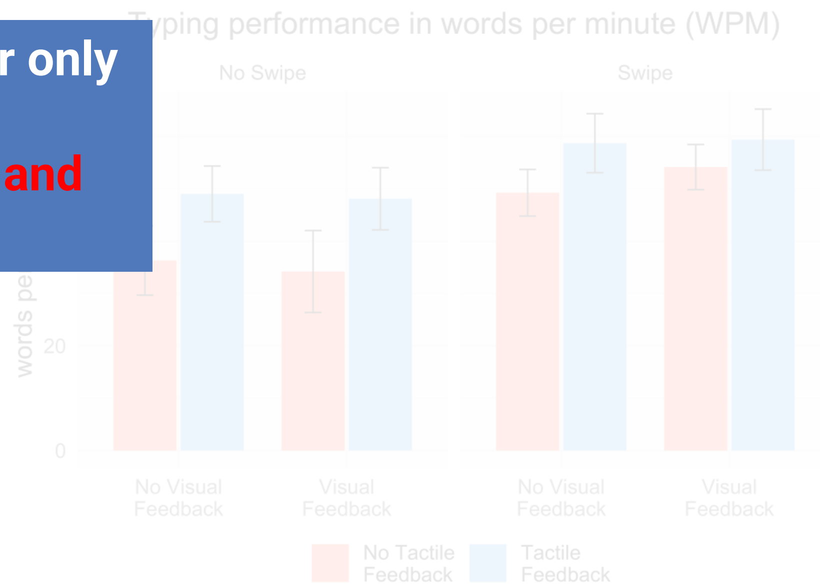
Three-Way RM ANOVA

R

```
library(rstatix)
anova <- anova_test(data = df, dv = wpm, wid = subjectID, within = c(swipe, visual, tactile), effect.size = "pes")
get_anova_table(anova, correction = "auto")
```

Once you have your study design, you later only need to change the dependent variable!
-Of course, you **check again for data type and normal distribution**, etc.

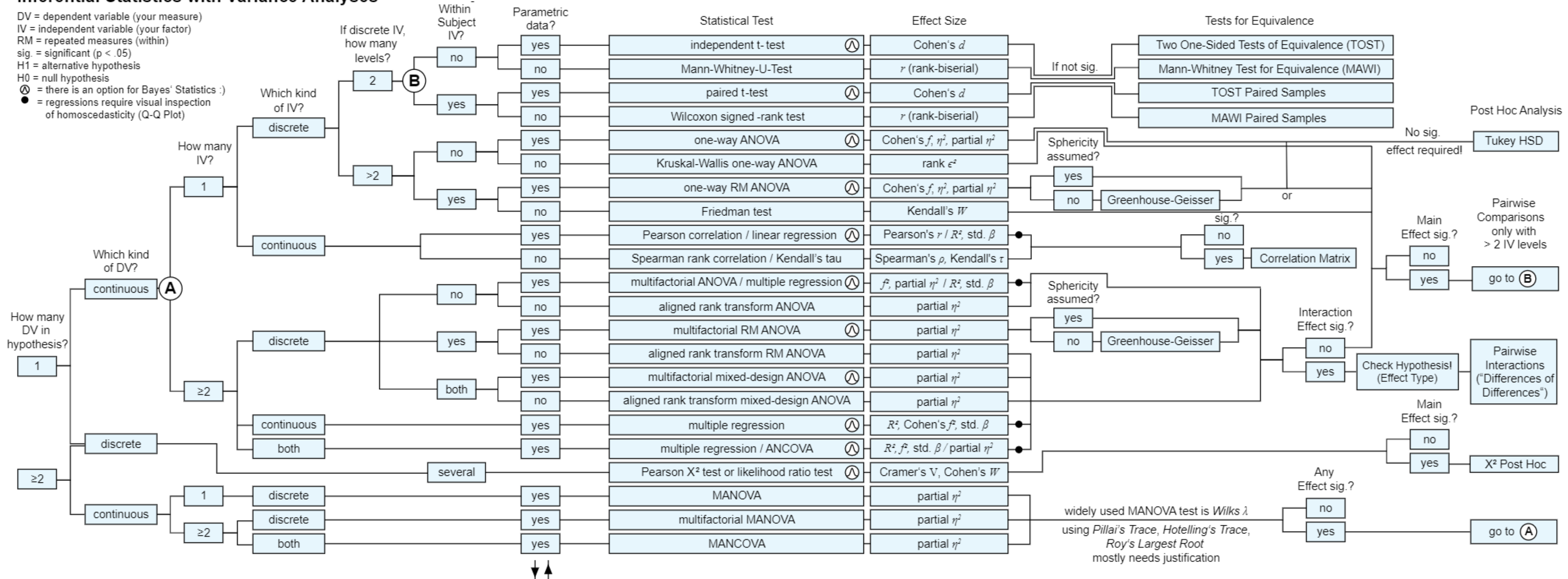
Effect	DFn	DFd				
1	swipe	1				
2	visual	1				
3	tactile	1				
4	swipe:visual	1	15	0.952	3.45e-01		0.060
5	**swipe:tactile**	**1**	**15**	**7.227**	**1.70e-02**	*	**0.325**
6	visual:tactile	1	15	0.132	7.22e-01		0.009
7	swipe:visual:tactile	1	15	0.308	5.87e-01		0.020

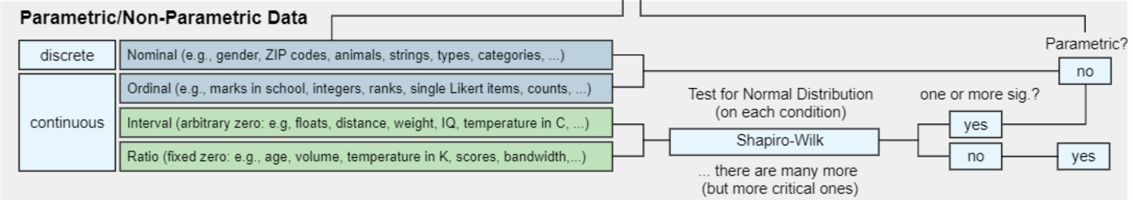


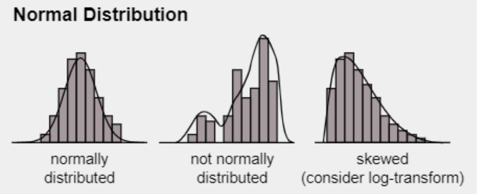
- Main Effect of Swipe. Main Effect of Tactile. No main effect of Visual.
- Interaction Effect: Swipe and Tactile combined increases the WPM!

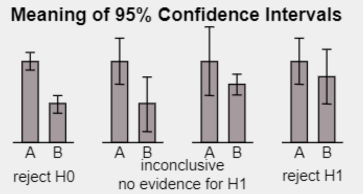
Inferential Statistics with Variance Analyses

DV = dependent variable (your measure)
IV = independent variable (your factor)
RM = repeated measures (within)
sig. = significant (p < .05)
H1 = alternative hypothesis
H0 = null hypothesis
Ⓐ = there is an option for Bayes' Statistics :)
● = regressions require visual inspection of homoscedasticity (Q-Q Plot)









Interpretation of Effect Sizes

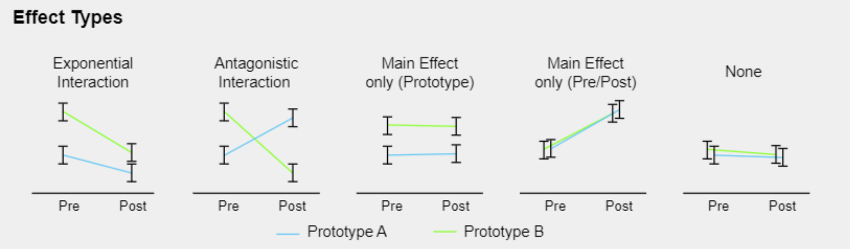
	small	medium	large
Cohen's d	0.20	0.50	0.80
r, Cohen's/Kendall's W	0.10	0.30	0.50
Cohen's f	0.10	0.25	0.40
Cohen's f^2	0.02	0.15	0.35
η^2	0.02	0.13	0.26
rank ϵ^2, partial η^2	0.01	0.06	0.14

Kendall (1939), Cohen (1988)

Interpretation of Correlations / Fits

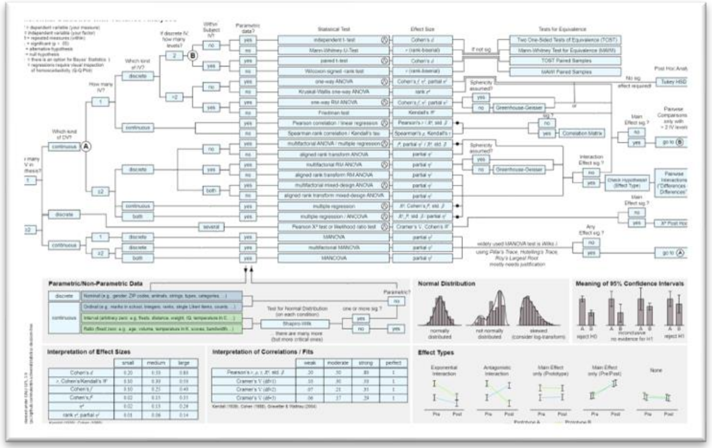
	weak	moderate	strong	perfect
Pearson's r, ρ, τ, R^2, std. β	.20	.50	.80	1
Cramer's V (df=1)	.10	.30	.50	1
Cramer's V (df=2)	.07	.21	.35	1
Cramer's V (df=3)	.06	.17	.29	1

Kendall (1939), Cohen (1988), Gravetter & Wallnau (2004)



Exercise: Find your Test

- **When you want answer your research question with quantitative methods**, use our statistical decision tree to find your test
  - https://hci-studies.org/ → Statistical Decision Tree
  - The tree is the ultimate thing
- **Discuss in your team**:
  - Which measures do you have?
  - Which statistical test is the correct one for the measure?
  - Write down all the tests (e.g., „Friedman test“)
  - Google the tests (e.g., „Friedman test in R“)
    - https://www.datanovia.com/en/lessons/friedman-test-in-r/
    - If there is a result, you can be sure that you can evaluate your user study!



What‘s next? – An Example

- **Precision vs Accuarcy**
  - Why is low accuracy better than low precision?
  - What can we do with that knowledge?

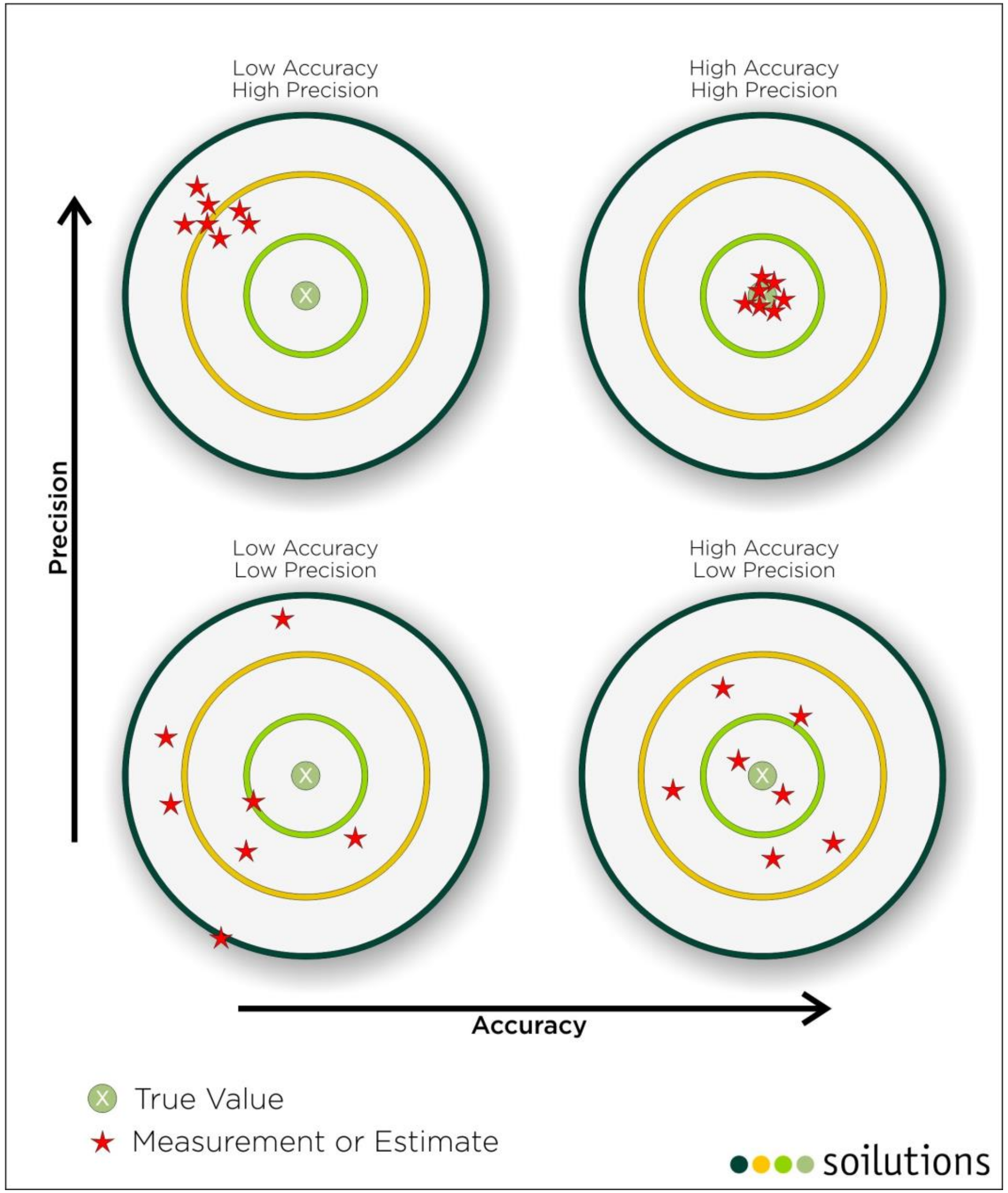


Image from: https://www.soilutions.co.uk/tag/precision-verses-accuracy/

Did we forget something?

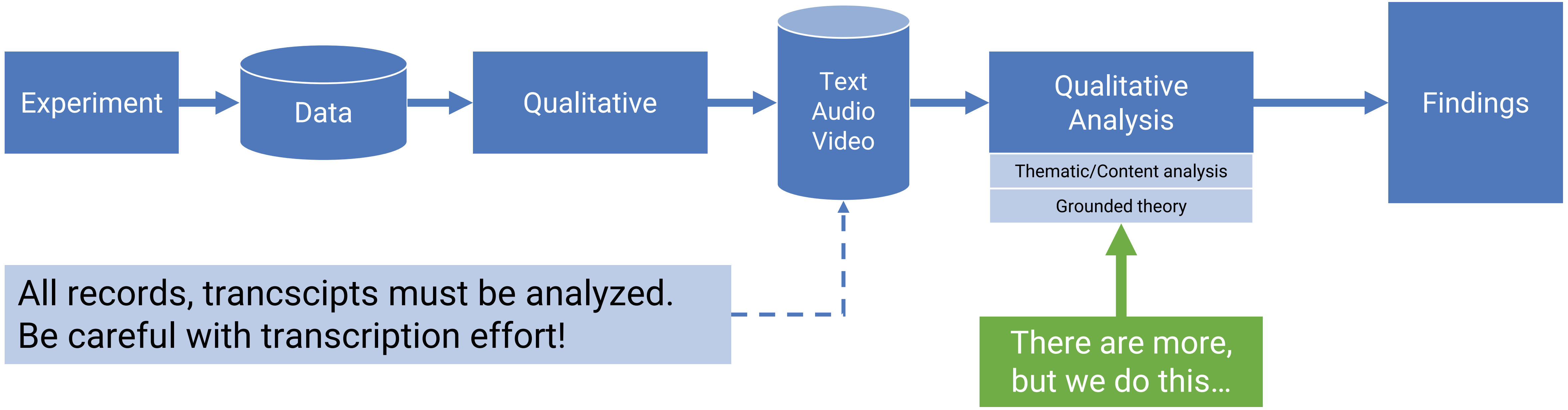
Evaluating Qualitative Data

Human-Computer Interaction Exercise

Qualitative Data

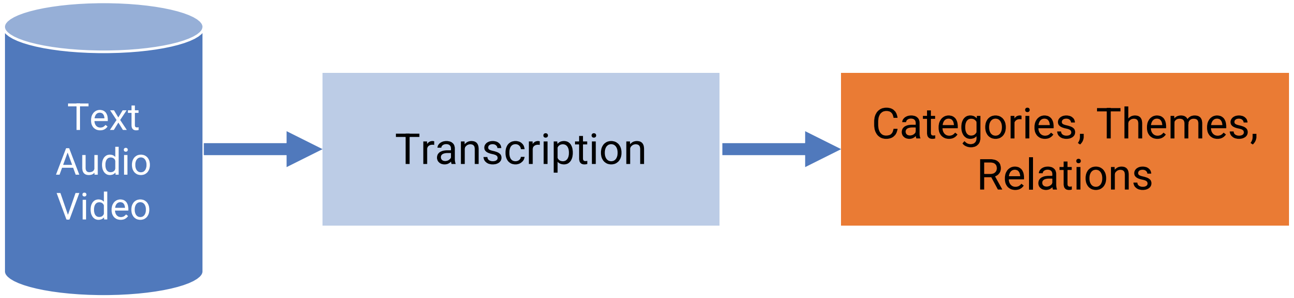
- Are **not based on hypotheses** and **cannot be tested with quantitative methods**
  - You can „count“ stuff (e.g., opinions, answers, words, utterances, observations, …) but this belongs to a quantitative evaluations or additionally include a **content analysis**!
- Contain
  - **feedback**: perspectives, opinions, thoughts, subjective impressions, anticipations, …
  - **observations**: behavior, gestures, emotions, person-object or person-person relations,…
- Explain
  - **why something happens**
  - **what people think**
  - **relationsships in your data**
- Important: **the analysis does not discriminate how often (or rarely) cases in the data occur**!

Qualitative Data Analysis

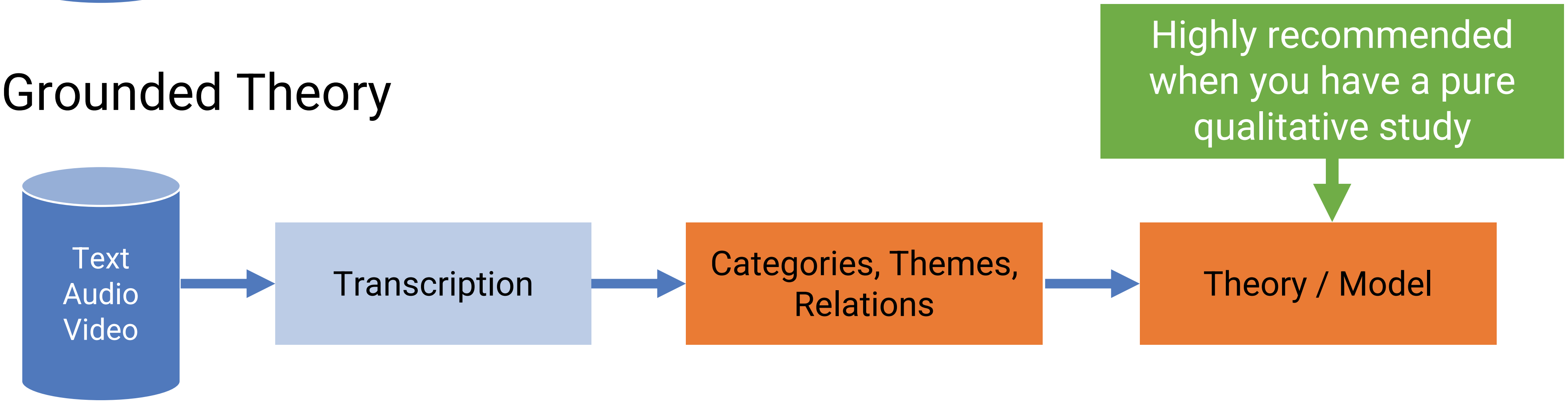


Thematic/Content Analysis vs Grounded Theory

- Thematic Analysis



- Grounded Theory



Transcribing

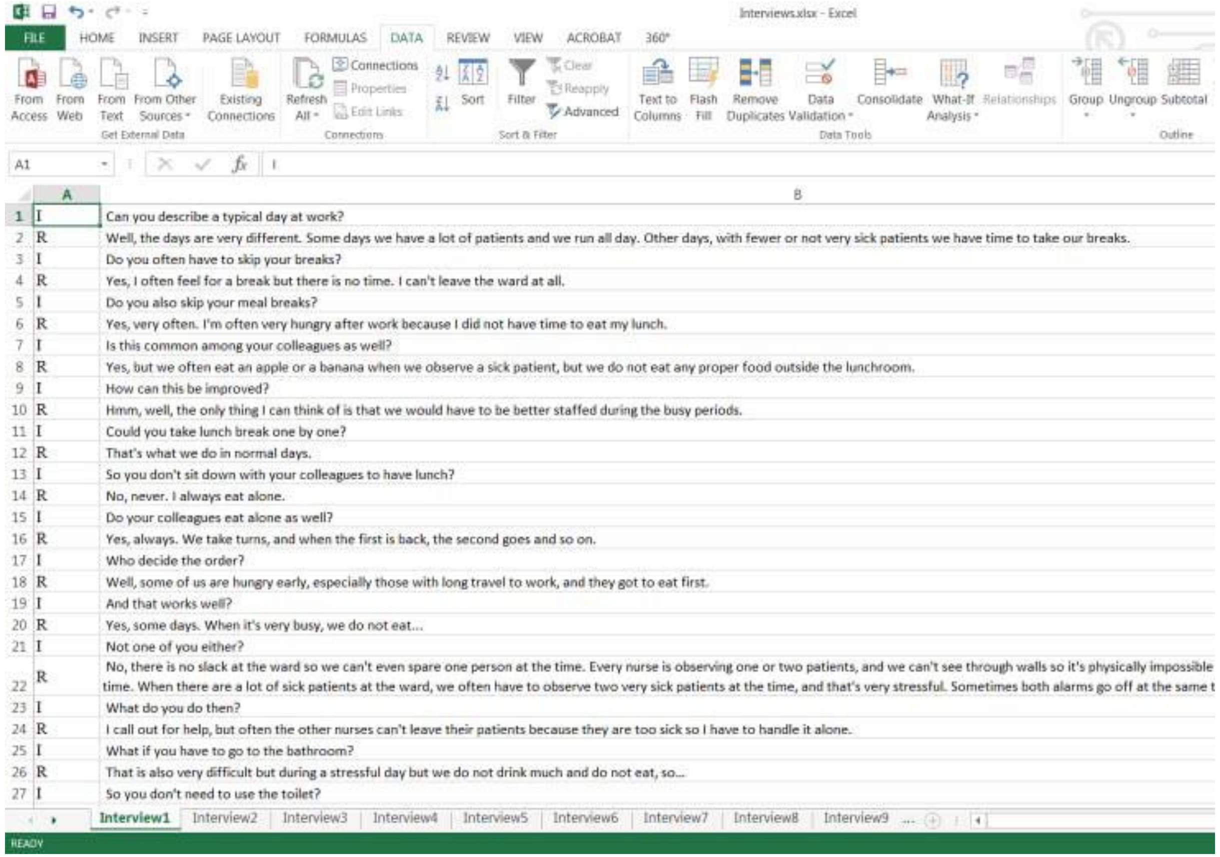
- **Only transcripts can be coded and analyzed**
  - sentence-wise, line-wise, paragraph-wise
- **Speech** must be transcripted as clear, readable verbatim text
  - e.g., audio raw data → [timecode, participant, condition, **text**] → table
- **Observations** must be annotated
  - e.g., video raw data → [timecode, participant , condition, **annotation**] → table
- **Transcribed text** must be translated into English
  - You can translate **while** transcription – but be consistent
  - You can translate **after** transcription e.g., automatically by a native speaker or professional program
    - but someone who can speak English must proofread it
- More Tips: http://www.qualitative-researcher.com/qualitative-analysis/using-excel-for-qualitative-data-analysis/

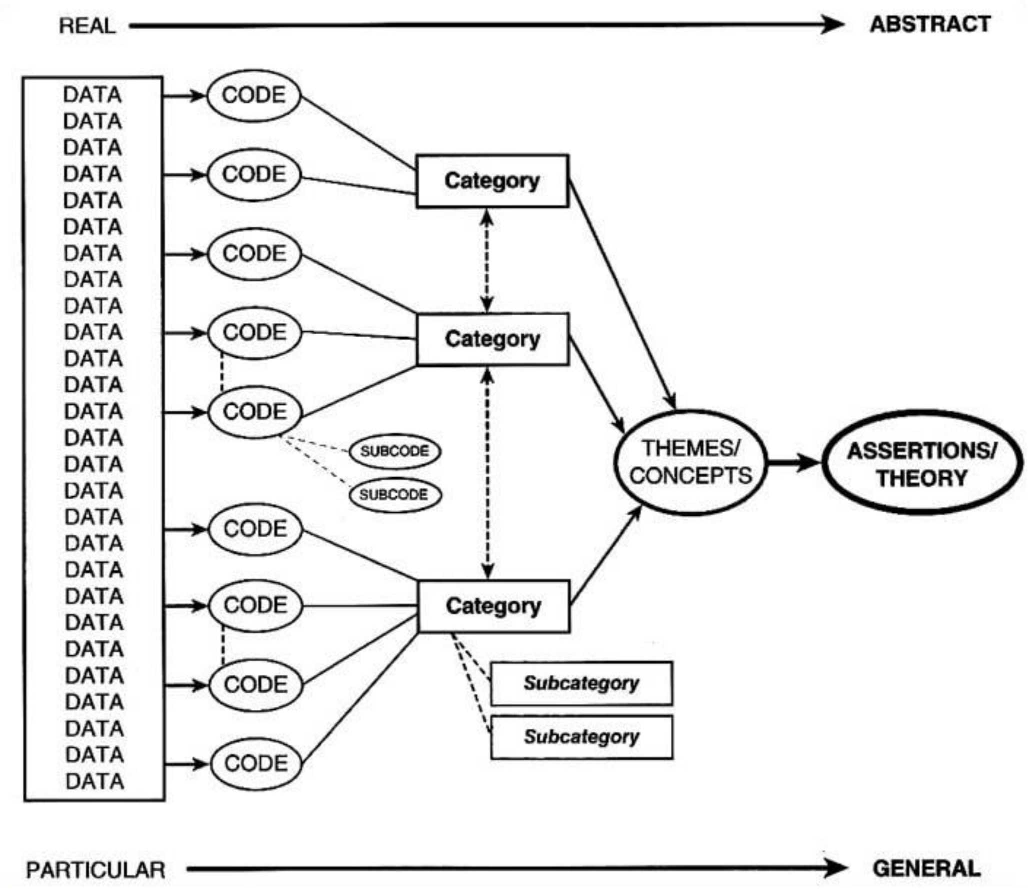
Transcribed Video Protocol: Example

Add translation, codes, themes

ID	Time	Q/R	SubjectID	Gender	Condition	Transcribed Utterance (Video)	Observed Action (Video)
...							
130	13:12	R	4	m	Male	Unterseite der männlichen Hände wirkt sehr überzeugend. Sie wirken wie die eigenen.	Turning arms around several times.
131	13:13	R	4	m	Male	Die Behaarung der Hände bricht die Überzeugung.	Grabs the underside of the right arm.
132	13:13	R	4	m	Male	Die Sounds steigern Immersion immens.	
...							
137	13:28	R	4	m	Toon Hands	Die Hände gefallen eher mittelmäßig. Schlechter als die Männerhände. Fühlen sich überhaupt nicht an wie die eigenen Hände. Hände werden nach kurzer zeit "ausgeblendet". Leap bugs irritieren stark.	Shake hands in between.
...							
153	13:32	R	4	m	Abstract Hands	Abstraktionsgrad fühlt sich unnatürlich an. Nicht als eigene Hände akzeptierbar. Fehlende Flächen innerhalb der Hand irritieren enorm. Die schwebende Kugel innerhalb der Hand irritiert: da nicht befestigt.	
154	13:33		4	m	Abstract Hands		Pausing to take off glasses.
...							
167	13:46	R	4	m	Robot Hands	Besser als die Kapseln und Toonhände, dank dem „Volumen" der Hände. Zwischen abstrakt und realistisch. Männerhände fühlen sich echter an.	
...							
177	13:57	R	4	m	Androgyne Hands	Glaubwürdiger, da näher an eigenen Händen.	
...							
187	13:05	R	4	m	Female Hands	Wirken weniger glaubwürdig als androgyne. Fingernägel irritieren enorm. Sogar weniger glaubwürdig als Robohände. Sound zur implikation.	

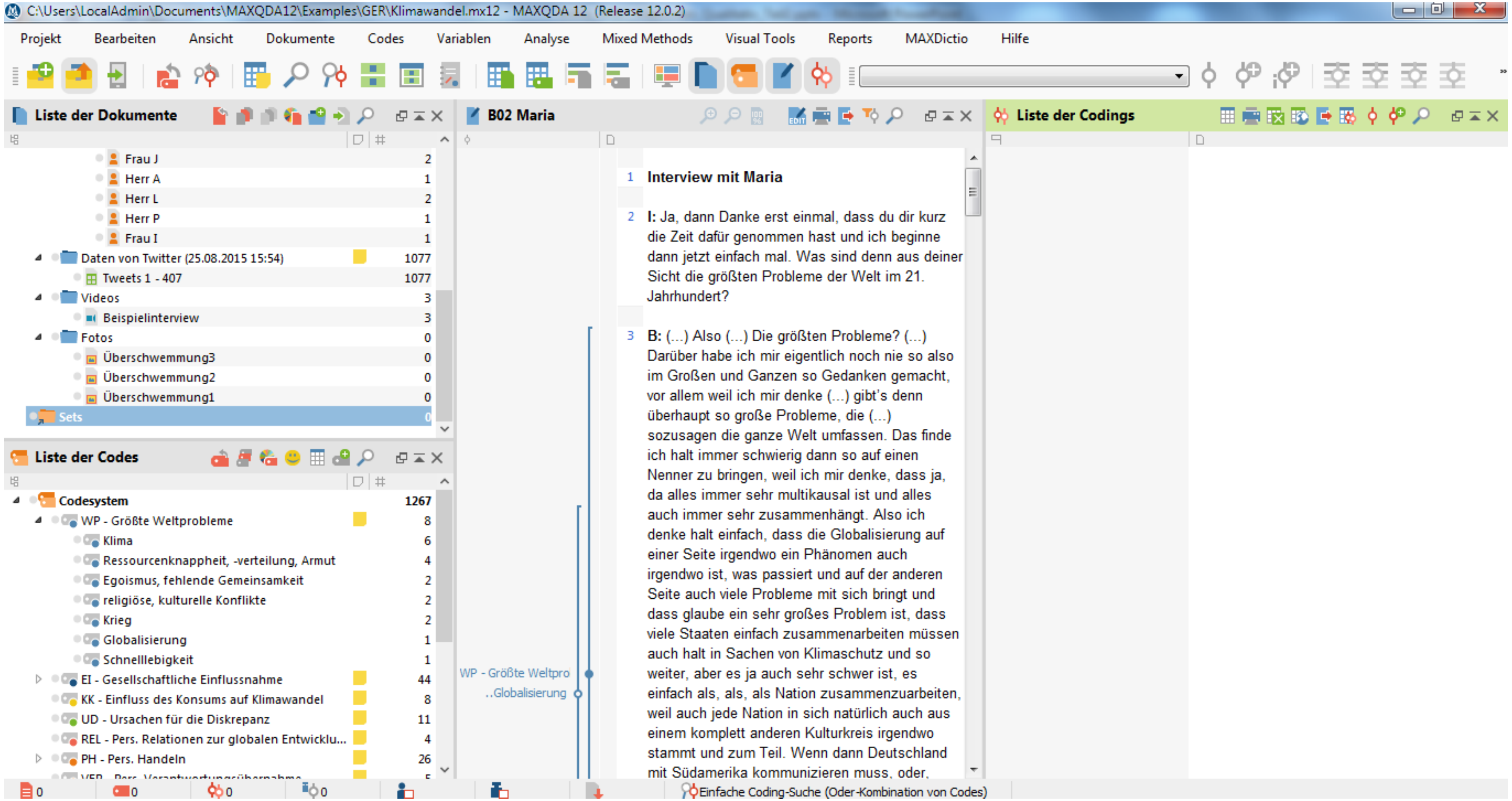
Example of Coding in Excel



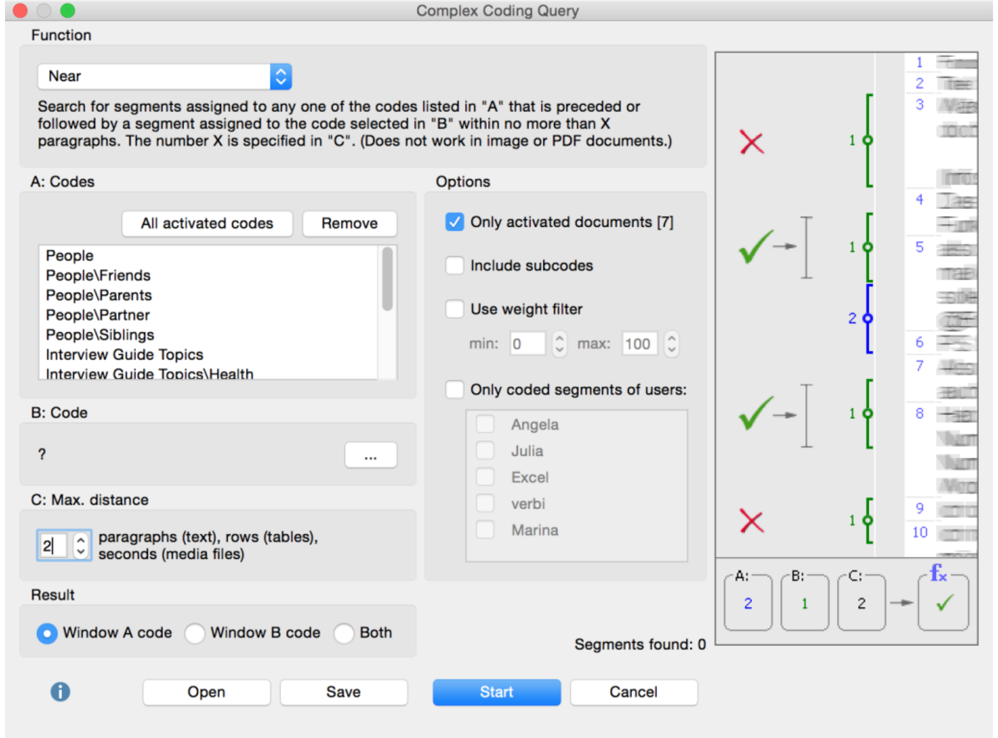


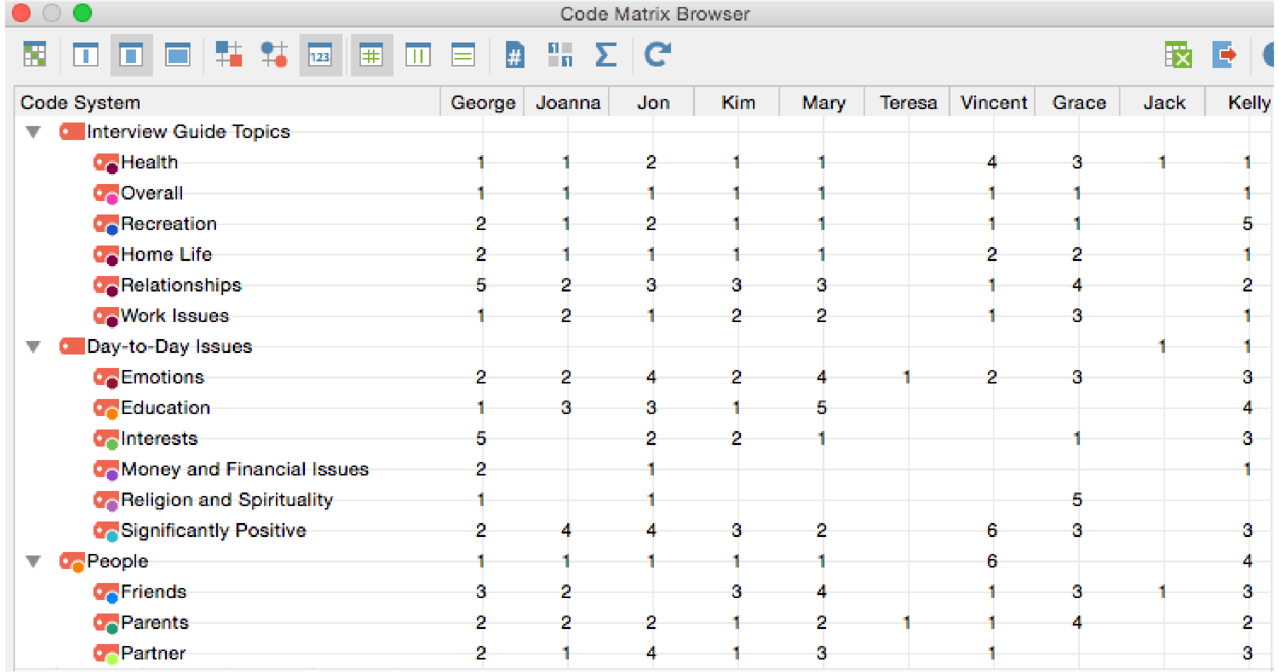
Ose, S. O. (2016). Using Excel and Word to Structure Qualitative Data. Journal of Applied Social Science, 10(2), 147–162. https://doi.org/10.1177/1936724416664948
Saldaña, J. (2015). The coding manual for qualitative researchers (Third ed.): Sage.

MAXQDA (Expensive)



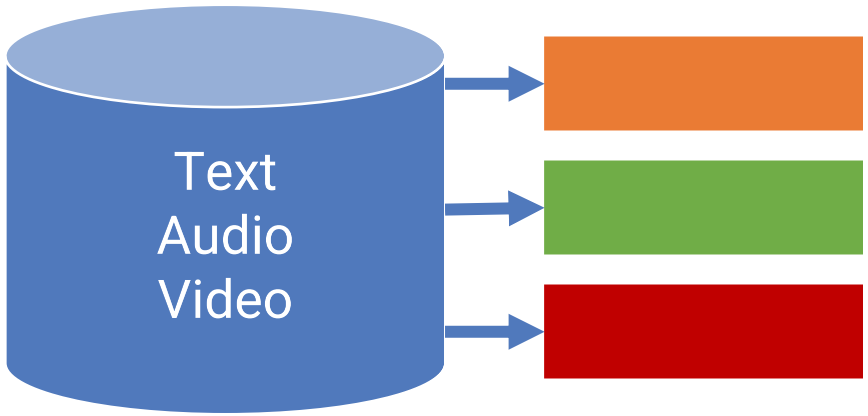
MAXQDA: Code-Matrix-Browser

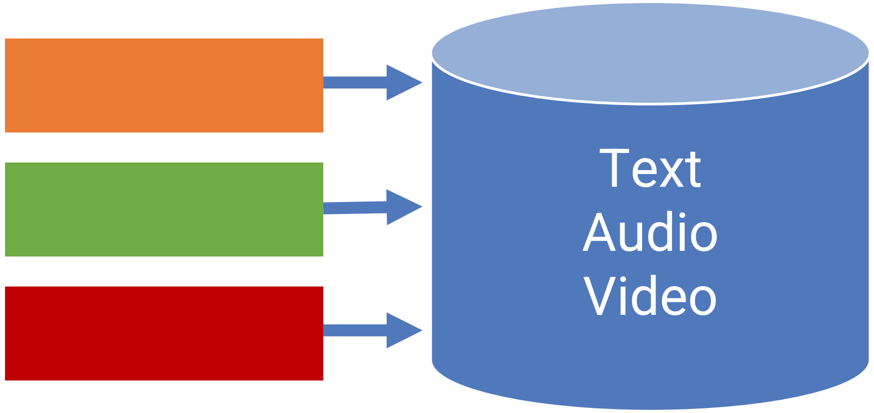




Inductive vs Deductive Approach

- **Inductive Analysis**
  - Very easy and the traditional qualitative anaylsis approach
  - Categories are derived from the data
  - You do not need any prior knowledge
- **Deductive Analysis**
  - Structure of the analysis & categories based on prior knowledge
  - Mostly used to test a theory
    - To test a hypothesis, we must gain a quantified outcome (e.g., a categorization matrix)
  - If prior knowledge (or the theory) does not match, we need an inductive analysis again





Example: Inductive Analysis

- A qualitative study investigating the acceptability of the Google Glass eyewear computer to people with Parkinson's disease (PD) [1]
  - 'Hands on the Glass' workshop: 5 patients, 2 therapeutists
  - 5-day field study with 4 patients: 5 tasks, interviews on the experiences (via phone)
  - *The workshop and interviews were audio recorded, which were transcribed and anonymised for later analysis.*
  - Target concept: acceptability
  - Target group: people with parkinson

P3 (m): *"my voice wasn't always working…it came up saying 'try again'"*

P5 (f): *"the fact that it wasn't recognising what I wanted was very irritating and very frustrating"*

P2 (m): *"it's better than a phone. With Parkinson's you can't text because you can't hit the buttons. With the glass you would just talk, you can see what you're doing, it's just instant"*

[1] McNaney, R., Vines, J., Roggen, D., Balaam, M., Zhang, P., Poliakov, I., & Olivier, P. (2014). Exploring the Acceptability of Google Glass As an Everyday Assistive Device for People with Parkinson's. https://dl.acm.org/doi/10.1145/2556288.2557092

Example: Inductive Analysis

- *“We conducted an inductive thematic analysis [4] on transcribed data by coding it at the sentence to paragraph level and drawing out themes across the data set. The quotes used to describe themes illustrate themes drawn across all participants and data, with the exception of instances where we focus on individual differences (e.g., the ‘Wearing the Glass out and about’ section).”* [1]
  - Workshop findings:
    - Issues and Frustrations, Confidence and Safety, Security and Vulnerability, Privacy
  - Field study findings:
    - Wearing Glass while ‘Out and About’, Frustrations, Appreciating Glass Wearing
  - Code common themes of the data set: Overview about the findings

[1] McNaney, R., Vines, J., Roggen, D., Balaam, M., Zhang, P., Poliakov, I., & Olivier, P. (2014). Exploring the Acceptability of Google Glass As an Everyday Assistive Device for People with Parkinson’s. https://dl.acm.org/doi/10.1145/2556288.2557092

Deductive Analysis

- **Categories are given**
  - e.g., by another researcher, the theory, the research question
- **Allocation** of the answers **to the codes** (“categories” or “themes”)
  - a categorization matrix can quantify the outcome
    - see **repertory grid**
  - useful in **elicitation studies**
    - see **agreement rate**
- What to do with content that is not in the matrix?
  - Create extra categories analogous to the inductive procedure and repeat
  - Ignore it

Quality Criteria

- **Everything is recorded**, clearly **documented**, and **justified** (process steps and code criteria)
  - Path from data to criteria is transparent (traceability)
  - Multiple data sources confirm codes (triangulation)
    - **Intra-encoder Reliability:** Consistency when the encoder encodes the same data again
    - **Inter-encoder Reliability:** Consistency when different encoders encode the same data
- Dimensions of **Quality Assessment**
  - **Agreement** : Matched Cases / Total Cases
  - **Cohen's Kappa**: Agreement relativized by the coincidence that is possible by chance
  - K = (Pa - Pc)/(1 - Pc)
    - Pa = Percentage of Matches
    - Pc = Percentage of matches that can come about by chance

Example: Deductive Analysis

- Part of an ongoing project: LGBTQ+ individuals' information behaviour in online communities
  - *"**We applied the method of deductive thematic analysis** informed by Aronson [1]. Deductive thematic analysis was chosen because it facilitates the interpretation of identifiable themes and patterns of behaviour**. The data corpus was read a number of times and were copied into a document. A detailed reading was carried out where initial thoughts were noted.** These notes are related to concepts and phrases that the researcher considered relevant to relatedness needs. The initial **notes were transformed into the main themes** of each posts and a list of specific themes were generated. At this stage the data were given to another rater who also generated a list of themes from the data. The **two raters then discussed and negotiated the findings until an agreement was reached** as to the validity and appropriateness of each theme. The data were again re-read and the themes were refined into more specific clusters based on RMT. Statements from the raw data were extracted to provide evidence of the existence of each theme within certain categories."* [2]

[1] Aronson, J. 1994. A pragmatic view of thematic analysis. Qualitative Report, 2(1), 1-3.

[2] Romy Menghao Jia, Jia Tina Du and Yuxiang Chris Zhao. 2021. Needs for Relatedness: LGBTQ+ Individuals' Information Seeking and Sharing in an Online Community. In Proceedings of 2021 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR'21), March 2021, Canberra, Australia. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3406522.3446040

Example: Deductive Analysis

Table 1. Categories and emergent themes of relatedness needs

Main categories	Emergent themes	Frequency
Being cared about	Community belongingness	32
	Community support	22
	Others' acceptance	20
Caring for others	Advice and experience sharing	6
Building and maintaining relationships	Romantic or sexual relationships	51
	Looking for relationships (sex, friend, date)	25
	In total	156

Romy Menghao Jia, Jia Tina Du and Yuxiang Chris Zhao. 2021. Needs for Relatedness: LGBTQ+ Individuals' Information Seeking and Sharing in an Online Community. In Proceedings of 2021 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR'21), March 2021, Canberra, Australia. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3406522.3446040

Trustworthiness

- Evaluators can have **different interpretations** of the data
- Ways of establishing **trustworthiness**:
  - member check, interviewer corroboration, prolonged engagement, peer debriefing, negative case analysis
  - …
- **Concepts** on which the researchers do not agree with each other can be **resolved through discussion, a third researcher, or left open**

Grounded Theory

- Not a „theory“ but **a methodology** (≠ method)

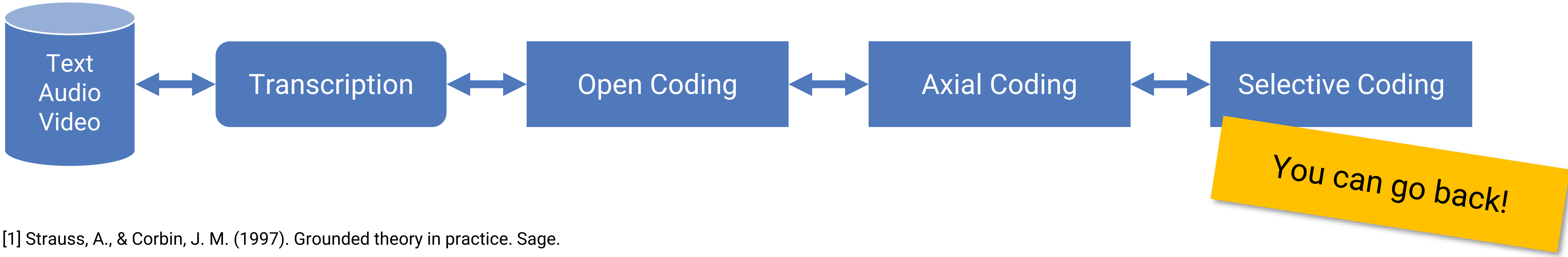
  → a systematic, theoretical analysis of qualitative (and sometimes even quantative) data
- Mostly used in the field of **data analysis of qualitative feedback**
- Grounded theory is **the process of deriving a high level of conceptual abstraction** by assigning general concepts (codes) to singular incidences, relating, and identifying the core concept.
- The goal of grounded theory is to systematically derive a clear and testable hypothesis or well-grounded theory

Grounded Theory and Thematic Analysis

- In both, **you take a body of data**, such as
  - interview transcripts, observations, research literature, any protocols, videos, audio, images, social media posts, observation, posts, articles, research papers,…
- In both, **you search for**
  - Opinions, phenomenons, artifacts in the data, events, data patterns
  - Any sampling does not depend on how often or how rarely cases in the data occur!
    - **One single incidence can lead to a new theory!**
- In both, **you identify common categories or themes** in your data
  - **Grounded Theory** continues the process inductively and deductively to derive relationship and **a new theory and more meaning** around the related categories or themes
    - **Requires a high level of abstraction**

Analysis Process

- Grounded theory analysis involves the following basic steps:
  1. **Data Collection** and Transcription
  2. **Open Coding**: Conceptual Labeling and Coding text
  3. **Axial Coding**: Finding Relationships between the Categories
  4. **Selective Coding**: Selection of Core Phenonenom and Theory



[1] Strauss, A., & Corbin, J. M. (1997). Grounded theory in practice. Sage.

Analysis Process: Example

RQ: „Why do people not click on some buttons on my interface?“

#1 „I did not click on the button because it has no functionality in this moment.“
#2 „The button seemed to be useless.“

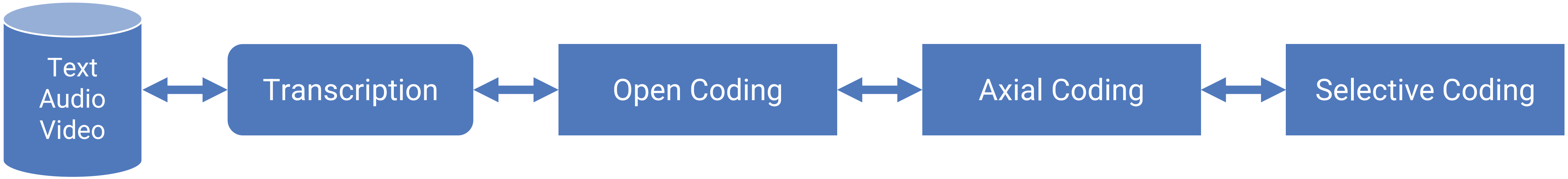
Code: UI Element Visibility (CF)
Code: Context awareness (CA)

Button should communicate functionality or must be invisible.

The button should only be visible if needed.
(Mini/Sub-Theory)

CF required when CA confirmes functionality. Visibility signals functionality.

Theoretical framework: “We only need the functionality in an interface that we need during the interaction.” (Core Concept)



Open Coding

- **Segment data** into meaningful expressions and describe them in a concept or theme
  - Existing annotations and concepts are attached to these expressions or create new relations (open coding)
- **Break down, understand the concept** and **develop categories** using open (W) questions
  - **What**? - Identify the underlying issue and the phenomenon
  - **Who**? - Identify the actors involved and the roles they play
  - **How**? - Identifying the aspects of phenomenon
  - **When**? How long? Where? - Time, course, location
  - **How much**? How long? - The intensity or duration
  - **Why**? - Identifying the reasons causing the phenomenon

Axial Coding

- Deductive analysis and focusing on the **phenomenon(s) under study using a template**
- Name all conditions and incidences related to that phenomenon
  - Context and structural conditions
  - Causes
  - Exceptions
- Assign all actions and interactional strategies directed at managing or handling the phenomenon
- **Consequences of the actions/interactions** related to the phenomenon

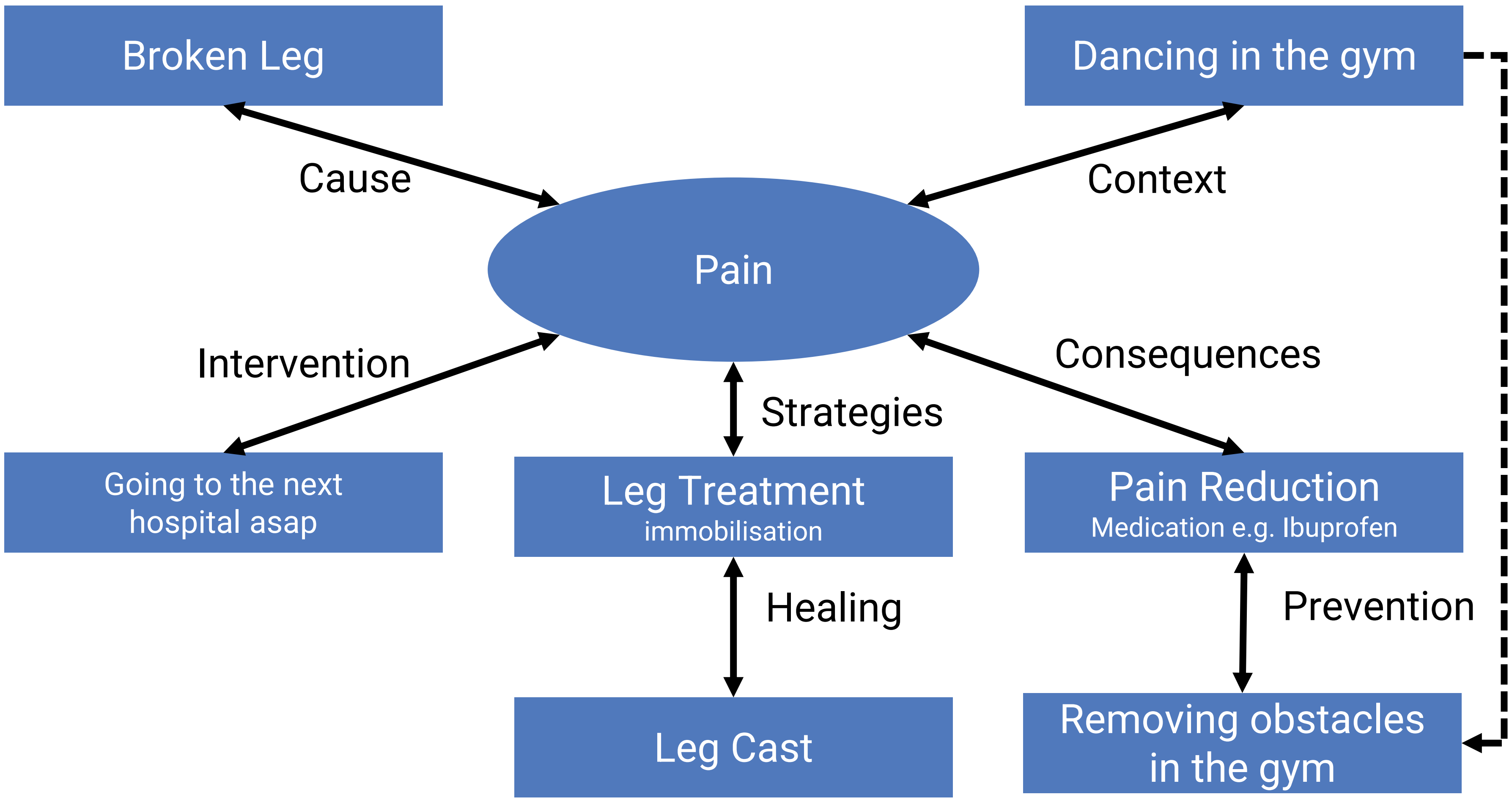
Raw Transcript. *"D6: The readability here is awful, but there is no way to escape from this (implementation). That is the standard (implementation). (...) indeed, it (the class) is not easy to ready"*
Code 1. developer mentions that the class readability is awful
Code 2. developer mentions that there is no way to escape from the analyzed implementation
Code 3. developer mentions that the analyzed implementation is the standard implementation
Code 4. developer accepts that the class is hard to read

We related the codes through *axial coding*. In this procedure, the codes were merged and grouped into more abstract categories, and the type of relation [44] was established. For instance, the previous codes were grouped into the following two categories:

Category 1. analysis of a non-functional requirement
Category 2. explanation for the existence of the symptom

Leonardo Sousa, Anderson Oliveira, Willian Oizumi, Simone Barbosa, Alessandro Garcia, Jaejoon Lee, Marcos Kalinowski, Rafael de Mello, Baldoino Fonseca, Roberto Oliveira, Carlos Lucena, and Rodrigo Paes. 2018. Identifying design problems in the source code: a grounded theory. In Proceedings of the 40th International Conference on Software Engineering (ICSE '18). Association for Computing Machinery, New York, NY, USA, 921–931. https://doi.org/10.1145/3180155.3180239

Example: Axial Coding of Pain



See Strauss/Corbin 1990, 99-107

Selective Coding

- **Integrate the different categories** that have been developed during axial coding into one cohesive theory or framework
- **Results from axial coding are further elaborated, integrated, and validated on an abstract level**
  - Is there an **overarching theory**?
  - What is the **overarching theory**?
- **Choose the core category** and **relate it with the other categories from axial coding**
  - If the core category is found, the story line of the research is set, and the researcher knows the central phenomenon of the research and can finally answer the research question and name the theory.

4.1 Properties of Emotional Exploration

The core category of 'Emotional Exploration' has six properties, which are summarised and explored below:

- It requires *expectations* to be appropriately set.
- The emotional landscape needs some level of *challenge* to be understood or 'traversed' (as we would a 'mechanical' landscape).
- One of the key tools for this is *ambiguity*.
- Emotional exploration results in a *mixed affect* emotional experience.
- The data here suggests that this emotional experience is *potentially more powerful* than that experienced in other media due to the player's participation in the diegesis and a certain level of *interactive vulnerability*.
- Emotional Exploration satisfies a need for *relatedness* in the player.

Leonardo Sousa, Anderson Oliveira, Willian Oizumi, Simone Barbosa, Alessandro Garcia, Jaejoon Lee, Marcos Kalinowski, Rafael de Mello, Baldoino Fonseca, Roberto Oliveira, Carlos Lucena, and Rodrigo Paes. 2018. Identifying design problems in the source code: a grounded theory. In Proceedings of the 40th International Conference on Software Engineering (ICSE '18). Association for Computing Machinery, New York, NY, USA, 921–931. https://doi.org/10.1145/3180155.3180239

When Grounded Theory?

- If you want to **derive theories** and even **testable hypotheses**:

  „Compared to the smartphone application, the voice interface improves the user experience, but increases the frustration level when errors occur.”

  “Women perceive higher levels of immersion than men with decreasing levels of avatar realism in virtual reality.”

  “Our machine-learning based model outperforms linear prediction in terms of precision and accuracy.”

- If you want to **derive a theoretical framework**
- When you have a mixed-method study and no idea how to structure the feedback
  - No. Then do a thematic analysis.
- When you only have a qualitative study and no idea what your contribution is
  - Yes. Then do it.

Summary

- Evaluating the results **is an essential part** of your study
- **Report all results** and **measures**
- **Quantitative Data**: objectively are being reported in descriptive **and** inferential statistics
  - **Descriptive Statistics**: text, plot, or table
  - **Inferential Statistics**: the correct statistical test and post hoc comparison for each of your measure
- **Qualitative Data**: subjectively analyze the feedback and observations using
  - **Thematic Analysis:** okay in mixed-methods
  - **Grounded Theory: h**ighly recommended to extent the contribution

Exercise: Find your Qualitative Analysis

- **Do you want to answer your research questions using qualitative methods?**
- Which concepts do you investigate?
- Create an example transcript fitting to your user study.
- How do you analyze the result?
- Who will be responsible for transcribing, coding, counter-checking?
- **Discuss in your team**

References

- Elo & Kyngäs (2008). The qualitative content analysis process. Journal Advanced Nursing. 62(1):107-15.
- Eliot, Susan (2011). Using Excel for Qualitative Data Analysis. Retrieved from http://www.qualitative-researcher.com/qualitative-analysis/using-excel-for-qualitative-data-analysis/
- Charmaz, K. (2006). Constructing Grounded Theory. A Practical Guide Through Qualitative Analysis. London: Sage.
- Glaser, B. & Strauss, A.(2010). Grounded Theory. Strategien qualitativer Forschung. Bern: Huber.
- Hof, C. & Weingarten (1979). Qualitative Sozialforschung. Stuttgart: Klett.
- Mayring, P. (2008). Qualitative Inhaltsanalyse. Grundlagen und Techniken. Weinheim: Beltz.
- McNaney, R., Vines, J., Roggen, D., Balaam, M., Zhang, P., Poliakov, I., & Olivier, P. (2014). Exploring the Acceptability of Google Glass As an Everyday Assistive Device for People with Parkinson’s. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 2551–2554). New York, NY, USA: ACM. doi:10.1145/2556288.2557092
- Krippendorf , K. (1980). Content Analysis. An Introduction to its methodology. Newbury Park: Sage Publications.
- Razavi, M. N., & Iverson, L. (2006). A grounded theory of information sharing behavior in a personal learning space. Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work - CSCW ’06, 459. doi:10.1145/1180875.1180946