Workshop Research Methods and Statistical Analysis

Session 1 - Introduction

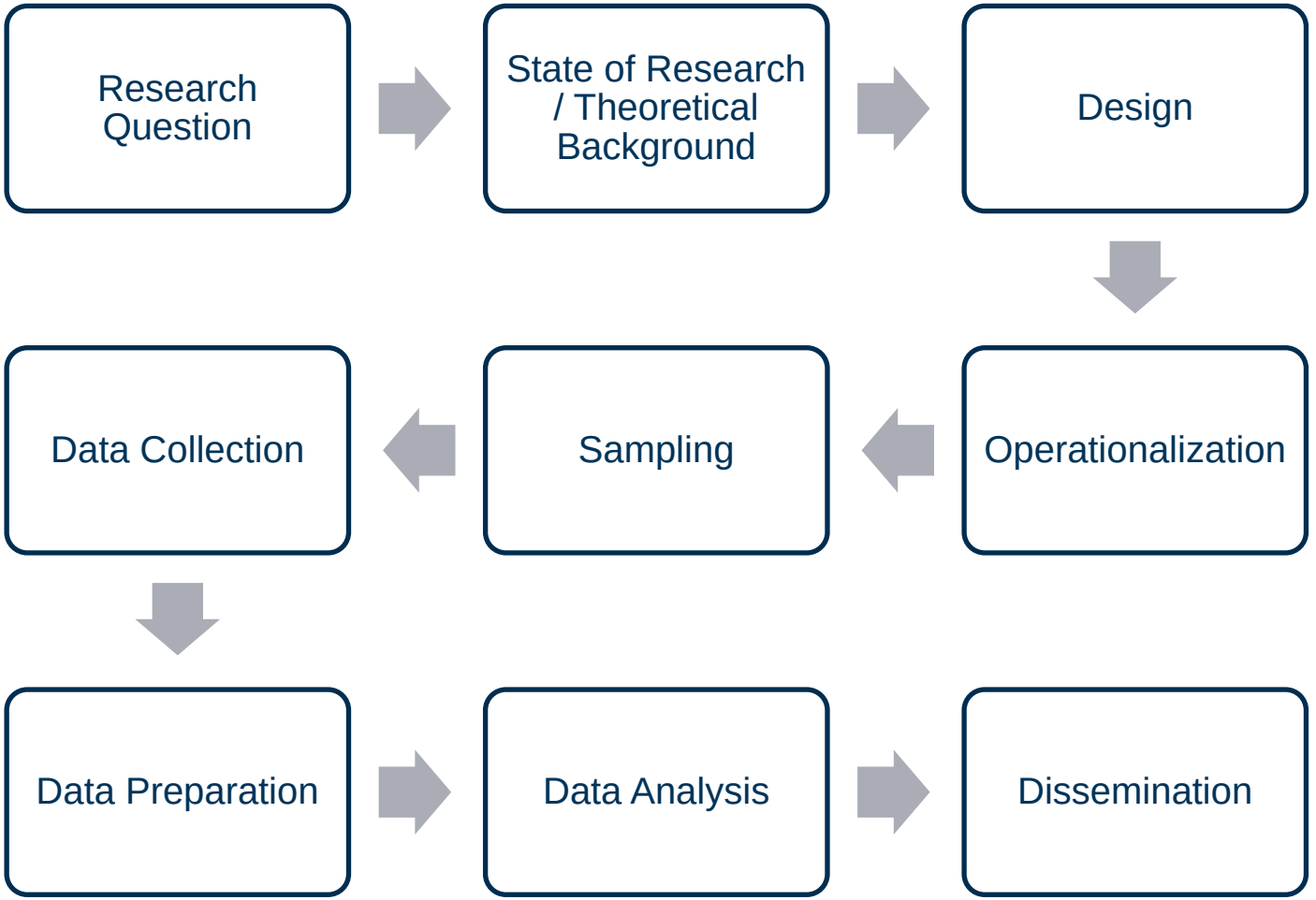
Sandra Poeschl

Agenda

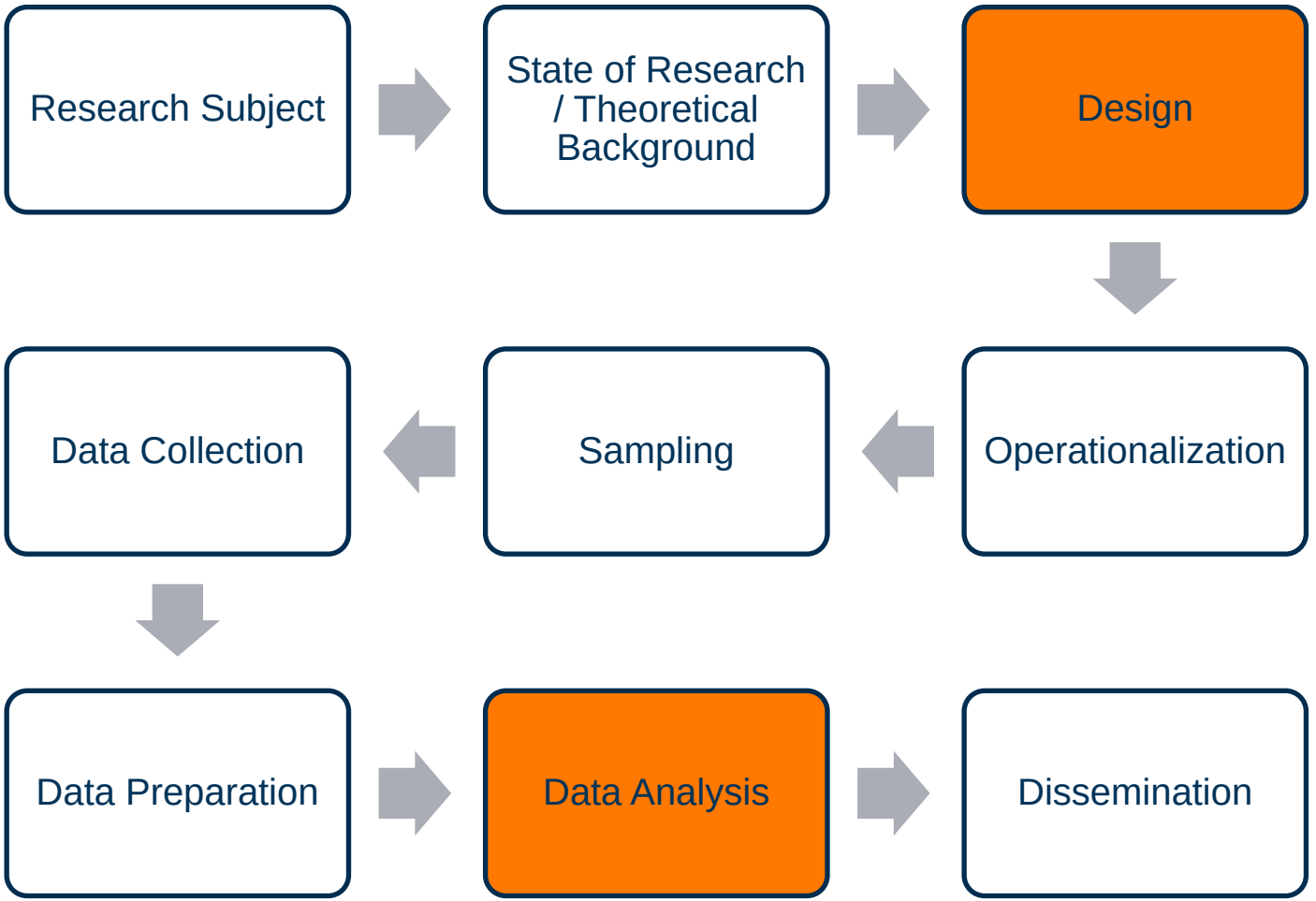
- Empirical Research Process
- Research Designs
- Effect Sizes & Power Analysis

EMPIRICAL RESEARCH PROCESS

Research Process



Research Process



RESEARCH DESIGNS

Research Design in General

Design Characteristic	Design Alternatives
Philosophy of science approach	Qualitative study **Quantitative study** Mixed Methods study
Goal	Basic Research study Applied Research study
Purpose	Theoretical Study / Research review Methodological Study **Empirical Study** •Original Study •Replication Study
Data Basis	**Primary analysis** Secondary analysis Meta-analysis

Research Design in General

Design Characteristic	Design Alternatives
Interest	exploratory study population descriptive study **explanatory study**
Treatment of Groups	**experimental study, true experiment** quasi-experimental study non experimental study
Location	**laboratory study** field study
Frequency of measurements	**cross-sectional design** **repeated measurements design** longitudinal design
Number of research objects	single participant study group study •**sample study** •population study

Experimental Designs

- Testing for differences between groups
- experimental study
  - At least 2 groups
  - Randomization
  - experimental manipulation of treatment (independent variables, causes)
  - Measurement of dependent variable/s (effects).
- quasi-experimental study
  - No randomization, existing groups

Causal influences

- Independent variable(s) → dependent variable(s)
- internal validity: results allow clear causal interpretation of effects
- Exclusion of alternative explanations
- Confounder (influences on dependent variable beyond independent variables)
  - Subject-related confounders (randomization)
  - Study-related confounders (standardization, (double-) blind trials)

Common variants

- 2-Groups (Treatment / Control):
  - 1 IV, dichotomous, 1 DV, metric (univariate)
  - Cross-sectional or repeated measures
  - t-Test (independent / dependent samples)

No sound	Spatial sound

IV: sound (no sound/spatial sound
DV: time to complete a orientation task

Common variants

- One-way, univariate:
  - 1 IV, more than 2 levels (nominal), 1 DV (metric)
  - Cross-sectional or repeated measures
  - One-way, univariate ANOVA (repeated measures)

FOR 20	FOR 90	FOR 270

IV: FOR (20 degrees/90 degrees/270 degrees)
DV: error rate in search task

Common variants

- Multi-factorial, univariate:
  - At least 2 IV, 1 DV, metric (univariate)
  - Cross-sectional or repeated measures
  - Interaction effects
  - Multi-factorial, univariate ANOVA

	No head-tracking	Head-tracking
No stereoscopy		
stereoscopy		

IV 1: head-tracking (yes/no)
IV 2: stereoscopy (yes/no)
DV: error rate in spatial judgement task

Multivariate

At least 2 DV (metric),One- and multi-factorial MANOVA

EFFECT SIZES & POWER ANALYSIS

A problem with significance tests (NHST)

Novice	Experts
2	2
3	3
4	4
1	1
2	2
3	3
2	2
3	3
4	4
5	**4**
M = 2.90 SD = 1.20 n = 10	M = 2.80 SD = 1.03 n = 10

- IV: Experience
- DV: Error rates in visualization task
- H_1: $\mu_{\text{no experience}} > \mu_{\text{experience}}$
- H_0: $\mu_{\text{no experience}} \leq \mu_{\text{experience}}$
- $n = 20$: $t_{\text{emp (df=18)}} = .20$, $p = .42$

- Can we have a minimal effect that is statistically significant?

A problem with NHSTs

Novice	Experts
2	2
3	3
4	4
1	1
2	2
3	3
2	2
3	3
4	4
5	**4**
M = 2.90 SD = 1.20 n = 10	M = 2.80 SD = 1.03 n = 10

- n = 20: t(df = 18) = .20, p = .42 n.s.
- n = 40: t(df = 38) = .29, p = .38 n.s.
- n = 80: t(df = 78) = .42, p = .33 n.s.
- n = 160: t(df = 158) = .59, p = .27 n.s.
- n = 320: t(df = 318) = .84, p = .20 n.s.
- n = 640: t(df = 638) = 1.19, p = .11 n.s.
- n = 1280: t(df = 1278) = 1.68, p = .05*

- NHSTs will always lead to significant results if n is large enough, even when effects are minimal and of no practical significance!

Effect size

- Reporting statistical significance (test statistics, p-value) + standardized effect size
- Statistically significant results do not automatically have to be of practical relevance.
- Practical relevance has to be decided with regards to content.
- Absolute effect sizes are hard to compare across studies (mean difference of 0.8 errors between groups …) → standardized effect sizes

Some standardized effect sizes

Type of effect size measure	Small effect	Medium effect	Large effect
Difference between 2 groups Cohen‘s **d**	0,2	0,5	0,8
Correlations Bivariate Pearson‘s correlation **r**	0,10	0,3	0,5
Variance explained (Partial) Eta Squared η^2	0,01	0,06	0,14

Cohen, 1988

Another problem with NHSTs

- H1: Higher simulation fidelity leads to higher number of saves in a VR goalkeeper training simulation.
- Could we have an effect of practical relevance without statistical significance?

	Low SF	High SF
M	2.40	2.70
SD	.94	.98
n	20	20

t(df = 38) = .99, p = .33
Mean difference = .30
SE of difference = .30

Another problem with NHSTs

- H1: Higher simulation fidelity leads to higher number of saves in a VR goalkeeper training simulation.
- We can have medium and large effects without statistical significance, if n is too small!

	Low SF	High SF
M	2.40	2.70
SD	.94	.98
n	20	20

t(df = 38) = .99, p = .33
Mean difference = .30
SE of difference = .30

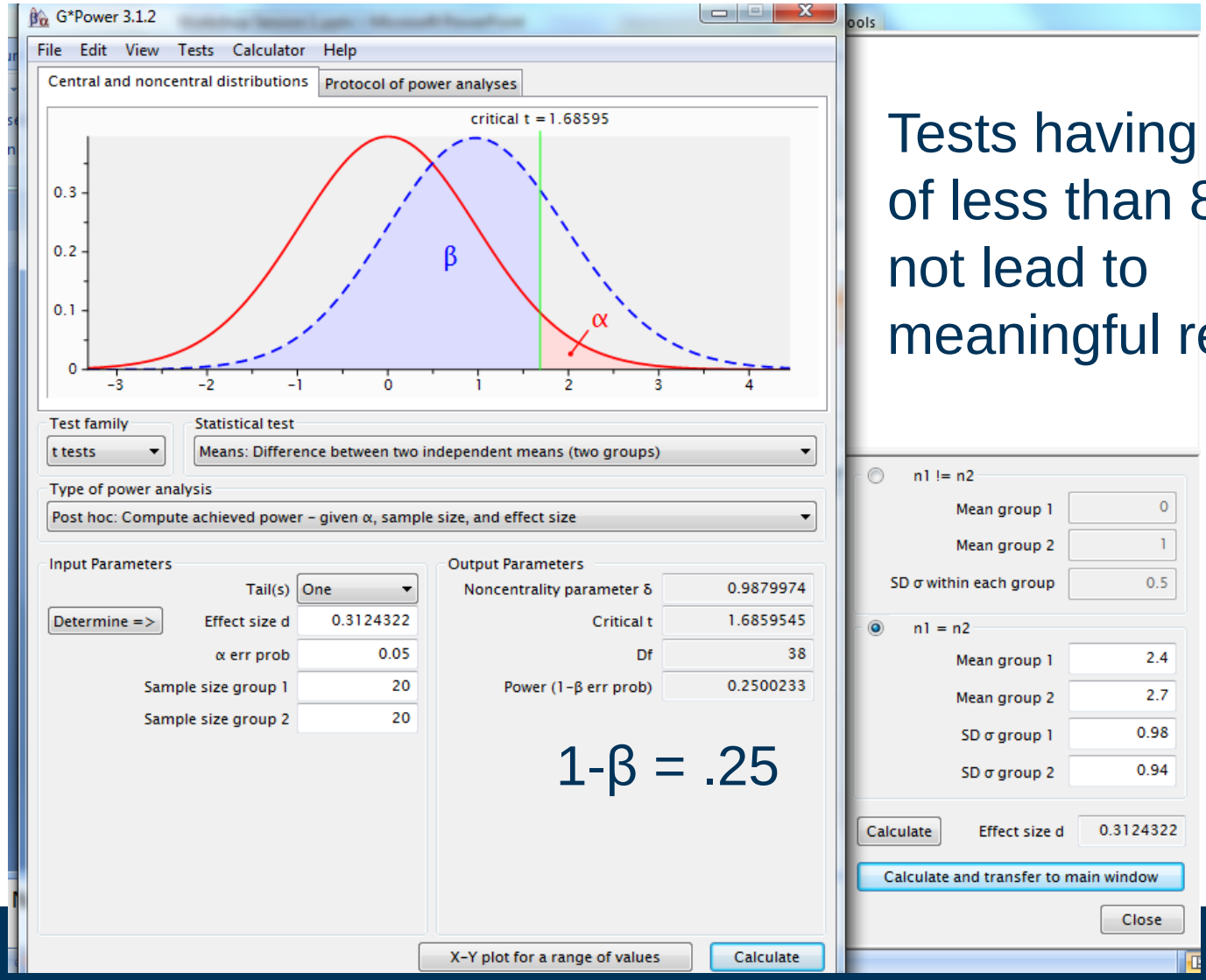
d = .31

Power

- Do a power analysis if you only find non-significant effects
- Even strong effects fail to gain statistical significance if the sample size is too small.
- Power (1-β) is the probability to find a significant population effect.
- Power should be at least 80 % (1-β ≥ .80).
- G*power for power analysis

  http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/

Power Analysis with g*power



Tests having a power of less than 80 % do not lead to meaningful results.

Power

- Power (1-β) increases with
  - Increase of sample size n
  - Increase of population effects
  - Increase of significance level

- We can‘t influence population effects and can‘t change the significance level.
- Controlling power → controlling sample size

In a nutshell

- **Small effects can be statistically significant** (especially when n is large): report and classify standardized effect sizes.
- **Relevant effects can be non-significant**: do a post-hoc power analysis with g*power. If power is less than 80 %, the results are not meaningful. Studies should be replicated with larger samples.

Required sample size

- To avoid problems with the NHST, we should take required sample sizes into account before data collection.
- Required sample sizes are construed to detect a priori determined effect sizes (small, medium, large) with test power of $1-\beta = .80$ and a significance level of $\alpha = .05$.
- You can use g*power to compute required sample sizes.

Example SF

- H1: Higher simulation fidelity leads to higher number of saves in a VR goalkeeper training simulation.
- What n do we need to detect an effect of d = .30 (small to medium)?

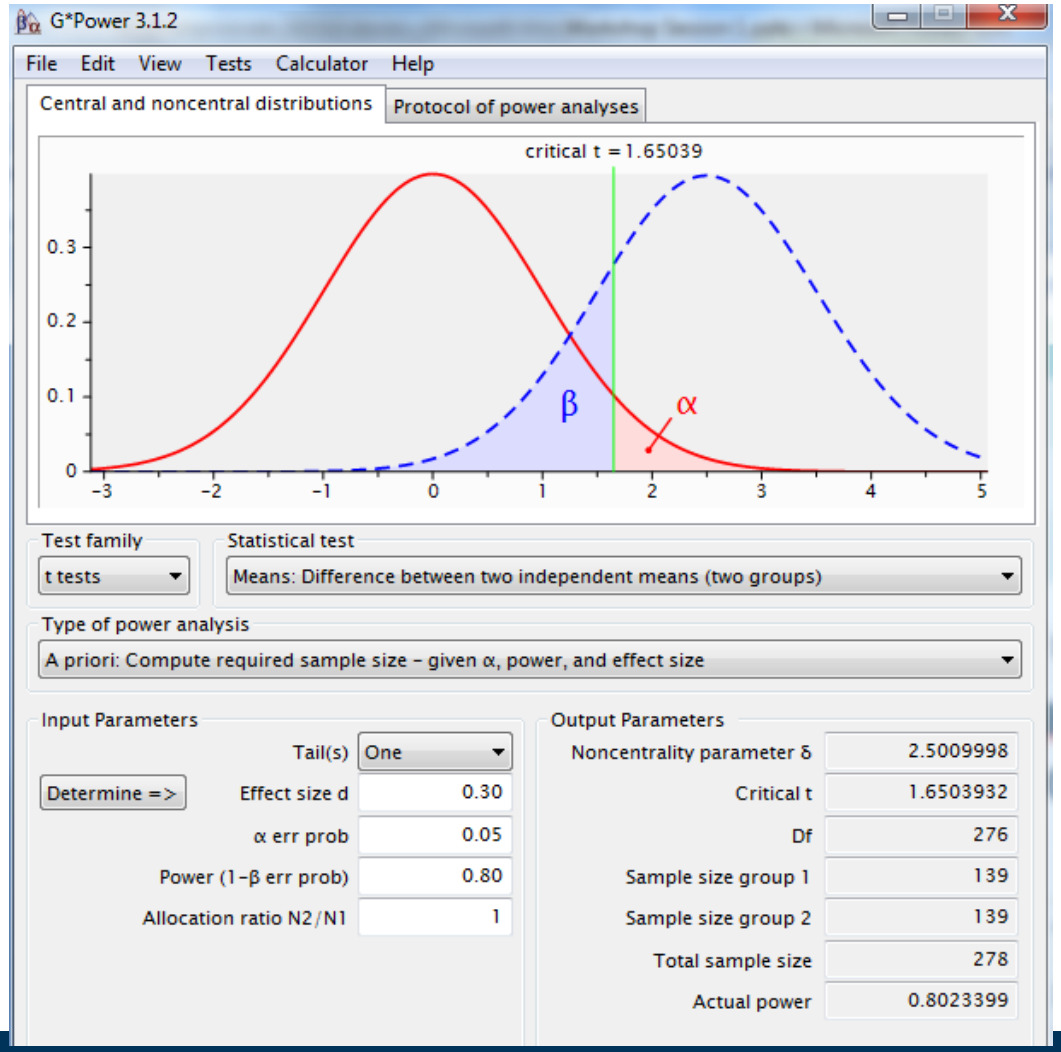
	Low SF	High SF
M	2.40	2.70
SD	.94	.98
n	20	20

$t(df = 38) = .99, p = .33$
Mean difference = .30
SE of difference = .30

$d = .31$

$1-\beta = .25$

Example SF



$n_{1requ} = n_{2requ} = 139$

Further Reading

- Kantowitz, B., Roediger, H., & Elmes, D. (2008). *Experimental Psychology*, International Edition (9th ed.). Andover: Cengage Learning Emea.
- Marques de Sá, J. P. (2007). *Applied Statistics Using SPSS, STATISTICA, MATLAB and R*. (2nd ed.) Berlin: Springer.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and Quasi-Experimental Design for Generalized Causal Inference*. Wadsworth.