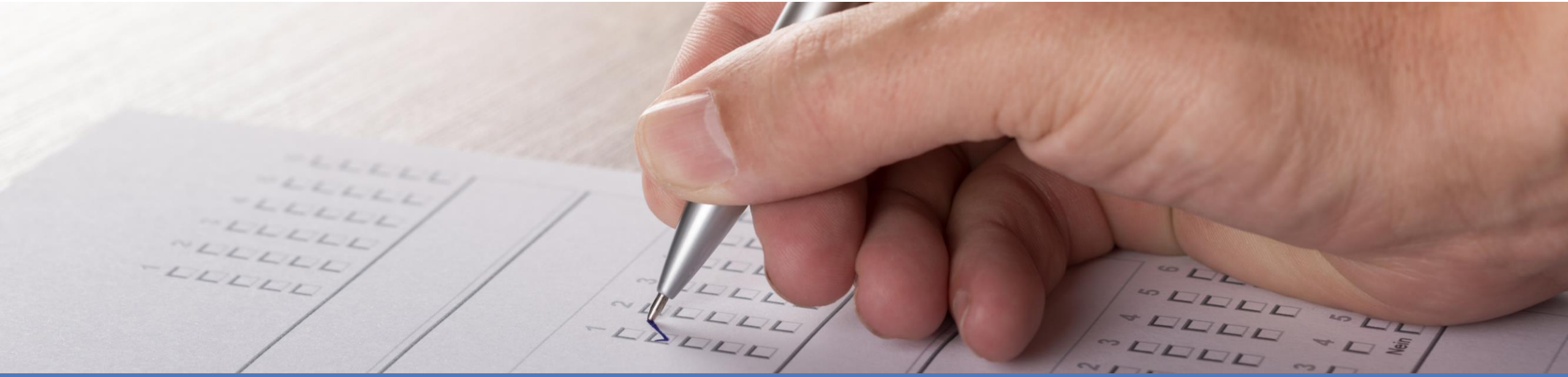
FRANKFURT UNIVERSITY OF APPLIED SCIENCES

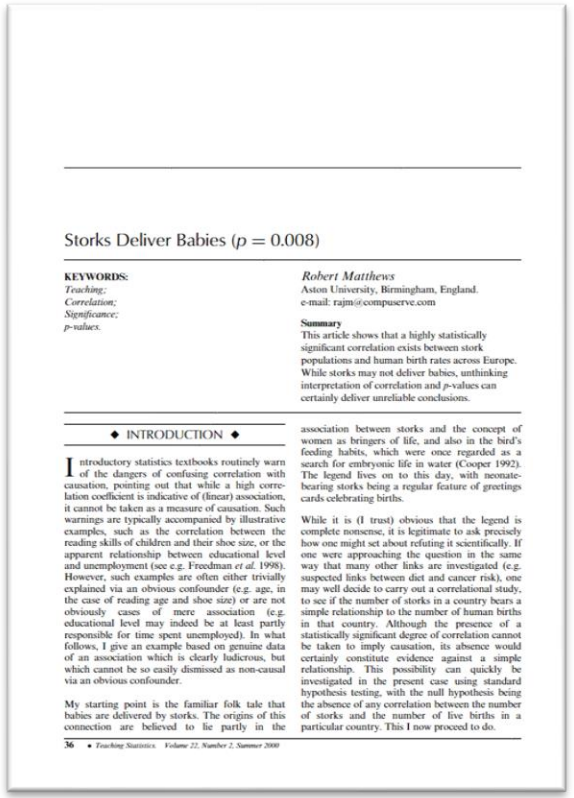


Empirical Research

Human-Computer Interaction Exercise

Image from https://pxhere.com/en/photo/855019

Storks Deliver Babies



Storks Deliver Babies ($p = 0.008$)

KEYWORDS:
Teaching;
Correlation;
Significance;
p-values.

Robert Matthews
Aston University, Birmingham, England.
e-mail: rajm@compuserve.com

Summary
This article shows that a highly statistically significant correlation exists between stork populations and human birth rates across Europe. While storks may not deliver babies, unthinking interpretation of correlation and *p*-values can certainly deliver unreliable conclusions.

◆ INTRODUCTION ◆

Introductory statistics textbooks routinely warn of the dangers of confusing correlation with causation, pointing out that while a high correlation coefficient is indicative of (linear) association, it cannot be taken as a measure of causation. Such warnings are typically accompanied by illustrative examples, such as the correlation between the reading skills of children and their shoe size, or the apparent relationship between educational level and unemployment (see e.g. Freedman *et al.* 1998). However, such examples are often either trivially explained via an obvious confounder (e.g. age, in the case of reading age and shoe size) or are not obviously cases of mere association (e.g. educational level may indeed be at least partly responsible for time spent unemployed). In what follows, I give an example based on genuine data of an association which is clearly ludicrous, but which cannot be so easily dismissed as non-causal via an obvious confounder.

My starting point is the familiar folk tale that babies are delivered by storks. The origins of this connection are believed to lie partly in the association between storks and the concept of women as bringers of life, and also in the bird's feeding habits, which were once regarded as a search for embryonic life in water (Cooper 1992). The legend lives on to this day, with neonate-bearing storks being a regular feature of greetings cards celebrating births.

While it is (I trust) obvious that the legend is complete nonsense, it is legitimate to ask precisely how one might set about refuting it scientifically. If one were approaching the question in the same way that many other links are investigated (e.g. suspected links between diet and cancer risk), one may well decide to carry out a correlational study, to see if the number of storks in a country bears a simple relationship to the number of human births in that country. Although the presence of a statistically significant degree of correlation cannot be taken to imply causation, its absence would certainly constitute evidence against a simple relationship. This possibility can quickly be investigated in the present case using standard hypothesis testing, with the null hypothesis being the absence of any correlation between the number of storks and the number of live births in a particular country. This I now proceed to do.

For the classroom

New evidence for the Theory of the Stork

Thomas Höfer[a], Hildegard Przyrembel[b] and Silvia Verleger[c]
[a]Federal Institute for Risk Assessment, Berlin, [b]Office of the National Breast Feeding Committee at BfR, Berlin, and [c]Independent Midwife, Berlin, Germany

Correspondence:
Dr Thomas Höfer, Federal Institute for Risk Assessment, Thielallee 88–92, D-14195 Berlin, Germany.
E-mail: thomas.hoefer@bfr.bund.de

Summary
Data from Berlin (Germany) show a significant correlation between the increase in the stork population around the city and the increase in deliveries outside city hospitals (out-of-hospital deliveries). However, there is no correlation between deliveries in hospital buildings (clinical deliveries) and the stork population. The decline in the number of pairs of storks in the German state of Lower Saxony between 1970 and 1985 correlated with the decrease of deliveries in that area. The nearly constant number of deliveries from 1985 to 1995 was associated with an unchanged stork population (no statistical significance). However, the relevance of the stork for the birth rate in that part of Germany remains unclear, because the number of out-of-hospital deliveries in this area is not well documented. A lack of statistical information on out-of-hospital deliveries in general is a severe handicap for further proof for the Theory of the Stork.

The intended value (disclaimer): This article is not intended to disprove the value of serious epidemiological investigations. It is an example of how studies based on popular belief and unsubstantiated theory, seconded by low quality references and supported by coincidental statistical association could lead to apparent scientific endorsement. Insofar it is a humorous case study for education in perinatal epidemiology.

Introduction

Background

Two different theories exist concerning the origin of children: the Theory of Sexual Reproduction (ThoSR) and the Theory of the Stork (ThoS).

For many people the stork has been and still is the 'bringer of new life' or the 'baby-carrier'. During pregnancy people say, 'the stork has been visiting'. Old German scientists therefore named the stork odebero (in dutch ooievaar), which means 'bringing luck'.

Nowadays, many people believe in the theory of reproduction, simply because they have been taught this theory in school, although it is a scientific theory, not a truth (Leisti T, personal communication via leisti@cc.helsinki.fi, 2001). A number of the world's leading scientists are still in favour of ThoS. Some recent scientific work (1997–2001) has shown new evidence for ThoS.

Status of ThoS

A search conducted on the internet revealed that more than 20 internet domains are communicating about aspects of ThoS (Table 1). This is evidently important for the future of the ThoS's relevance.

According to work of Prof Erkki Aalto[1] from the University of Helsinki the evidence supporting ThoS is based on six facts:

(i) storks exist
(ii) unexplained features of fetal development
(iii) a newborn is new-born
(iv) sexual intercourse without delivery
(v) positive correlation of birth rate to stork population
(vi) scientific studies.

Aalto's work focussed not only on aspects in the field of gynaecology and obstetrics:

Aspect no. (i): It is a scientifically established fact that the stork exists, which is confirmed by ornithologists.

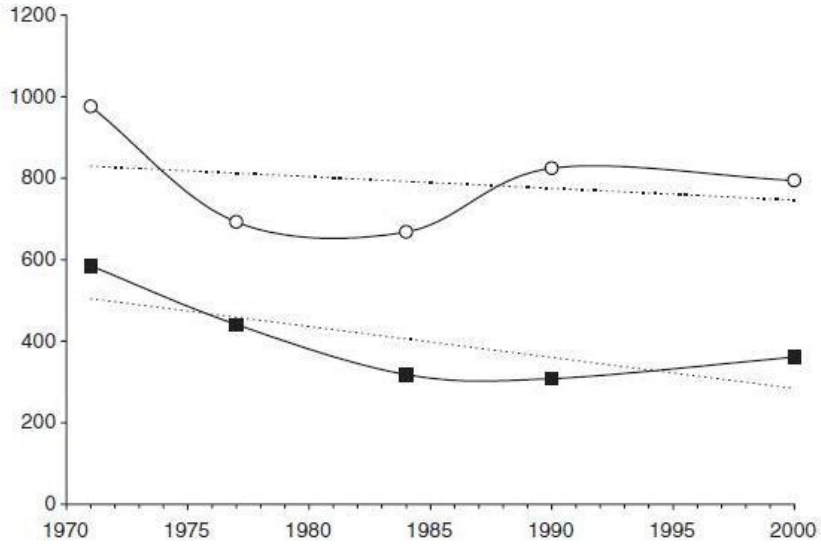


Figure 1. Storks and the birth rate in Lower Saxony, Germany (1971–2000). Open circles show yearly birthrates in hundreds in Lower Saxony. Full squares show numbers pairs of storks in Lower Saxony. Dotted lines represent linear regression trend ($y = mx + b$).

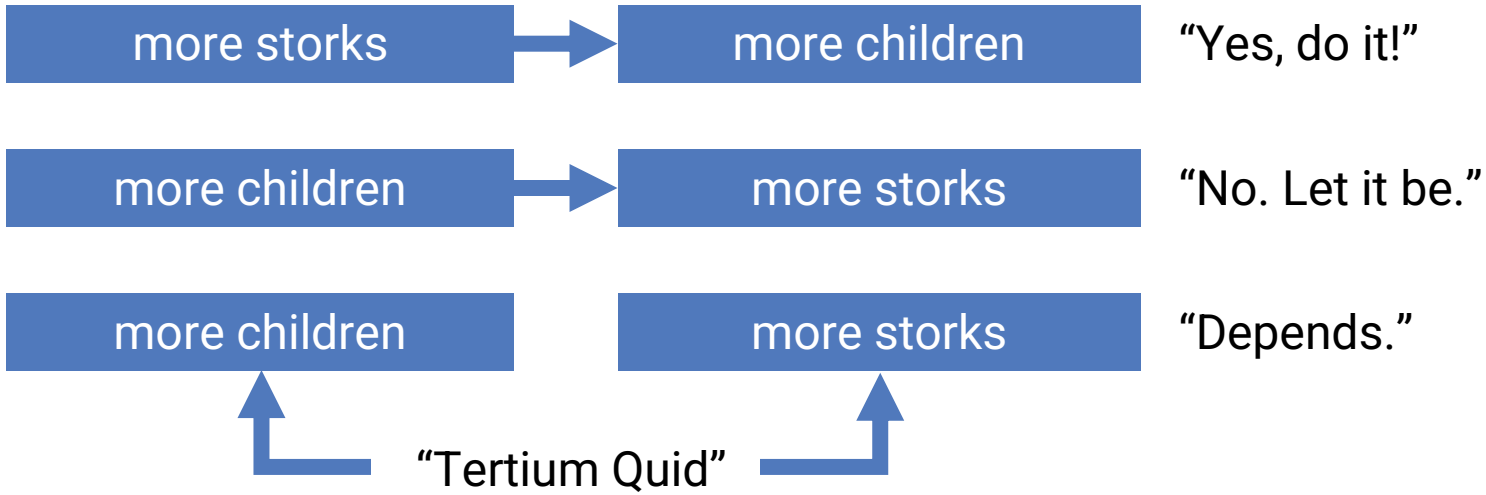
Matthews, R. (2000), Storks Deliver Babies (p= 0.008). Teaching Statistics, 22: 36-38. doi:10.1111/1467-9639.00013

Höfer, Thomas & Przyrembel, Hildegard & Höfer, Silvia. (2004). New evidence for the Theory of the Stork. Paediatric and perinatal epidemiology. 18. 88-92. 10.1111/j.1365-3016.2003.00534.x.

Image from https://pxhere.com/en/photo/855019

Causation and Correlation

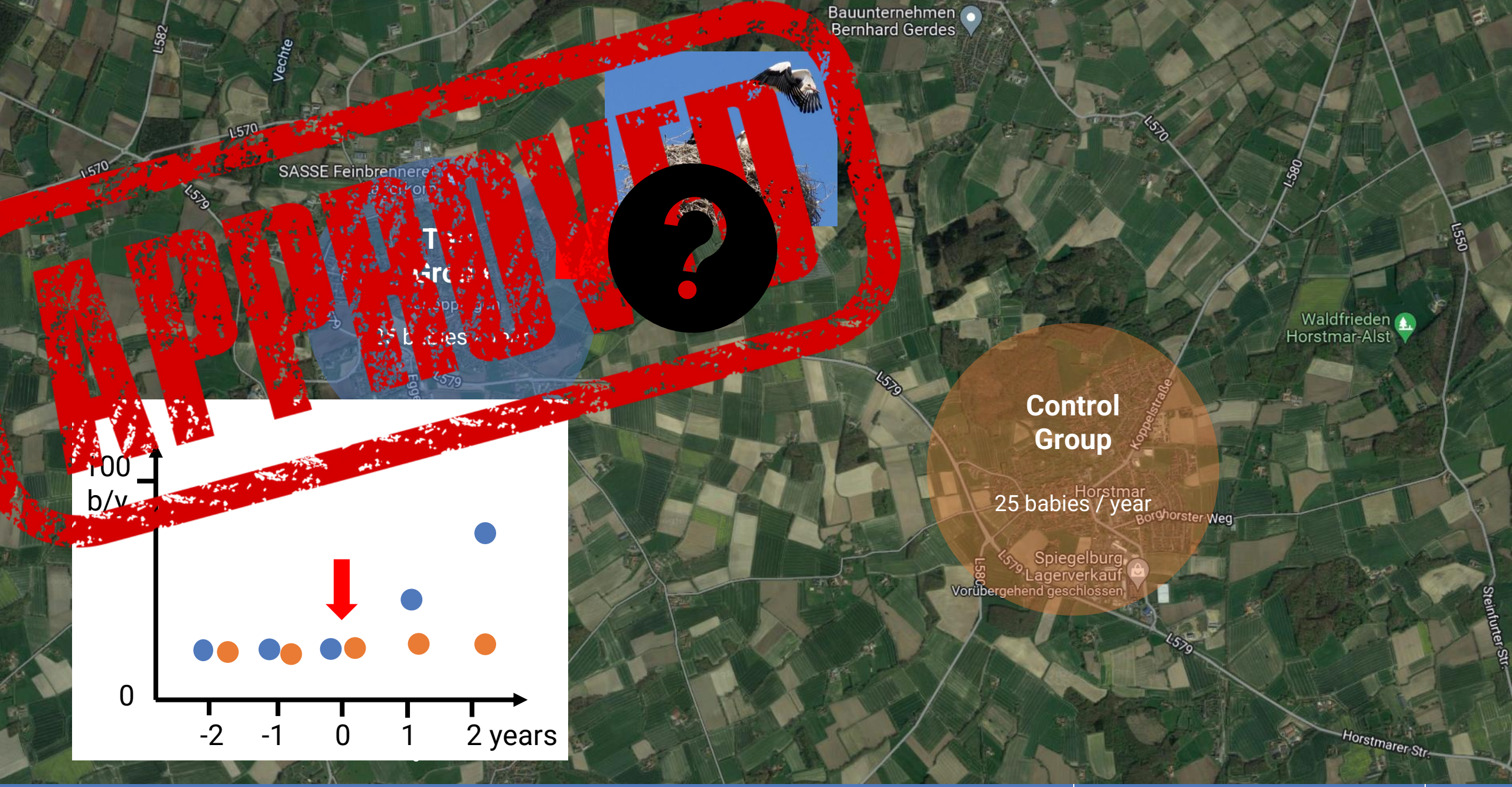
- A correlation cannot separate cause from effect, however,…
- … it's still a fact: birthrate and number of stork correlate. Any implications?
  - Question: "If I want more babies, should I move to an area with many storks?"

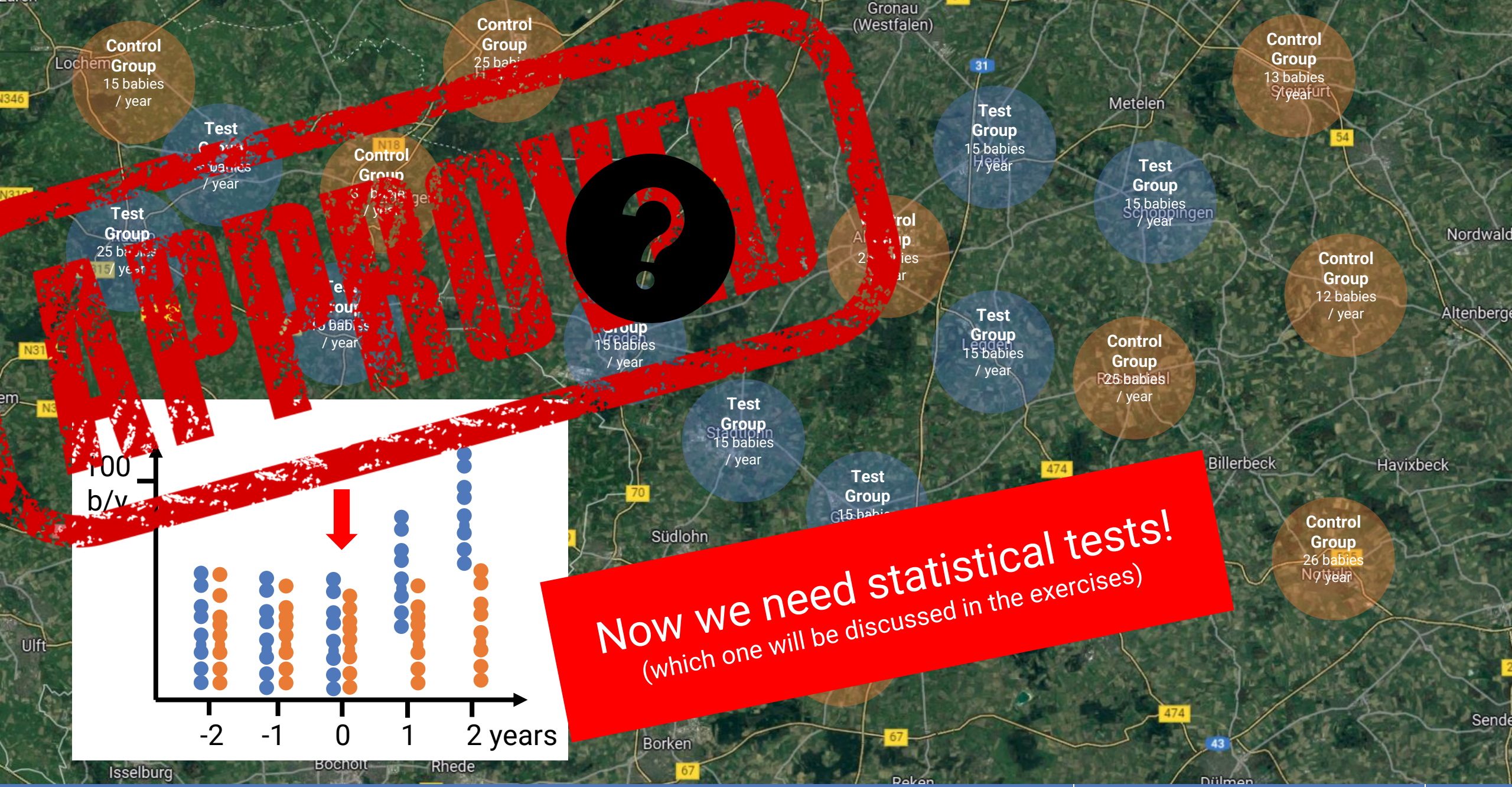


Everything starts with a research question…

“If you want to increase the likeability to have more babies, should you move to an area with many storks?”

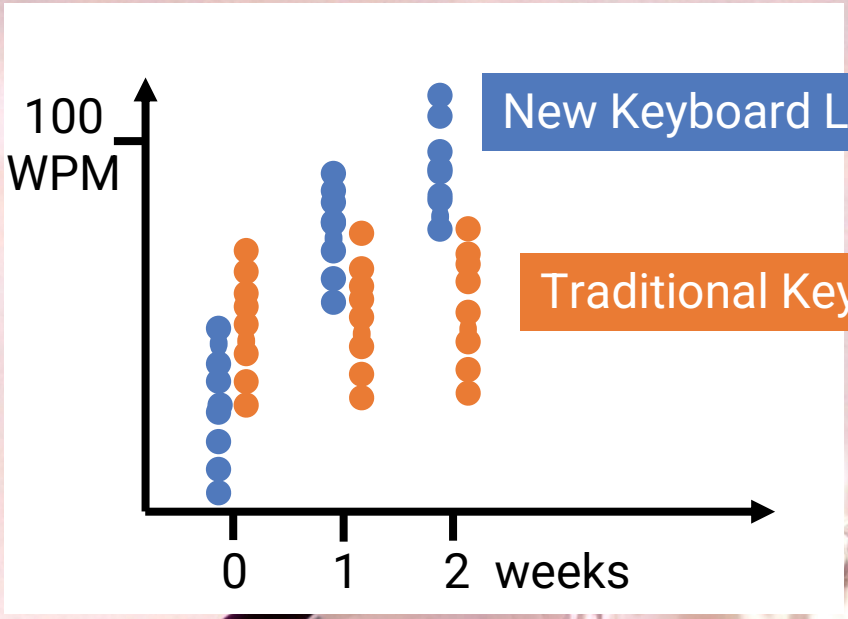
How would you answer the research question?


Bauunternehmen Bernhard Gerdes
L582
Vechte
L570
L579
L580
L550
SASSE Feinbrennerei
Waldfrieden Horstmar-Alst
Control Group
Horstmar
25 babies / year
Koppelstraße
Borghorster Weg
Spiegelburg Lagerverkauf
Vorübergehend geschlossen
Steinfurter Str.
Horstmarer Str.
APPROVED
?
100 b/y
0
-2
-1
0
1
2 years


Control Group 15 babies / year
Control Group 25 babies / year
Test Group 15 babies / year
Control Group 13 babies / year
Test Group 15 babies / year
Test Group 25 babies / year
Control Group 15 babies / year
Test Group 15 babies / year
Control Group 12 babies / year
Test Group 15 babies / year
Control Group 25 babies / year
Test Group 15 babies / year
Test Group 15 babies / year
Control Group 26 babies / year
Gronau (Westfalen)
Metelen
Schoppingen
Billerbeck
Havixbeck
Südlohn
Borken
Isselburg
Bocholt
Rhede
Ulft
Dülmen
Nordwalde
Altenberge
Sende
100 b/y
-2
-1
0
1
2 years
Approved
?
Now we need statistical tests!
(which one will be discussed in the exercises)

What the heck have storks to do with HCI?

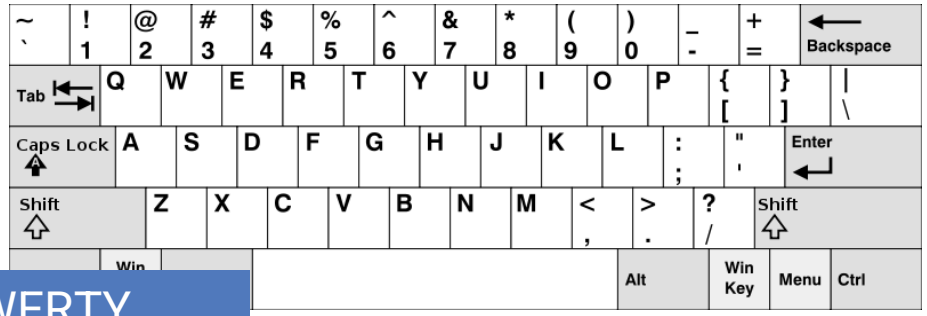
Discussion



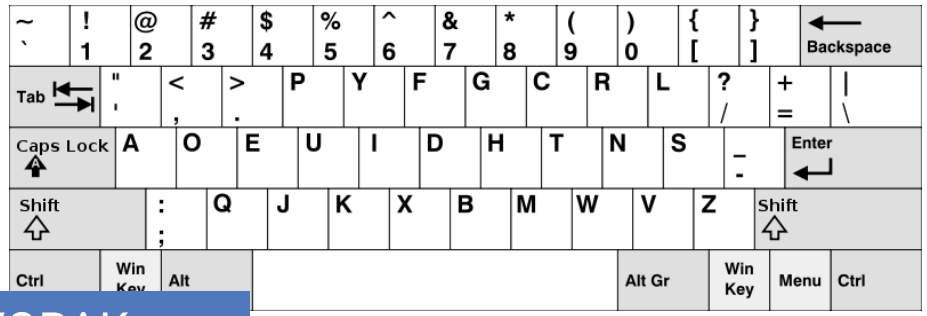
The results are just numbers.
Let‘s do it with keyboards…

Let‘s test our new keyboard layout!

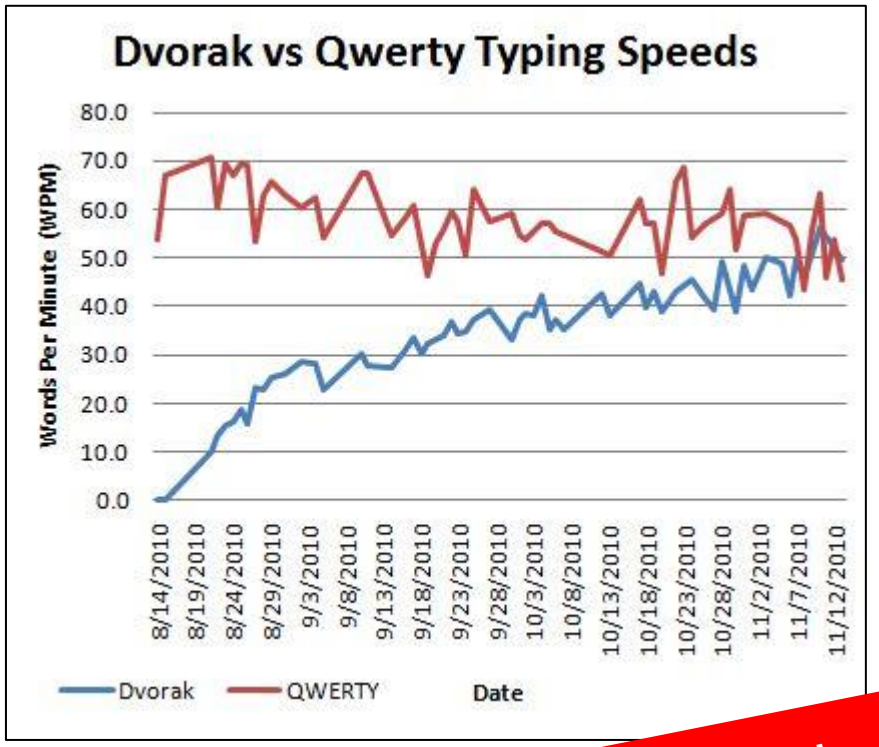
- What is the problem when we develop new keyboard layouts?



QWERTY



DVORAK

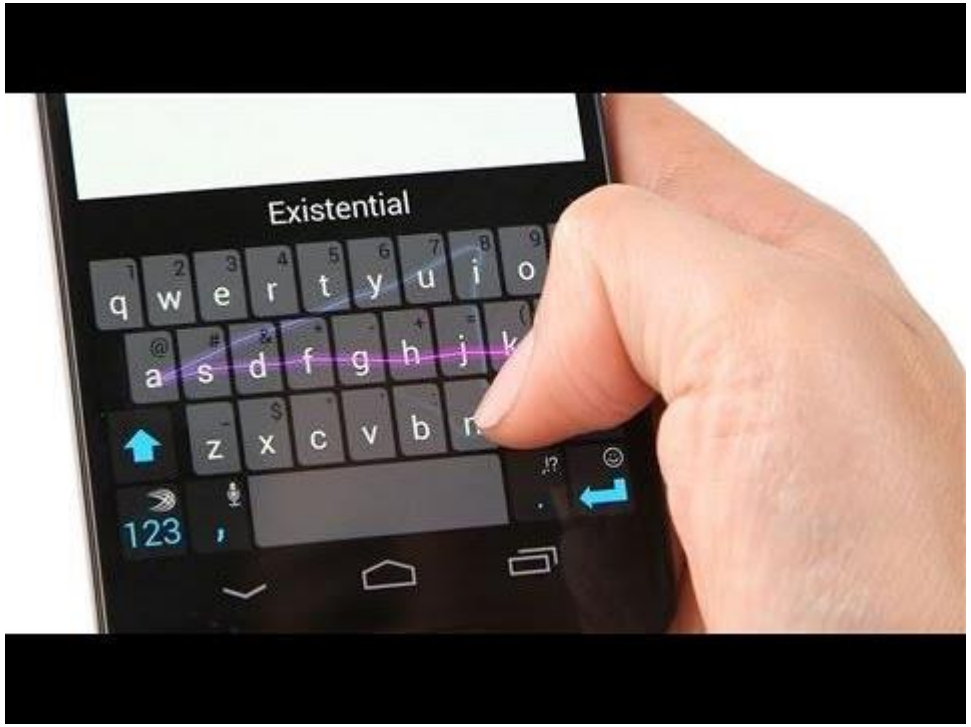


3 months training!

Images from https://www.boundlessat.com/Keyboards-Mice/Alternative-Keyboards

Inventing Swipe

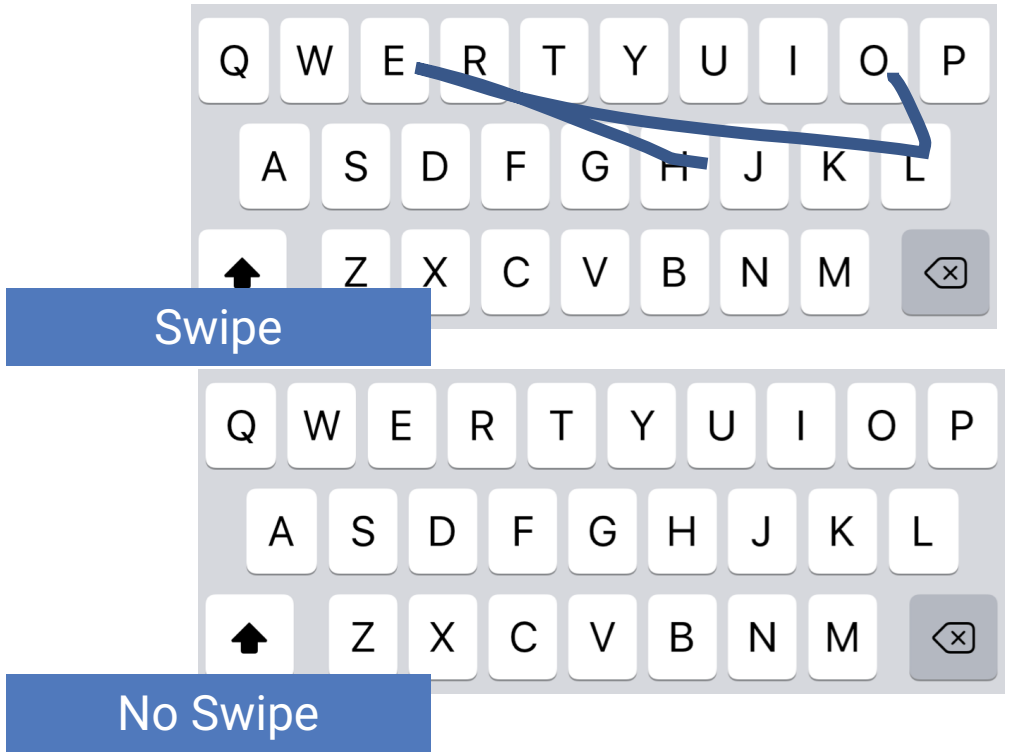
- Let‘s imagine we invented a gesture-based typing technique and call it SWIPE

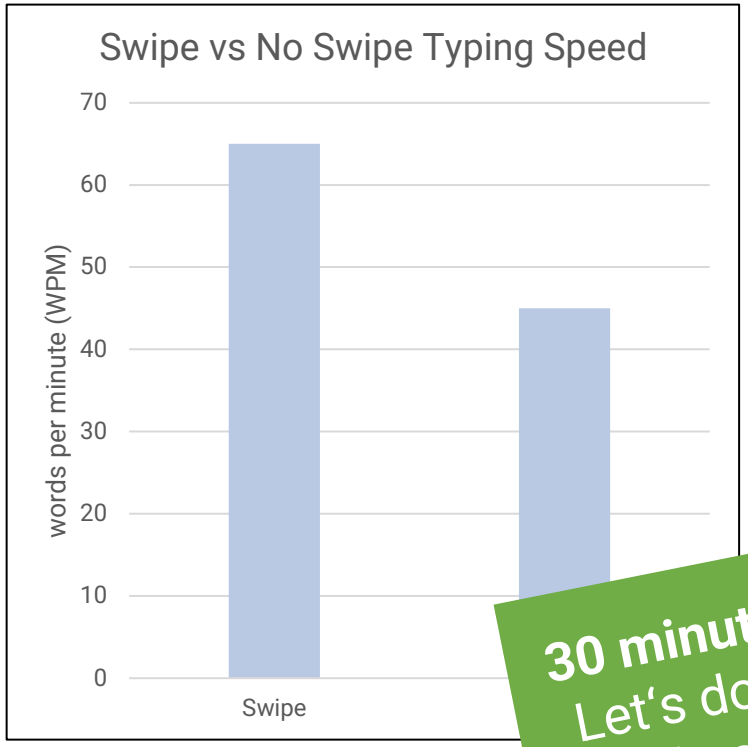


https://www.youtube.com/watch?app=desktop&v=1dr068gL03g
https://www.engadget.com/2019-06-03-apple-ios-13-quickpath-swipe-keyboard.html

Inventing Swipe

- First tests with us seem promising. But: Does it work with real participants?





30 minutes training!
Let‘s do a test with real participants…

Let‘s design an empirical study…

- The research question: **“Does Swipe improve the typing performance on touch displays (compared to a regular smartphone keyboard)?”**
- Do we need a **qualitative or quantitative research method?**
  - Quantitative because of three reasons:
    - **We can measure performance**
    - **We can formulate a hypotheses**:
      - H1a: Swipe improves the typing performance.
    - … and **we can falsify it** (the null hypothesis):
      - H0a: Swipe does not improve the typing performance

→ Let‘s use **a quantitative method** and measure performance

- We‘ll later learn, that qualitative methods can be great for our research, too.

Swipe vs No Swipe – whats better?

- We are assuming that “better” = “more performance”, which is not necessarily true.

No Swipe

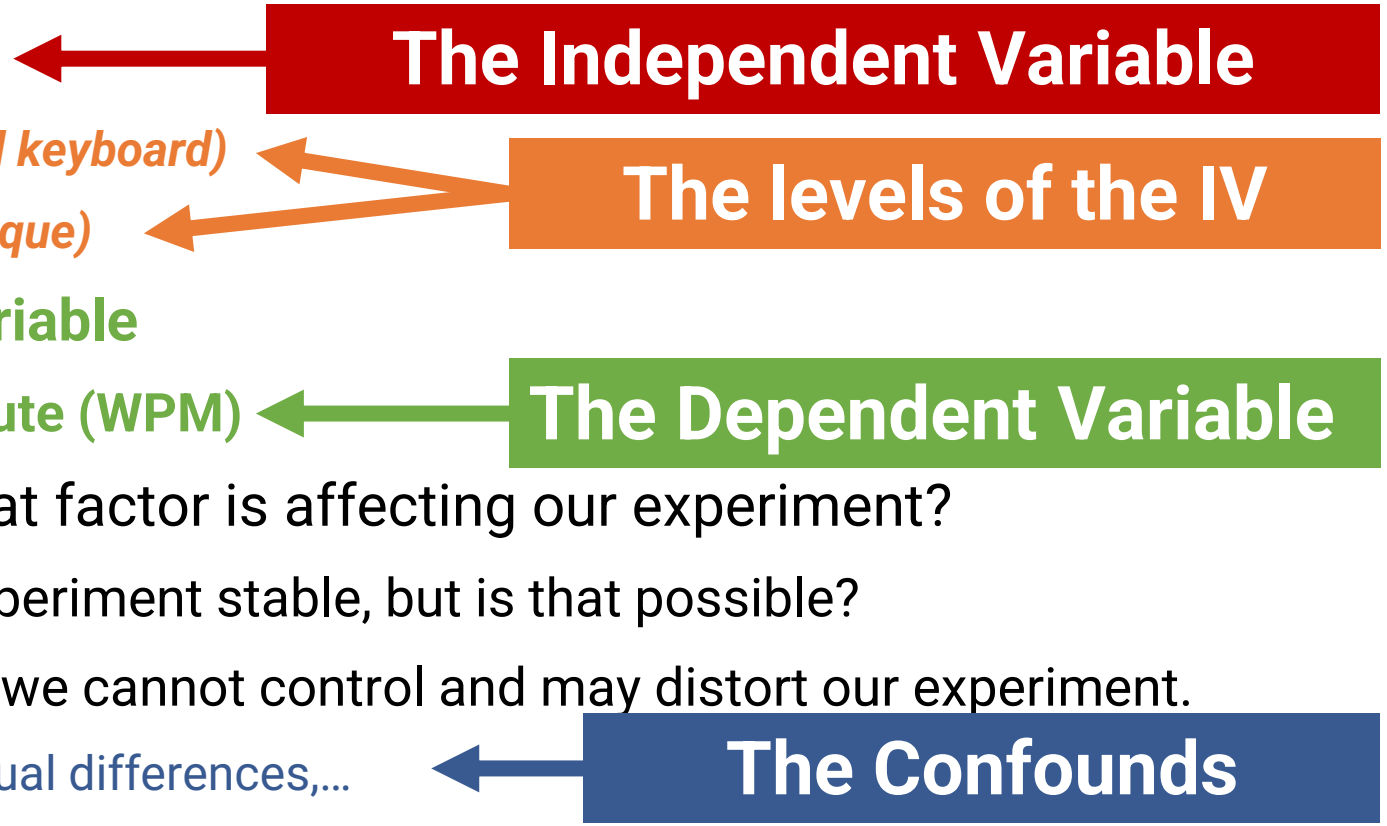
Swipe

the standard or „control“

the factor we change

Quantitative Studies

- **The factor you change is the Independent Variable**
  - **STORKS = SWIPE**
    - *Control Group = No Swipe (normal keyboard)*
    - *Test Group = Swipe (our technique)*
- **Your measure is the Dependent Variable**
  - **Births per year = Words per minute (WPM)**
- But how can we ensure that only that factor is affecting our experiment?
  - By keeping all other factors in our experiment stable, but is that possible?
  - No. We always have **Confounds** that we cannot control and may distort our experiment.
    - Environment, weather, training, individual differences,…



Independent Variables (IV) and their Levels

- An IV (here: SWIPE) must have **at least 2 levels** that allow us **to control that factor**
  - We define these levels. This is why we called it the „independent“ variable
  - Levels are (nominal) categories: e.g., *No Swipe (normal)*, *Swipe (our awesome technique)*
- **We must assign** these **levels** to our participants
  - When we do that, levels become **conditions** ← **The Condition**
- Either we assign participants…
  - **randomly to only one level of the IV** (e.g., in A/B testing or medical treatments)
    → This is called a **between-subject** (or between-groups) **variable**
  - **to each level of the IV** (e.g., in surveys or many lab studies)
    → This is called a **within-subject** (or within-groups) **variable**
    - In some cases, **within-subject IVs have indefinite levels and are numeric**
      - „number x is a function of y“

The Dependent Variable

- **The thing that you measure** (typically a metric or stuff you can count) should be part of your research question and hypothesis:
  - The reserach question: **„Does swipe improve the typing performance on touch displays?"**
    - H1a: „Swipe improves the typing performance."
    - H0a: „Swipe does not improve the typing performance."
    - H1b: „Swipe improves the usability."
    - H0b: „Swipe does not improve the usability."

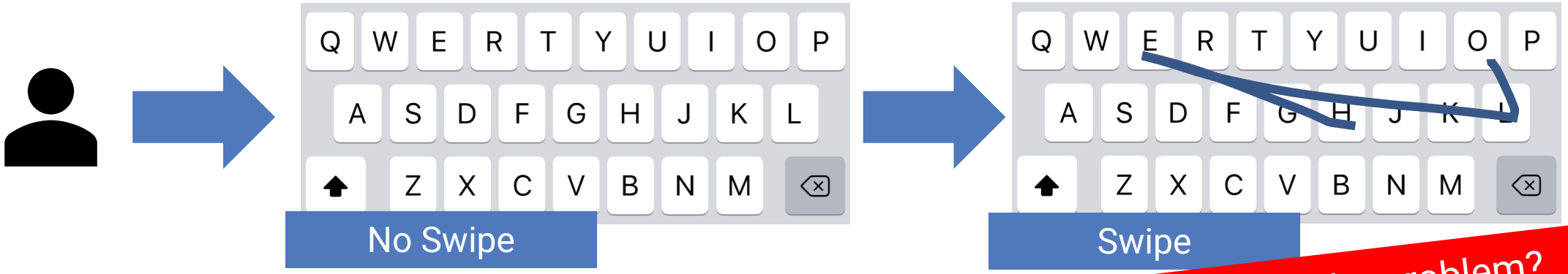
„improves" = directed hypothesis

„changes" = undirected hypothesis

- The **dependent variables** are…
  - **typing performance** that can be measured 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, etc.), …
  - **usability** that can be measured by:
    - System Usability Scale (SUS), NASA taskload index (TLX), …

The Confounds / The Covariates

- In theory, **we need to keep all confounding factors** of the experiment (environment, weather, training, intelligence, mood, …) **stable**
- This is not possible, but we can…
  - **record and describe all confounds (also called covariates)** in our experiment (gender, age, previous training, the location …) and
  - **draw a random/representative sample (our participants)** that must test both keyboards…



Whats the problem?

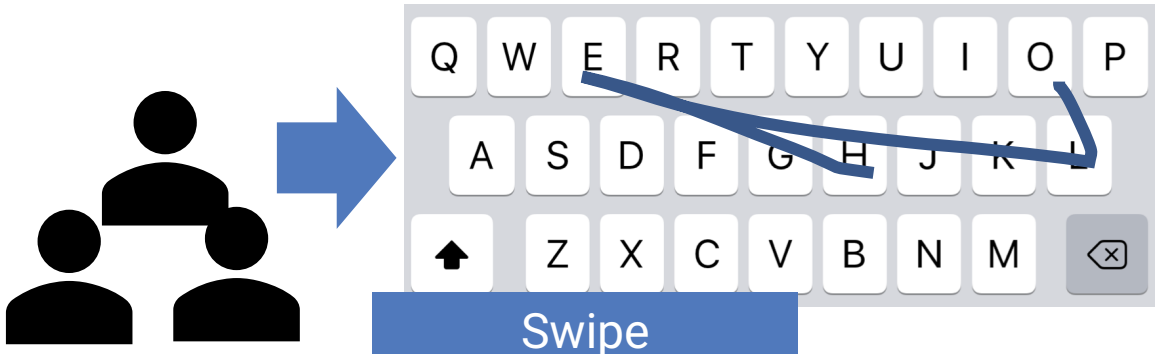
Carry-over effects!

Carry-Over Effects

- **Carry-over** (or sequence) **effects** refer to **the influence of a subsequent condition**
- These effects can be
  - physiological
  - psychological
  - behavioral
  - or may manifest as learning, fatigue, or sensitization
- **Carry-over effects are confounds** and make it unclear whether the observed effects are due to the experimental manipulation or the effects of previous conditions
  - **If you have them, your experiment is broken!**
  - We have them when people are using our smartphone and learn to type two times.
- How can we prevent carry-over effects?

Designing a Study

- We have two conditions (the two levels of our IV). To prevent carry-over effects we split them into two groups.

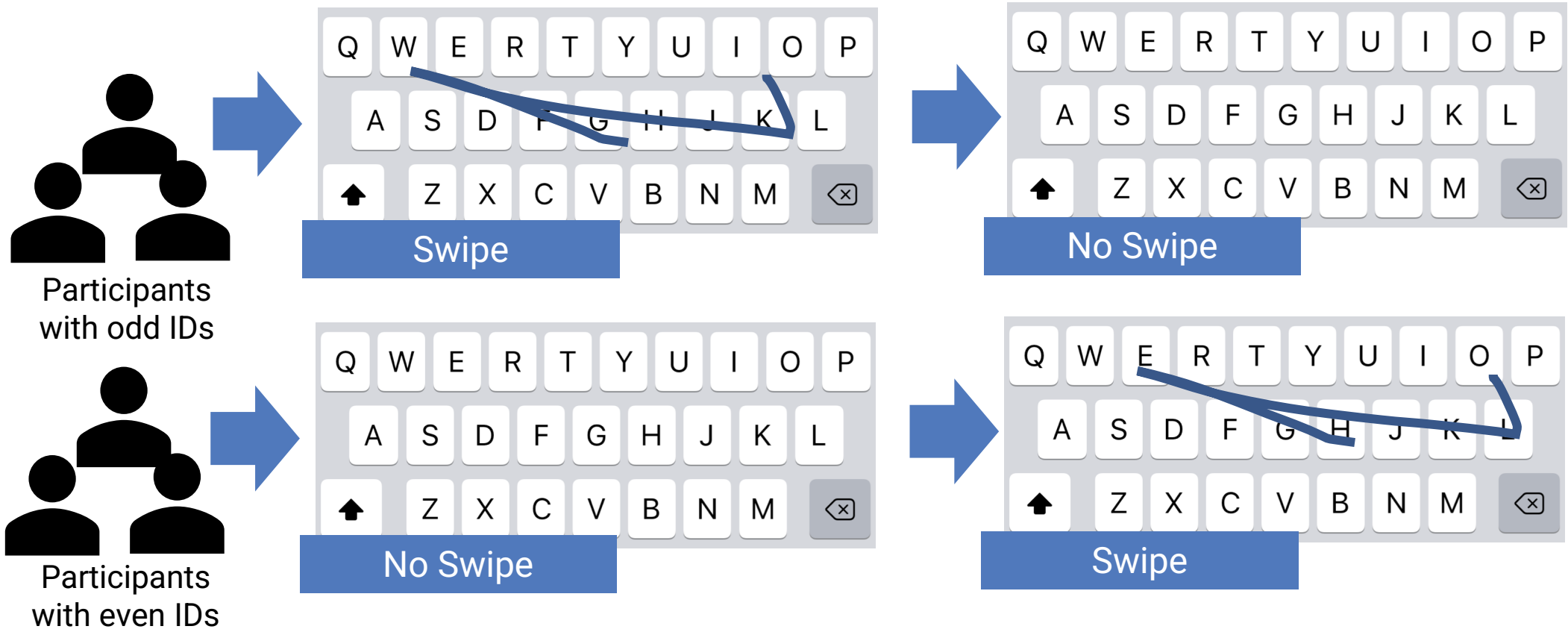


Such designs are very common in clinical studies.
But what is the problem with such designs?

We are poor.
We don‘t have hundreds of participants.

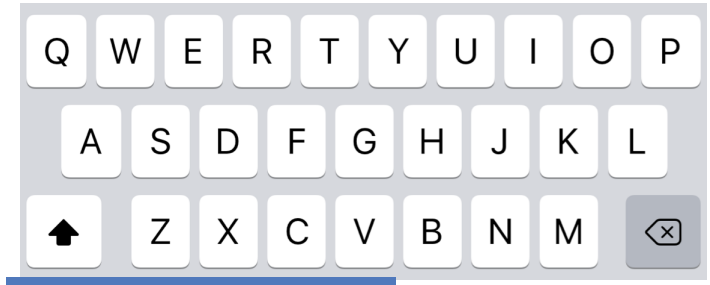
Counterbalancing

- Typically, we want that our participants do both conditions.



Counterbalancing 2 Conditions

- With two conditions we just keep continue switching the levels of the IV…



No Swipe

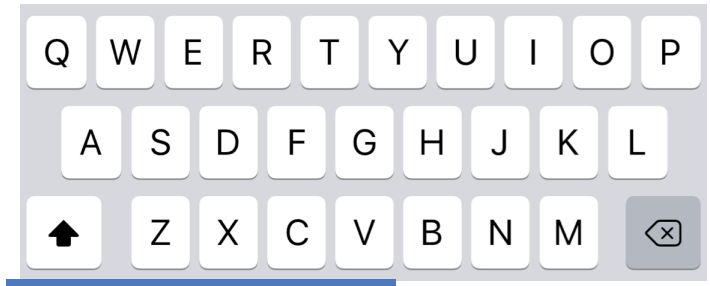
Swipe

	Condition 1	Condition 2
Participant #1	**Swipe**	**NoSwipe**
Participant #2	**NoSwipe**	**Swipe**
Participant #3	**Swipe**	**NoSwipe**
Participant #4	**NoSwipe**	**Swipe**
…	…	…

Continue switching…

Let add a third level of the IV: VibroSwipe

- Let‘s integrate some tactile feedback! How would you design the experiment now?



	Condition 1	Condition 2	Condition 3
Participant #1	**NoSwipe**	**Swipe**	**VibroSwipe**
Participant #2	**Swipe**	**NoSwipe**	**VibroSwipe**
Participant #3	**NoSwipe**	**VibroSwipe**	**Swipe**
Participant #4	**VibroSwipe**	**NoSwipe**	**Swipe**
Participant #5	**Swipe**	**VibroSwipe**	**NoSwipe**
Participant #6	**VibroSwipe**	**Swipe**	**NoSwipe**

Calculate the number of combinations:
N! → X:
2 → 2
3 → 6
4 → 24
5 → 120
6 → 720

… we should defintely avoid permutations with more than 4 conditions!
Or…?

D

Balanced Latin Square

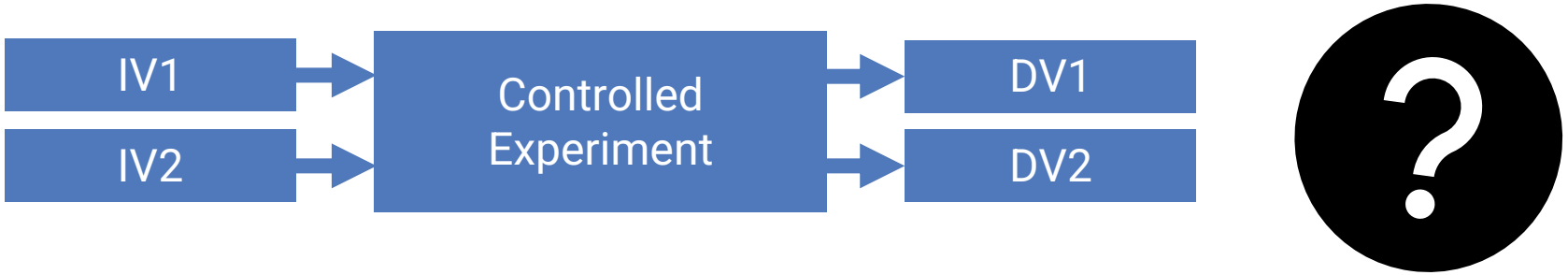
- A Latin square is **an N × N array** filled with n different symbols. They are useful to reduce order-effects when designing experiments with many conditions.
- **A condition will precede another exactly once** (or twice when condition no. is odd)
- Balanced Latin Square generator: https://hci-studies.org/balanced-latin-square/

	Condition 1	Condition 2	Condition 3	Condition 4	Condition 5	Condition 6
Participant #1	**A**	**B**	**F**	**C**	**E**	**D**
Participant #2	**B**	**C**	**A**	**D**	**F**	**E**
Participant #3	**C**	**D**	**B**	**E**	**A**	**F**
Participant #4	**D**	**E**	**C**	**F**	**B**	**A**
Participant #5	**E**	**F**	**D**	**A**	**C**	**B**
Participant #6	**F**	**A**	**E**	**B**	**D**	**C**

More Indepedent Variables?

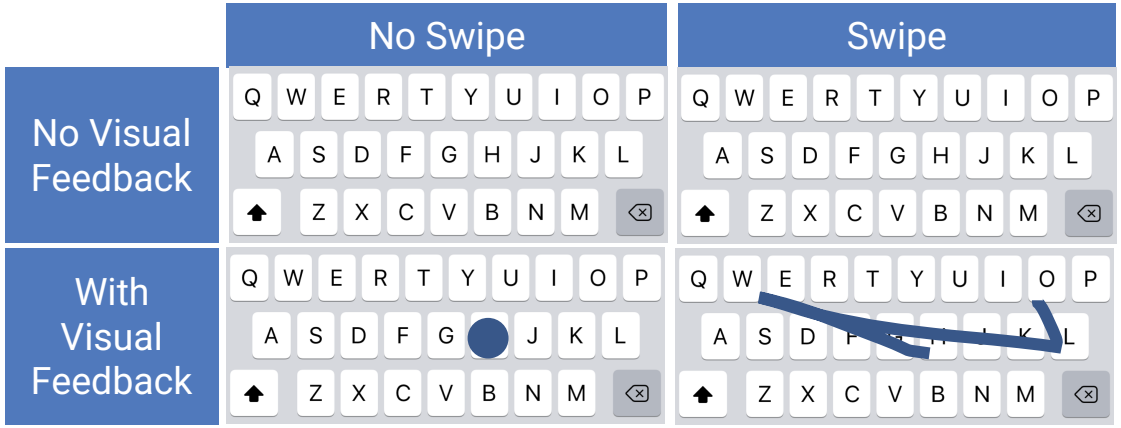
- **Controlled experiments are the only reliable mean to isolate cause** (of the independent variable) **from the effect** (on the dependent variable)

- What if there are potential two effects or if they potentially depend on each other?
  - **Can we consider multiple IVs and observe effects on multipe DVs?**



Full-Factorial Designs (2 Factors)

- Only in controlled experiments, **we can observe** the effects of **multiple factors** at the same time
- For example: **SWIPE × VISUAL FEEDBACK**



- What does the Balanced Latin Square looks like?

Balanced Latin Square of a 2×2 Full-Factorial Design

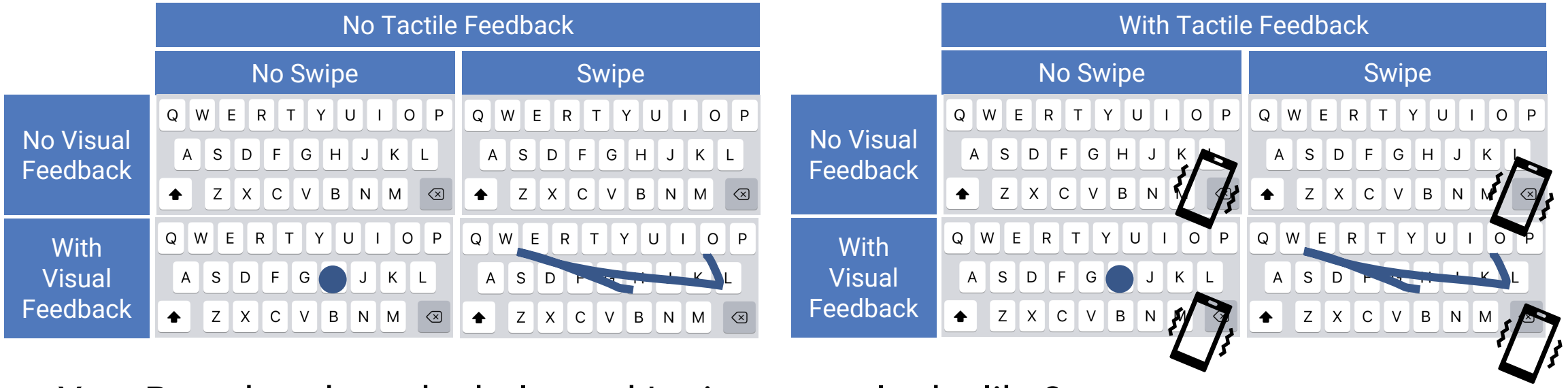
- **The combination of each level becomes now a condition** of the experiment…

	Condition 1	Condition 2	Condition 3	Condition 4
Participant #1	**NoSwipe-NoVisualFeedback**	**NoSwipe-VisualFeedback**	**Swipe-VisualFeedback**	**Swipe-NoVisualFeedback**
Participant #2	**NoSwipe-VisualFeedback**	**Swipe-NoVisualFeedback**	**NoSwipe-NoVisualFeedback**	**Swipe-VisualFeedback**
Participant #3	**Swipe-NoVisualFeedback**	**Swipe-VisualFeedback**	**NoSwipe-VisualFeedback**	**NoSwipe-NoVisualFeedback**
Participant #4	**Swipe-VisualFeedback**	**NoSwipe-NoVisualFeedback**	**Swipe-NoVisualFeedback**	**NoSwipe-VisualFeedback**
…	…	…	…	

→ Repeat the Latin Square!

Full-Factorial Designs (3 Factors)

- Okay, now things are getting complicated: **Can we observe the impact of 3 factors at once?**
- For example: **SWIPE × VISUAL FEEDBACK × TACTILE FEEDBACK**



- Yes. But what does the balanced Latin square looks like?

Balanced Latin Square of a 2×2×2 Full-Factorial Design

	Condition 1	Condition 2	Condition 3	Condition 4	Condition 5	Condition 6	Condition 7	Condition 8
Participant #1	None-None-None	None-None-Visual	None-Tactile-None	None-Tactile-Visual	Swipe-None-None	Swipe-None-Visual	Swipe-Tactile-None	Swipe-Tactile-Visual
Participant #2	None-None-Visual	None-Tactile-Visual	None-None-None	Swipe-None-Visual	None-Tactile-None	Swipe-Tactile-Visual	Swipe-None-None	Swipe-Tactile-None
Participant #3	None-Tactile-Visual	Swipe-None-Visual	None-None-Visual	Swipe-Tactile-Visual	None-None-None	Swipe-Tactile-None	None-Tactile-None	Swipe-None-None
Participant #4	Swipe-None-Visual	Swipe-Tactile-Visual	None-Tactile-Visual	Swipe-Tactile-None	None-None-Visual	Swipe-None-None	None-None-None	None-Tactile-None
Participant #5	Swipe-Tactile-Visual	Swipe-Tactile-None	Swipe-None-Visual	Swipe-None-None	None-Tactile-Visual	None-Tactile-None	None-None-Visual	None-None-None
Participant #6	Swipe-Tactile-None	Swipe-None-None	Swipe-Tactile-Visual	None-Tactile-None	Swipe-None-Visual	None-None-None	None-Tactile-Visual	None-None-Visual
Participant #7	Swipe-None-None	None-Tactile-None	Swipe-Tactile-None	None-None-None	Swipe-Tactile-Visual	None-None-Visual	Swipe-None-Visual	None-Tactile-Visual
Participant #8	None-Tactile-None	None-None-None	Swipe-None-None	None-None-Visual	Swipe-Tactile-None	None-Tactile-Visual	Swipe-Tactile-Visual	Swipe-None-Visual

Between vs Within-Subject Study Designs

- **Between-Subject Design**
  - Participants are assigned to **one level of the IV**
  - Advantages: very simple, no sequence effects, required when it is impossible for an individual to participate in all conditions (e.g., gender)
  - Disadvantages: expense (time, effort, and number of participants), very insensitive to experimental manipulations
- **Within-Subject Design**
  - Participants are assigned to **each level of the IV**
  - Advantages: economy, sensitiveness, cancelling out individual differences
  - Disadvantages: carry-over effects from previous conditions, conditions need to be counter-balanced or randomized

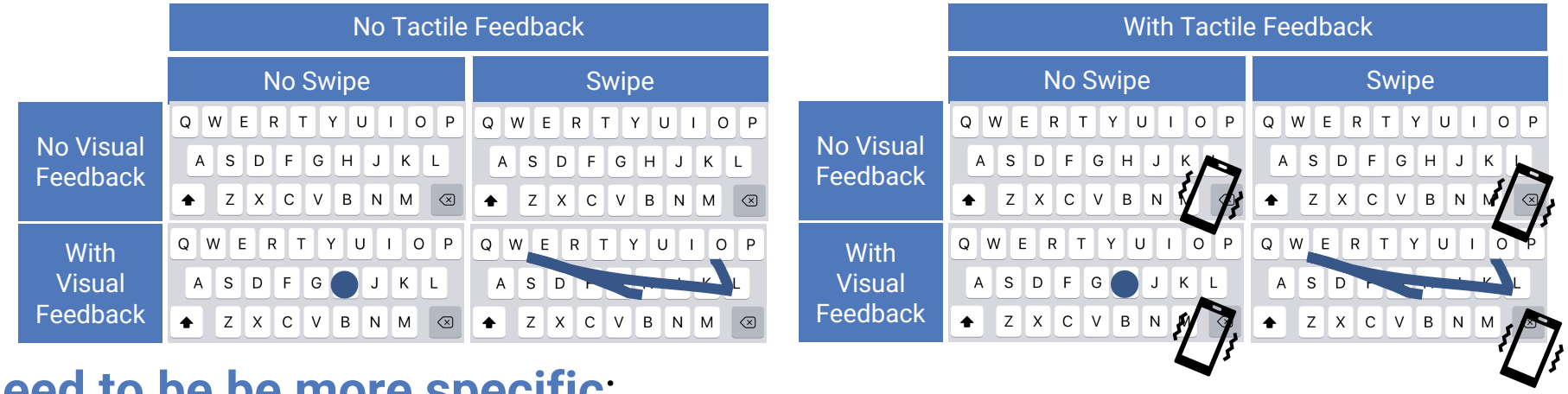
Mixed Designs

- **Your experiment can also include both types** of independent variables
  - between-subjects variable(s)
  - within-subjects variable(s)
  - e.g., in gender studies: do men or women type faster on my new swipe keyboard?
- Important: Participants must be **randomly assigned** to each level of the **between-subjects variable**(s)
  - Gender is random by birth
- **All participants are exposed** to each level of the **within-subjects variable**(s)
  - All men and women test my swipe keyboards. But, we must consider the order of conditions!
    - permutate the order (2 → 2, 3 → 6, 4 → 24, 5 → 120, 6 → 720, …)
    - counter-balancing using e.g., a balanced Latin-Square
    - fully random

Great. Can we now evaluate what‘s better?

What do you understand regarding „better“?

- Let‘s do the 8 conditions 4 times = 32 participants, measuring the WPM
- What‘s still the problem here?



- **We need to be be more specific**:
  - „Swipe improves the **typing performance**.“
    - **Are words per minute are identical to performance?**
  - What about „usability“?

Operationalization: From Concepts to Measures

- **What is „performance“? What is „usability“? Measures can be ambigous!**
  - Let‘s take „usability“
- Concepts such as usability can contain or refer to **other concepts**
  - e.g., „workload“, „efficiency“, „satisfaction“
- Concepts such as workload can be composed of **different variables**
  - e.g., „cognitive workload“ and „physical workload“
- Variables need **definitions** or **working definitions** (when we are not sure)
  - e.g., „Cognitive workload refers to the amount of mental effort that is exerted or required while reasoning and thinking, to process information, make decisions, and solve problems.“ [3]
- Finally, we need to **operationalize** the definition of **a concept** using scientific methods and consensus
  - e.g., NASA Taskload Index (TLX) [1], Cognitive Load Questionnaire [2]
    - And even cognitive load has more subconcepts: intrinsic, extraneaous, germane [3] (but we are happy with cognitive load)

[1] Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139-183). North-Holland.
[2] Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2016). Cognitive load measurement as a means to advance cognitive load theory. In *Cognitive Load Theory* (pp. 63-71). Routledge.
[3] Skulmowski, Alexander; Rey, Günter Daniel (2020). "Subjective cognitive load surveys lead to divergent results for interactive learning media". Human Behavior and Emerging Technologies. 2 (2): 149–157. doi:10.1002/hbe2.184

Standardized Tools

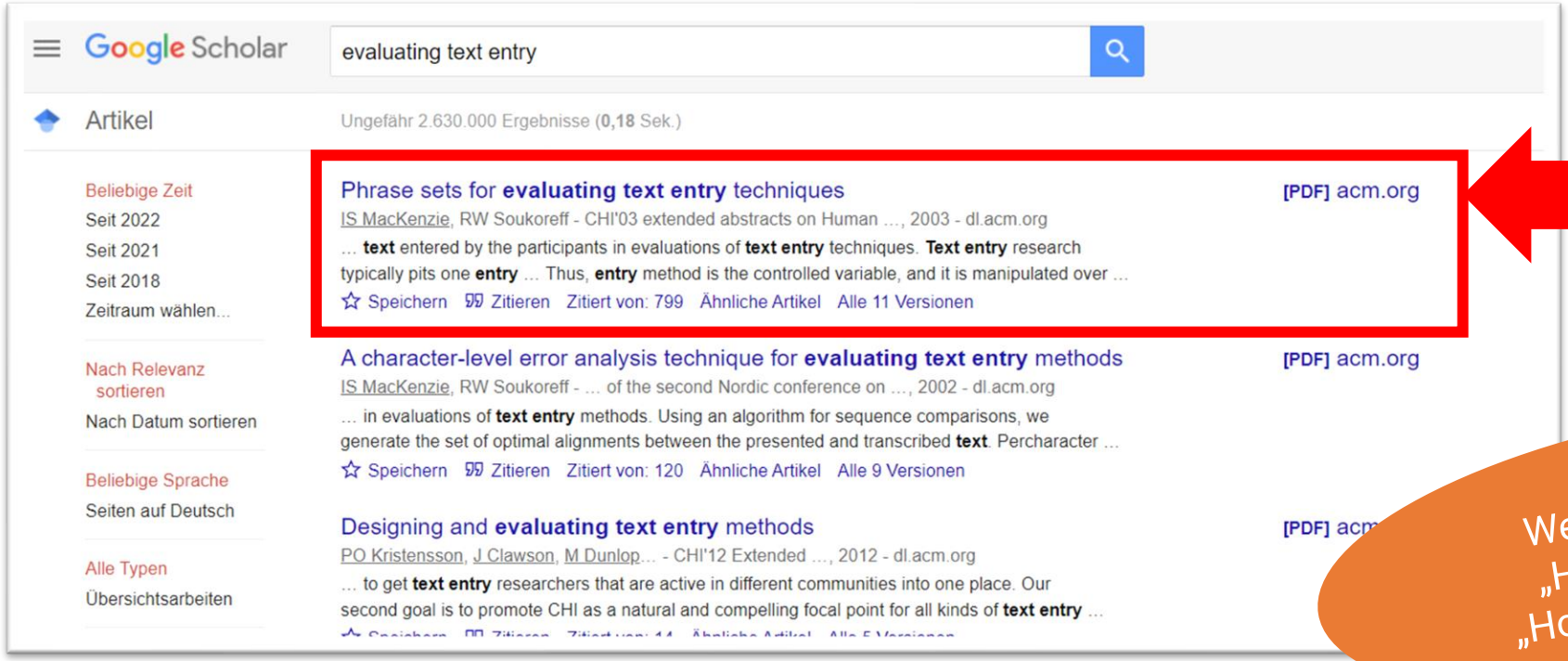
- The process of **defining the measurement for a concept** that is not directly measurable (even objective measures such as “performance”)
- Depends on theoretical definitions
- Makes a fuzzy concept (e.g., “emotions”, “user experience”, “likeability”, “memorability”, “usability”, “health”, …) **distinguishable**, **measurable**, and **understandable**
- Even helps infer the **existence of a concept** (e.g., realism) using a tool
  - If a standardized tool exist: take it
  - If multiple standardized tools exist: take the most appropriate or validated one
  - If no standardized tool exist: use own questions ← we’ll talk about that, later

Can we now evaluate what‘s ~~better?~~ faster/more usable?

- **Important:**
  - **Each concept must be part of your research question** (performance/usability)
  - **Each dependent variable must be part of your hypothesis** (WPM/CPM/error rate)
    - They can operationalized as **objective measure** such as **performance**: WPM, CPM, error rate
    - They can operationalized as **subjective measure** such as **usability**: SUS, TLX Score
- **Hypotheses should be precise!**
  - **H1a**: „Typing using Swipe increases the WPM on a touch display of smartphone.“
    H0a: „Typing using Swipe does not increase the WPM on a touch display of smartphone.“
  - **H1b**: „Typing using Swipe increases the SUS score on a touch display of smartphone.“
    H0b: „Typing using Swipe does not increase the SUS score on a touch display of smartphone.“
- What else do we need for our experiment?

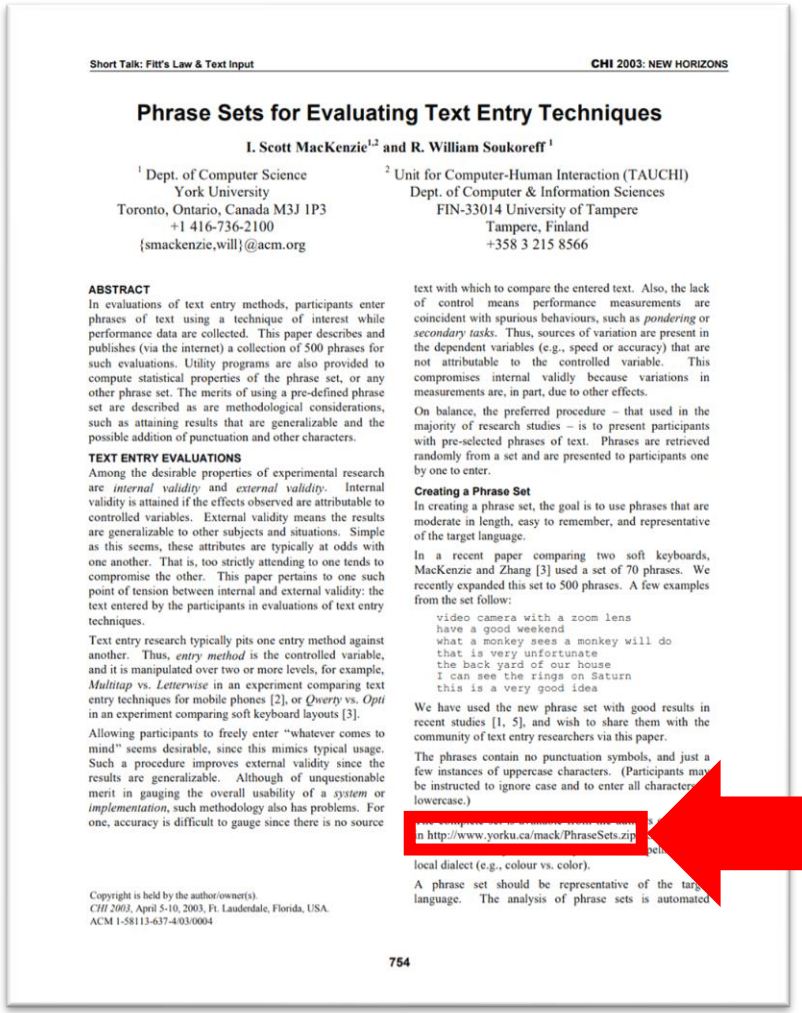
Tasks in HCI

- In HCI/science there are **standardized tasks** for many things (not everything)
  - This is the point where you need help (or Google): *„evaluating text entry“*



We discuss this tool in
„How to read a paper“
„How to review literature“

Science is strange, but nice, let‘s download it…



Short Talk: Fitt's Law & Text Input — CHI 2003: NEW HORIZONS

Phrase Sets for Evaluating Text Entry Techniques

I. Scott MacKenzie[1,2] and R. William Soukoreff [1]

[1] Dept. of Computer Science
York University
Toronto, Ontario, Canada M3J 1P3
+1 416-736-2100
{smackenzie,will}@acm.org

[2] Unit for Computer-Human Interaction (TAUCHI)
Dept. of Computer & Information Sciences
FIN-33014 University of Tampere
Tampere, Finland
+358 3 215 8566

ABSTRACT
In evaluations of text entry methods, participants enter phrases of text using a technique of interest while performance data are collected. This paper describes and publishes (via the internet) a collection of 500 phrases for such evaluations. Utility programs are also provided to compute statistical properties of the phrase set, or any other phrase set. The merits of using a pre-defined phrase set are described as are methodological considerations, such as attaining results that are generalizable and the possible addition of punctuation and other characters.

TEXT ENTRY EVALUATIONS
Among the desirable properties of experimental research are *internal validity* and *external validity*. Internal validity is attained if the effects observed are attributable to controlled variables. External validity means the results are generalizable to other subjects and situations. Simple as this seems, these attributes are typically at odds with one another. That is, too strictly attending to one tends to compromise the other. This paper pertains to one such point of tension between internal and external validity: the text entered by the participants in evaluations of text entry techniques.

Text entry research typically pits one entry method against another. Thus, *entry method* is the controlled variable, and it is manipulated over two or more levels, for example, *Multitap* vs. *Letterwise* in an experiment comparing text entry techniques for mobile phones [2], or *Qwerty* vs. *Opti* in an experiment comparing soft keyboard layouts [3].

Allowing participants to freely enter "whatever comes to mind" seems desirable, since this mimics typical usage. Such a procedure improves external validity since the results are generalizable. Although of unquestionable merit in gauging the overall usability of a *system* or *implementation*, such methodology also has problems. For one, accuracy is difficult to gauge since there is no source text with which to compare the entered text. Also, the lack of control means performance measurements are coincident with spurious behaviours, such as *pondering* or *secondary tasks*. Thus, sources of variation are present in the dependent variables (e.g., speed or accuracy) that are not attributable to the controlled variable. This compromises internal validly because variations in measurements are, in part, due to other effects.

On balance, the preferred procedure – that used in the majority of research studies – is to present participants with pre-selected phrases of text. Phrases are retrieved randomly from a set and are presented to participants one by one to enter.

Creating a Phrase Set
In creating a phrase set, the goal is to use phrases that are moderate in length, easy to remember, and representative of the target language.

In a recent paper comparing two soft keyboards, MacKenzie and Zhang [3] used a set of 70 phrases. We recently expanded this set to 500 phrases. A few examples from the set follow:

```
video camera with a zoom lens
have a good weekend
what a monkey sees a monkey will do
that is very unfortunate
the back yard of our house
I can see the rings on Saturn
this is a very good idea
```

We have used the new phrase set with good results in recent studies [1, 5], and wish to share them with the community of text entry researchers via this paper.

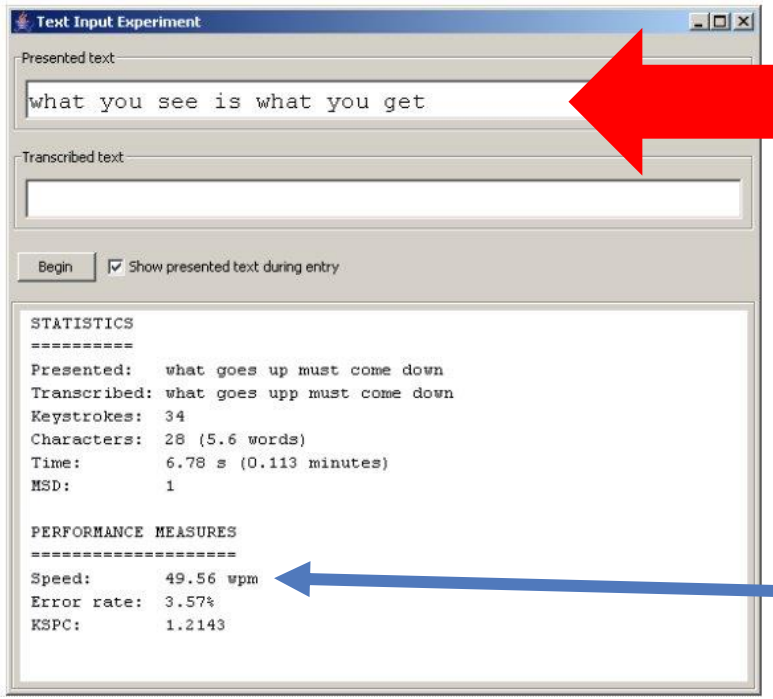
The phrases contain no punctuation symbols, and just a few instances of uppercase characters. (Participants may be instructed to ignore case and to enter all characters in lowercase.)

The complete set is available from the author in http://www.yorku.ca/mack/PhraseSets.zip

local dialect (e.g., colour vs. color).

A phrase set should be representative of the target language. The analysis of phrase sets is automated

CHI 2003, April 5-10, 2003, Ft. Lauderdale, Florida, USA.
ACM 1-58113-637-4/03/0004

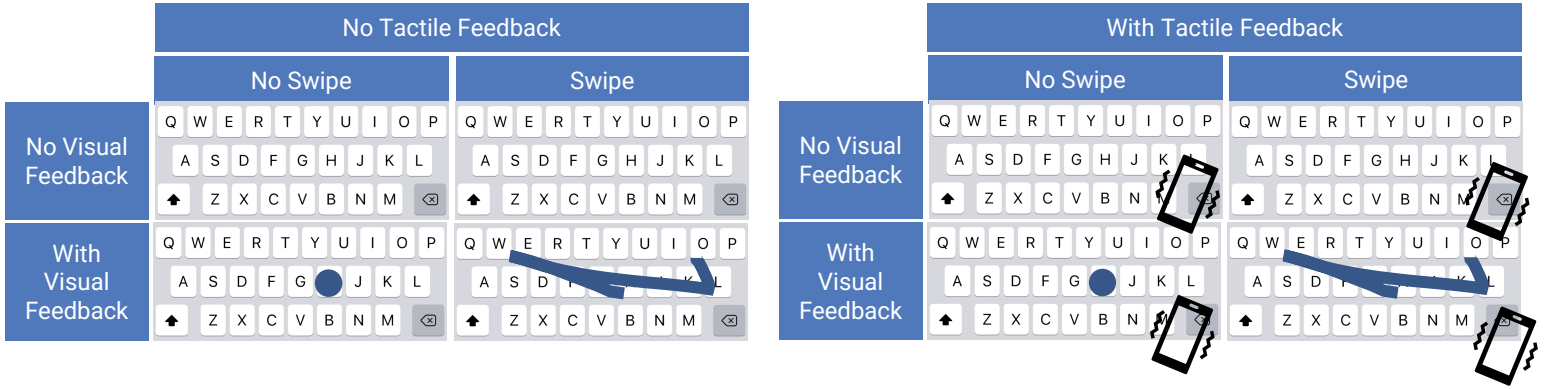


It‘s a Java program ☹
→ we need <u>that</u> on a smartphone

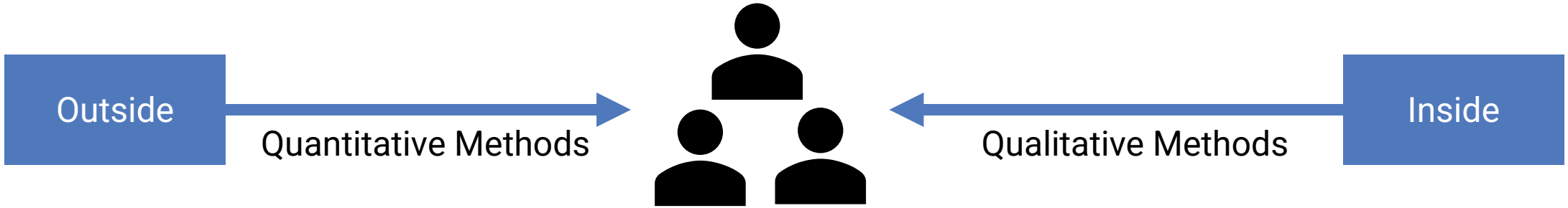
<u>Your task</u> in <u>your project</u> would be to develop and log everything correctly.

Can we now evaluate our keyboard?

- 2×2×2 Within-Subject Design, 8 Conditions, WPM, CPM, SUS, TLX, we have a task, …



- What else can we do?



Qualitative or Quantitative Method?

What?	Quantitative	Qualitative
When?	To confirm a hypothesis using a objectively or subjective measure.	Complex contexts, when little is known about the underlying mechanisms.
General Procedure	1. Selection of the quantitative methods 2. Invitation of representative population 3. Measurements in a controlled experiment 4. Statistical analysis	1. Selection of the qualitative methods 2. Selection of people (e.g., experts and/or users) 3. Data collection of feedback and/or observation 4. Transcription, translation, analysis
Data Analysis	Statistical (numerical data)	Interpretative (non-numerical data)
Method Examples	Controlled experiment, A-B testing	Semi-structured interview, group interviews (focus groups), field survey, observation, …
Advantages	▪ Confirmation of important assumptions ▪ Objective, unbiased data and evaluation ▪ Standardized procedures	▪ Generation of new knowledge ▪ Small sample is often sufficient ▪ Subjective, detailed, in-depth insights
Disadvantages	▪ Needs larger samples ▪ No deeper insights into causes ▪ Often inconclusive	▪ Often biased by the participant and researcher ▪ Needs interpretation by the researcher ▪ Researchers often tend to quantify feedback

We often need both!

Question

Are there independent or dependent variables in qualitative research?

Do we need a Balanced Latin Square in qualitative research?

IVs and DVs in Qualitative Studies?

- While **quantitative research** aim to understand the **cause-and-effect relationship**, **qualitative studies** aim to understand more **complex mechanisms behind**
  - Consquently, the **concept of IV and DV does not apply in the same manner**
  - **Qualitative research is not driven by hypotheses** and focuses on exploring and **open-ended results** gained by the collected **feedback or observations**
- **Independent variables** in qualitative research can be considered as the factor or **"subject of the investigation"**
  - The order of "discussion stimuli" should be counterbalanced, too (see: "How to conduct a user study")
- **Dependent variables** in qualitative research can be considered as the **analyzed "outcome", "feedback",** or **"responses"** from the participants
  - Results = DVs in quantitative studies = feedback or observations in qualitative studies

Quantiative and Qualitative („Mixed“) Methods: Example

- **Let’s assume:**
  - **H1a:** „Typing using Swipe improves the WPM on a touch display of smartphone.“

    → **H1a confirmed**
  - **H1b:** „Typing using Swipe decreases the SUS score on a touch display of smartphone.“

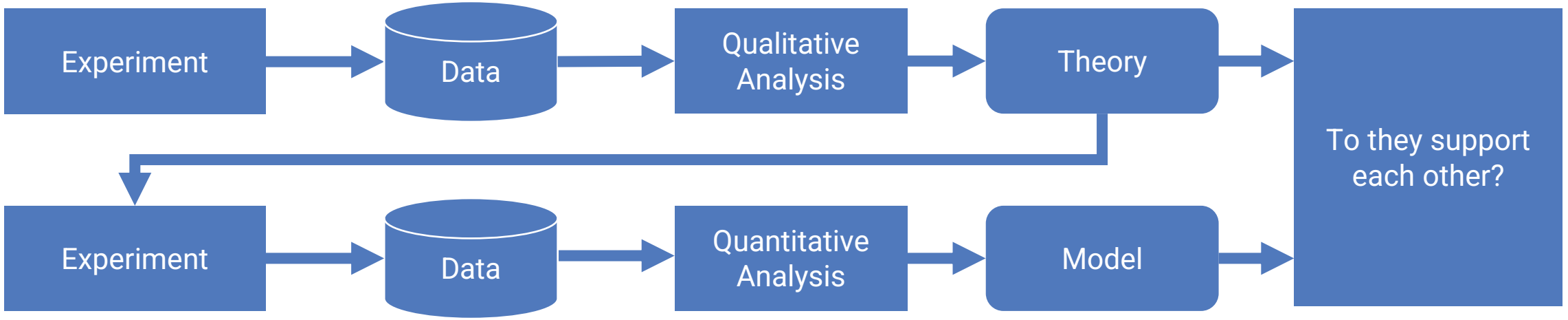
    → **H1b not confirmed**. The SUS score decreases. → Swipe has less usability!
- **Using quantitative methods** we only learned that higher typing performance (objectively) is negatively correlated to usability (subjectively).
- **Using qualitative methods** (e.g., in a semi-structured interview after each test) **we would learn that there is a reason behind that mechanism:**
  - *„The analysis of the qualitative feedback from showed that the participants perceived Swipe as faster, however, mentally more demanding **as its visual feedback was ‚distracting‘** (P1-P6, P8, and P12). P7 stated to **need „more training“** to master the Swipe technique.“*

    → They need higher mental workload because of the visual distraction and probably need more training.

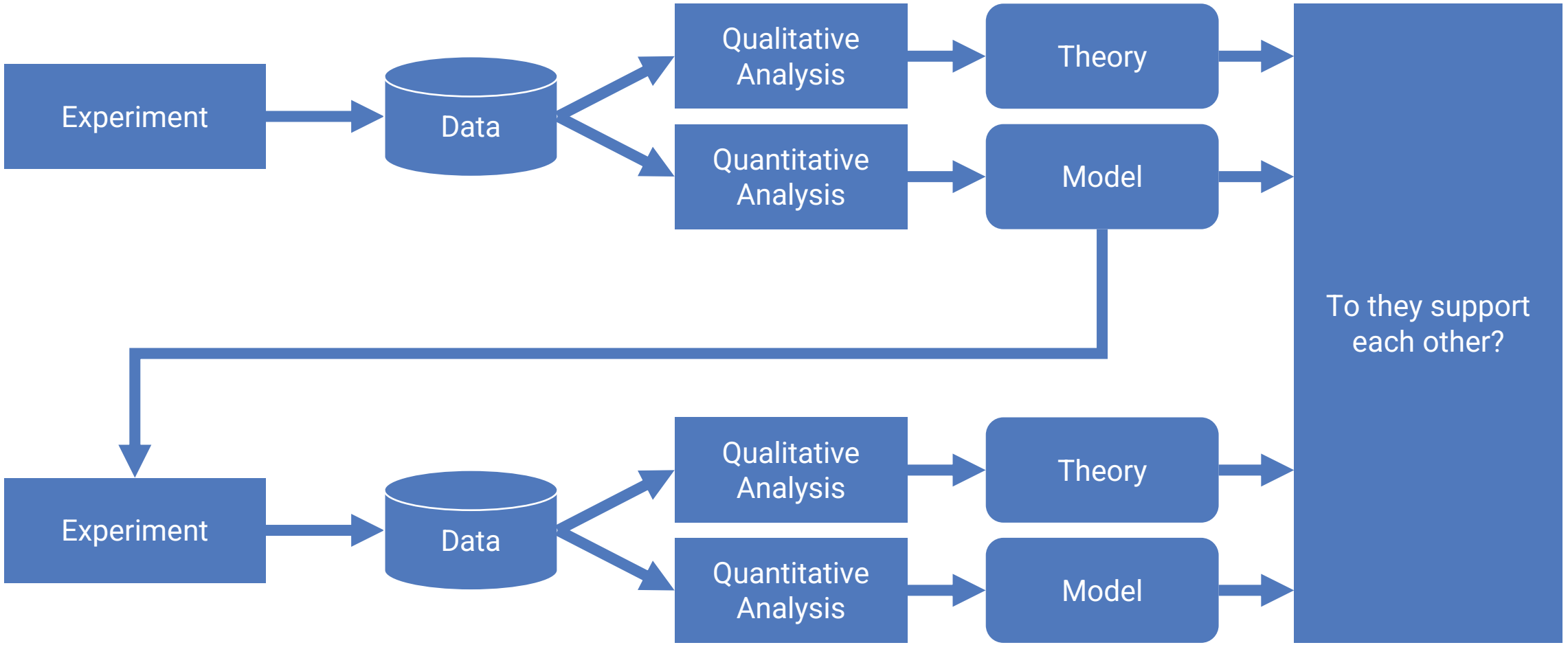
Study Designs

Our Study Design

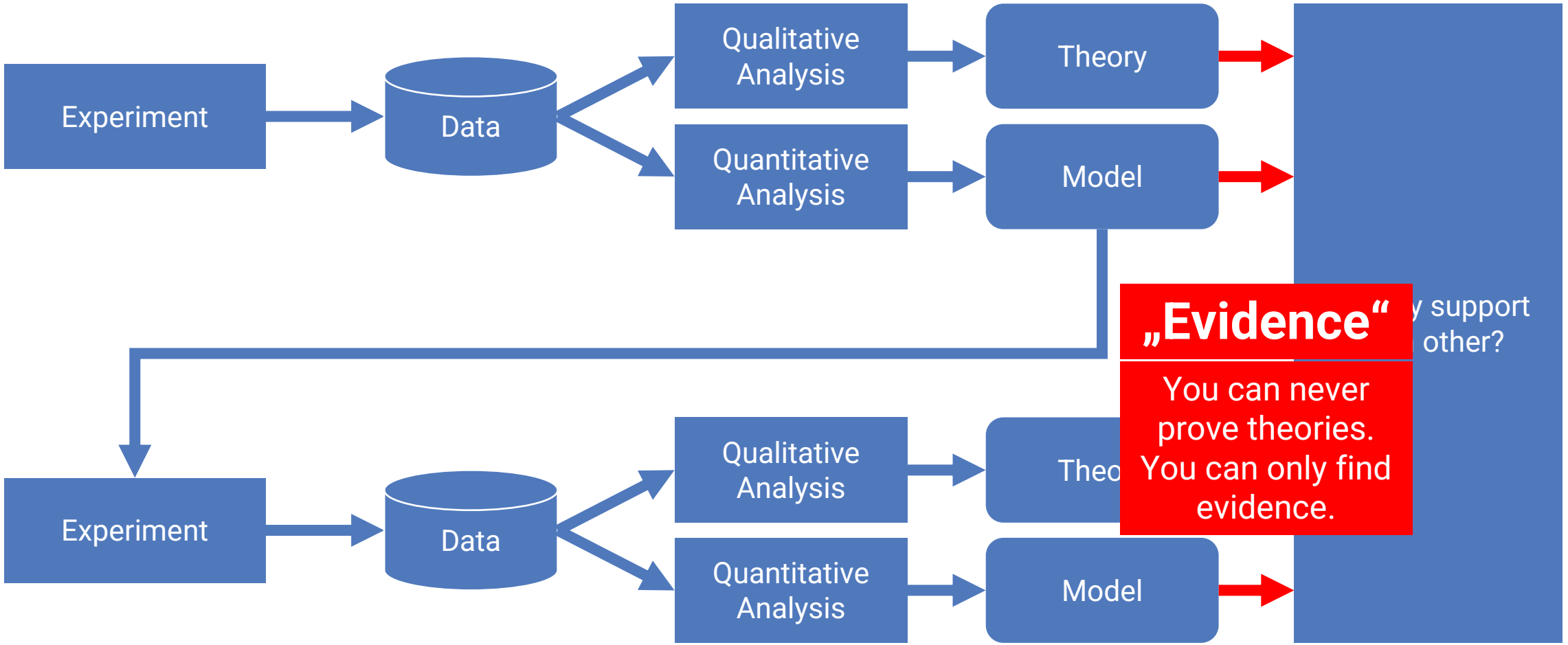
Alternative Study Design



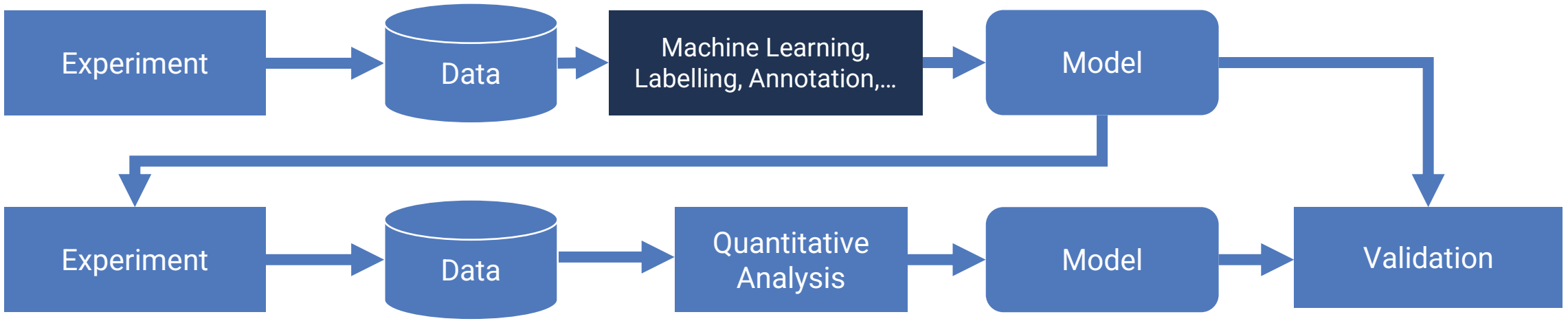
Alternative Study Design



Alternative Study Design



Data-Driven Model Validation in HCI



Objective Data (Examples)

Task Performance	▪ Task Completion Time (TCT) ▪ Success Rate ▪ Throughput (e.g., Fitts' Law) ▪ Word/Character input speed (e.g., typing) ▪ Error rate, Corrected error rate ▪ Accuracy, Precision ▪ Distance, Movements, Taps,…
Arousal/ Stress/Relaxation	▪ Galvanic Skin Response/Electrodermal Activity (GSR/EDA) ▪ Blood pressure/Heart rate (BP/HR) ▪ Electromyography (EMG)
Emotions	▪ Face recognition ▪ Electroencephalography (EEG) ▪ Heart Rate Variation (HRV)
Gaze	▪ Eye-Tracking ▪ Electroencephalography (EEG)
Sensual Acuity	▪ Discrimination/Just-noticeable-difference (JND) ▪ Lateral/Sensory Inhibition
Fitness / Ergonomics	▪ Grip Strength, ergometer Range ▪ Posture assessment

and many more…

Subjective Data (Examples)

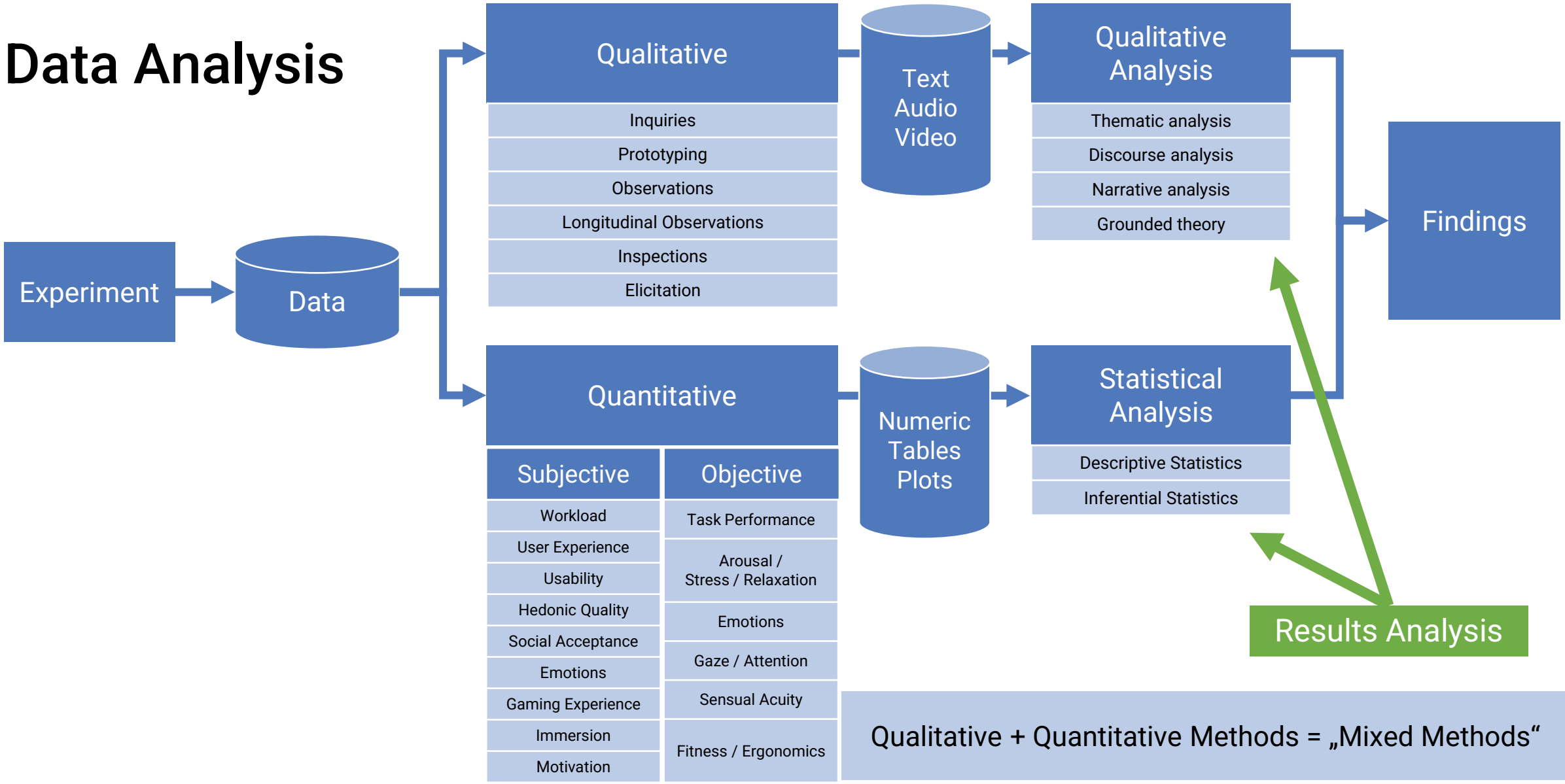
Workload	▪ Nasa Task Load Index (TLX) ▪ Subjective Mental Effort Questionnaire (SMEQ)
User Experience	▪ Usability Metric for User Experience (UMUX) ▪ Questionnaire for User Interaction Satisfaction (QUIS)
Usability	▪ System Usability Scale (SUS) ▪ Computer System Usability Questionnaire
Hedonic/Pragmatic Quality	▪ AttrakDIFF / AttrakDIFF mini
Acceptance	▪ Technology Acceptance Model (TAM) ▪ Stereotype Content Model (SCM)
Emotions	▪ Emotional Metric Outcomes (EMO) ▪ Differential Emotion Scale (DES) ▪ Positive and Negative Affect Schedule (PANAS)
Gaming Experience	▪ Game Experience Questionnaire (GEQ)
Immersion	▪ Presence Questionnaire (PQ) ▪ iGroup Presence Questionnaire (IPQ)
Motivation	▪ Intrinisic Motivation Inventory (IMI) ▪ Motivation Questionnaire (MQ)

and many more…

Qualitative Data (Examples)

Inquiries	▪ Structured, Unstructured, Semi-structured Interviews ▪ Task analysis ▪ Focus Groups ▪ Questionnaires / Surveys
Prototyping	▪ Rapid Prototyping / Paper Prototyping ▪ Tool Kit / Parts Kit
Observations	▪ Think Aloud Protocol ▪ Cooperative Evaluation ▪ Contextual Inquiries ▪ Case Study ▪ Activity / Task Analysis
Longitudinal Observations	▪ Experience Sampling ▪ Diary Studies
Inspections	▪ Cognitive Walktrough / Heuristic Evaluation ▪ Card Sorting / Tree Tests ▪ Activity Analysis ▪ Pluralistic / Consistency Inspection
Elicitation	▪ Agreement Rate (AR) ▪ Repertory Grid

Data Analysis

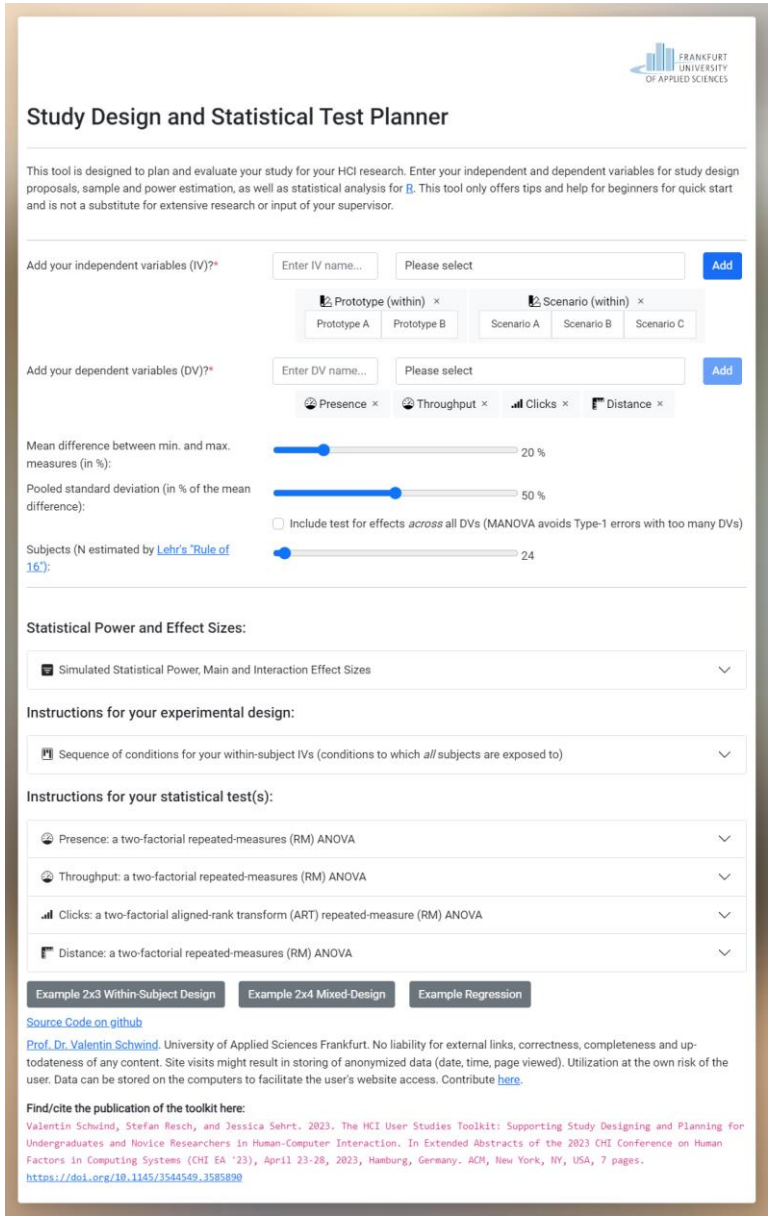


Data Analysis in HCI

- You don't need to know all of the data collection methods - just the ones you need to **answer your research question**
- **Every measure needs a method for its analysis**
  - **Objective and subjective measures** are being evaluated using **statistical tests**
    - Statistical tests depend on the hypothesis
    - We will talk about statistical tests later (you will love it)
  - **Observations and qualitative feedback**?
    - **Thematical Analysis**
    - **Content Analysis**
    - **Narrative Analysis**
    - **Grounded Theory**
    - **…**
- Is there more we can analyze?

Exercise: Study Design Plan

- **Use our tool to find and plan your study design**
  - Goto: https://hci-studies.org/study-design-planner/
- **Enter are your IVs, levels and DVs**
  - Ignore the sliders and statistics
  - Go to “Instructions for your experimental design”
- **Discuss in your group:**
  1. What is the proposed study design of your study?
  2. Does this work when you have a qualitative study?
  3. Is that here important when you do a literature review?



Tasks Next Time (TNT): Study Design / Study Plan

- **Update your problem statement, add a section called „Method“**
  - Write down in your own words what you found in „Instructions for your experimental design” or explain in your own words your study plan
- **Update and upload your presentation (PDF)** with **a study design** that includes:
  - The incorporated feedback and things you need to update
  - Your update research question(s) and describe the operationalization of your concept(s)
  - Describe your method in a few words. Depending on your research method …
    - quantitative: Name your IVs/DVs? within-subject? between-subject? levels? hypotheses?
    - qualitative: Name your method and study plan? (factor, desired feedback, …)
    - literature review: Explain what of the stuff here is important for you?
  - **Draw your study design / study plan!**