

Potsdam, 21.03.2024

Poweranalyse für experimentelle Designs in R

Workshop auf der Nachwuchstagung der GEBF 2024




Sophie E. Stallasch

Quantitative Methoden in den Bildungswissenschaften

✉ stallasch@uni-potsdam.de

Ablauf des Workshops

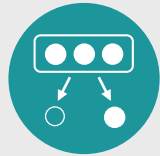
 ~ 20 Min.



KONZEPTUELLE GRUNDLAGEN

Das «Was» und «Warum» der statistischen Power(analyse)

 ~ 40 Min.



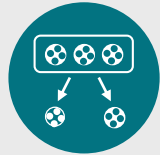
TEIL I: EINFACHE (NICHT-HIERARCHISCHE) EIN-EBENEN-DESIGNS

Poweranalysen für individuell-randomisierte Studien

 ~ 10 Min.


PAUSE

 ~ 70 Min.



TEIL II: KOMPLEXERE MEHREBENEN-DESIGNS

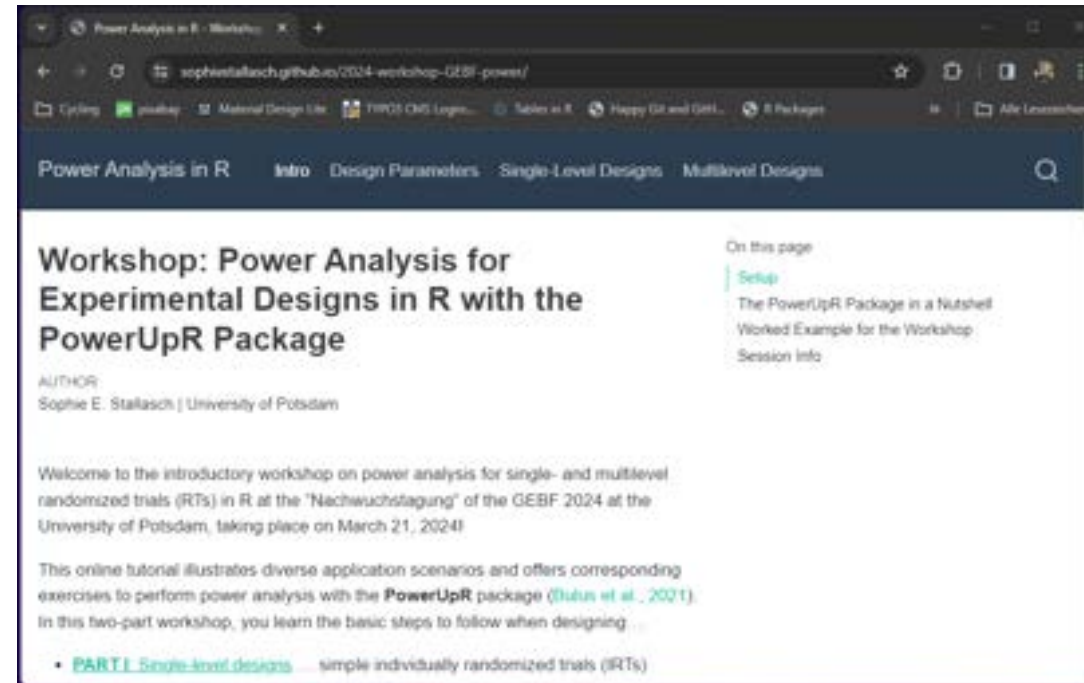
Poweranalysen für cluster-randomisierte Studien

 ~ 10 Min.



FRAGEN & DISKUSSION

Workshop-Website



<https://sophiestallasch.github.io/2024-workshop-GEBF-power>

Poweranalyse für experimentelle Designs in R



Konzeptuelle Grundlagen

Interventionseffekt: „Signal-Rauschen“-Verhältnis



Was ist statistische Power?

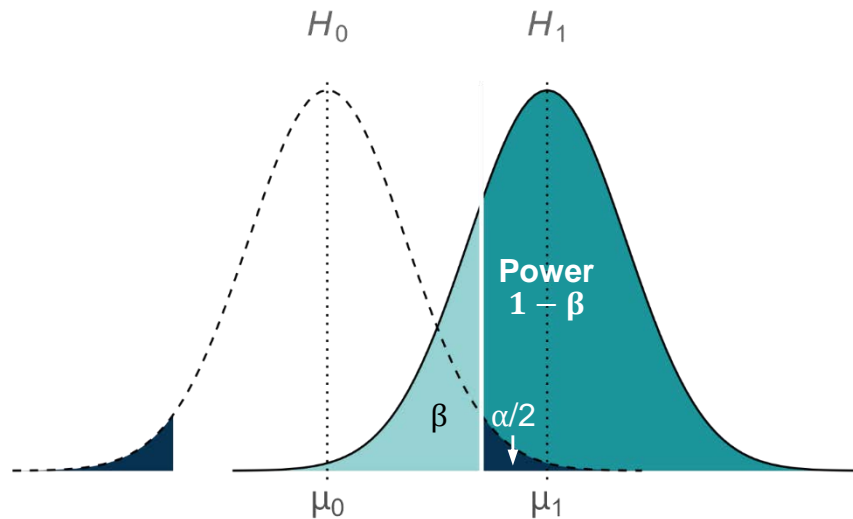
Wahrscheinlichkeit, einen Effekt zu finden, wenn dieser in der Population existiert.

„Teststärke [engl. statistical power], [FSE], die Teststärke eines stat. Tests (Signifikanztest) ist die Wahrscheinlichkeit $1-\beta$ eines signifikanten Testergebnisses bei Gültigkeit der Alternativhypothese (H_1).“

Erdfelder E. (2022). Teststärke. In M. A. Wirtz (Hrsg.), *Dorsch. Lexikon der Psychologie*.
<https://dorsch.hogrefe.com/stichwort/teststaerke>

Statistische Schlussfolgerungen

Fehler 1. Art und Fehler 2. Art



μ_0 ... Populationsmittelwert unter H_0 (Kontrollgruppe)

Test/Schlussfolgerung	Population	
	Effekt existiert ($\mu_0 \neq \mu_1$) H_0 falsch	Effekt existiert nicht ($\mu_0 = \mu_1$) H_0 wahr
Effekt signifikant ($p < \alpha$) H_0 verwerfen	✓ $1 - \beta$ (Power)	✗ Fehler 1. Art α „false positive“
Effekt nicht signifikant ($p \geq \alpha$) H_0 beibehalten	✗ Fehler 2. Art β „false negative“	✓ $1 - \alpha$

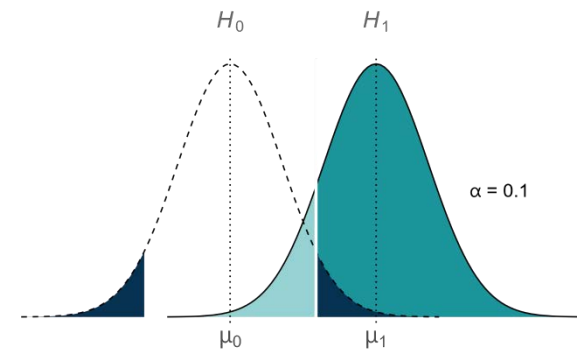
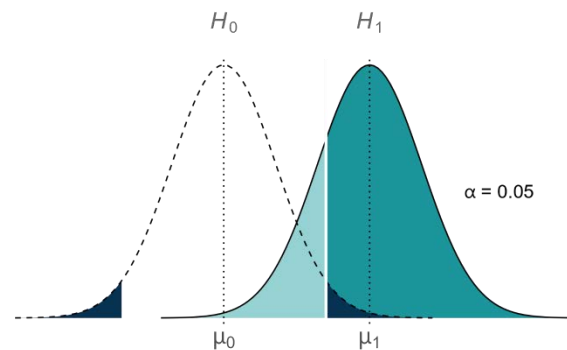
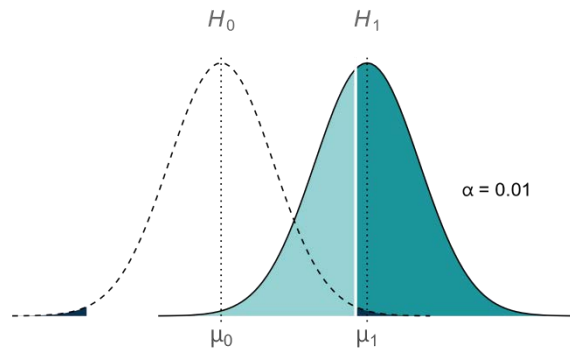
μ_0 ... Populationsmittelwert unter H_1 (Interventionsgruppe)

Statistische Power

Die wichtigsten Determinanten

α -LEVEL (SIGNIFIKANZNIVEAU)

Je größer α , desto höher die Power. (→ Einseitiger Test mehr Power als zweiseitiger.)



Statistische Power

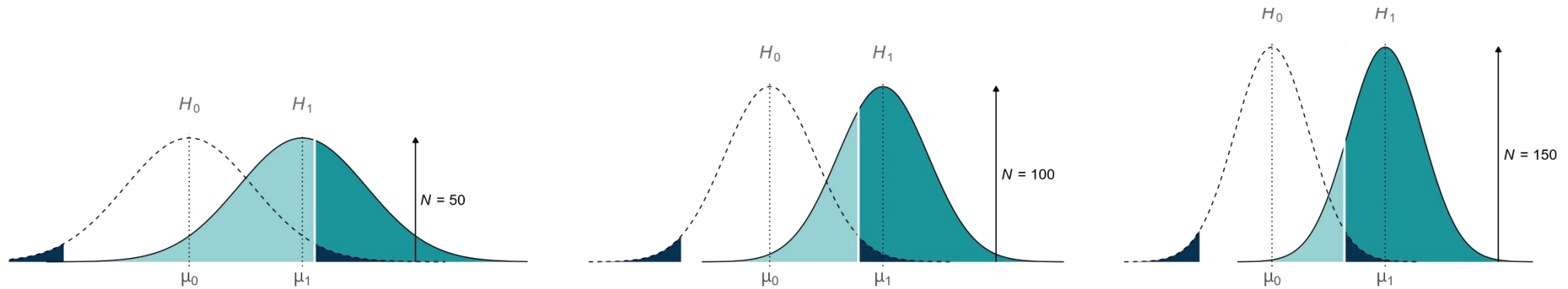
Die wichtigsten Determinanten

α -LEVEL (SIGNIFIKANZNIVEAU)

Je größer α , desto höher die Power. (→ Einseitiger Test mehr Power als zweiseitiger.)

STICHPROBENGROÖßE

Je größer N , desto höher die Power.



Statistische Power

Die wichtigsten Determinanten

α -LEVEL (SIGNIFIKANZNIVEAU)

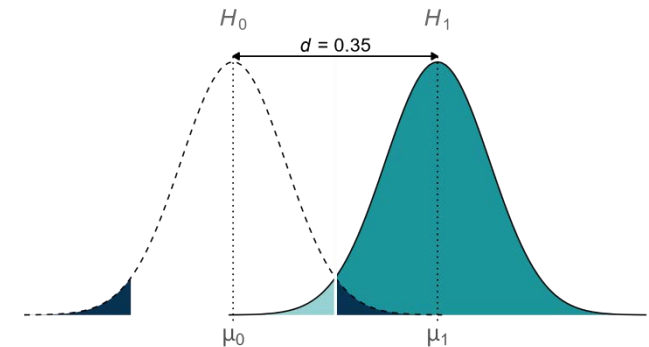
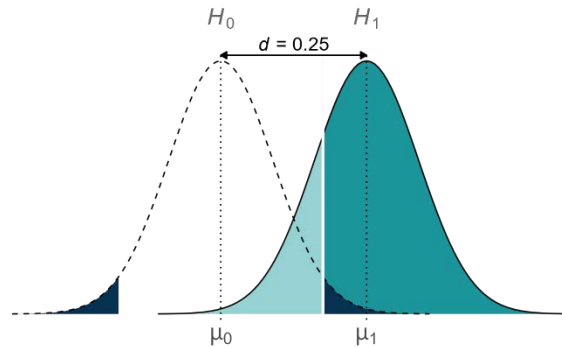
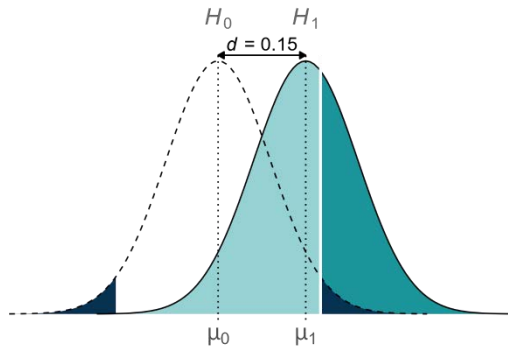
Je größer α , desto höher die Power. (→ Einseitiger Test mehr Power als zweiseitiger.)

STICHPROBENGROÖßE

Je größer N , desto höher die Power.

EFFEKTGRÖÖßE

Je größer d , desto höher die Power.



Statistische Power

Die wichtigsten Determinanten

α -LEVEL (SIGNIFIKANZNIVEAU)

Je größer α , desto höher die Power. (→ Einseitiger Test mehr Power als zweiseitiger.)

STICHPROBENGROÖßE

Je größer N , desto höher die Power.

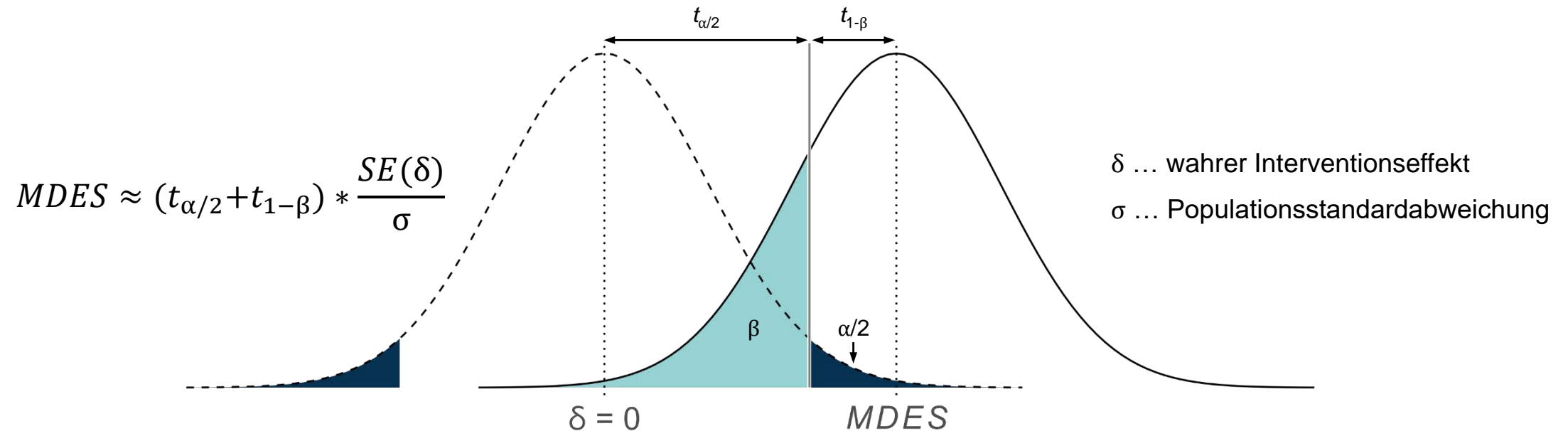
EFFEKTGRÖÖßE

Je größer d , desto höher die Power.

<https://rpsychologist.com/d3/nhst/>

Kleinstmöglich auffindbare (standardisierte) Effektgröße

Minimum Detectable Effect Size (*MDES*; Bloom, 1995, 2005)



MDES: Mehrfaches des *SEs* des Interventionseffektes = Maß für die Schätzgenauigkeit

(Mindestens) 6 gute Gründe, eine Poweranalyse (a priori) durchzuführen

Vermeidung von sowohl unter- als auch überpowererten Studien



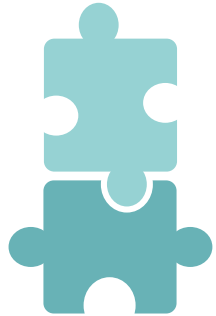
AUSSAGEKRÄFTIGE BEFUNDE

Informativ mit Blick auf ein bestimmtes inferenzstatistisches Ziel (Hedberg, 2018; Lakens, 2022)

→ Figure 3 (p. 164) in “Rigorous Large-Scale Educational RCTs Are Often Uninformative: Should We Be Concerned?” by H. Lortie-Forgues and M. Inglis, 2019, *Educational Researcher*, 48(3), 158–166 (<https://doi.org/10.3102/0013189X19832850>). Copyright 2019 by AERA.

(Mindestens) 6 gute Gründe, eine Poweranalyse (a priori) durchzuführen

Vermeidung von sowohl unter- als auch überpowererten Studien



AUSSAGEKRÄFTIGE BEFUNDE

Informativ mit Blick auf ein bestimmtes inferenzstatistisches Ziel (Hedberg, 2018; Lakens, 2022)

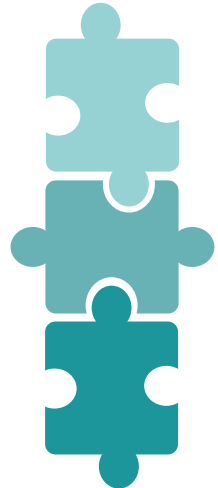
KORREKTE BEFUNDE

Größe und Richtung des Effekts (Gelman & Carlin, 2014; Ioannidis, 2005, 2008; Sims et al., 2022)

→ Figure 2 (p. 644) in “Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors” by A. Gelman and J. Carlin, 2014, *Perspectives on Psychological Science*, 9(6), 641–651 (<https://doi.org/10.1177/1745691614551642>). Copyright 2014 by the author(s).

(Mindestens) 6 gute Gründe, eine Poweranalyse (a priori) durchzuführen

Vermeidung von sowohl unter- als auch überpowererten Studien



AUSSAGEKRÄFTIGE BEFUNDE

Informativ mit Blick auf ein bestimmtes inferenzstatistisches Ziel (Hedberg, 2018; Lakens, 2022)

KORREKTE BEFUNDE

Größe und Richtung des Effekts (Gelman & Carlin, 2014; Ioannidis, 2005, 2008; Sims et al., 2022)

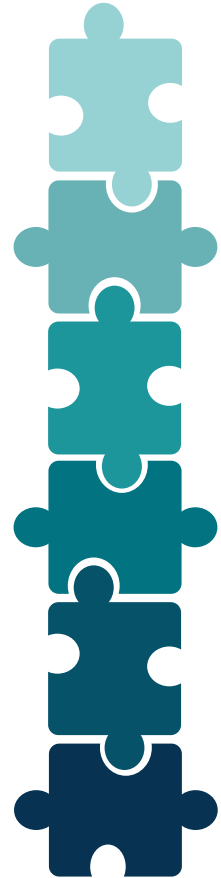
REPRODUZIERBARKEIT & REPLIZIERBARKEIT

(Open Science Collaboration, 2015)

→ Figure 1 (p. 610) in “At What Sample Size Do Correlations Stabilize?” by F. D. Schönbrodt and M. Perugini, 2013, *Journal of Research in Personality*, 47, 609–612
(<http://dx.doi.org/10.1016/j.jrp.2013.05.009>). Copyright 2013 by Elsevier Inc.

(Mindestens) 6 gute Gründe, eine Poweranalyse (a priori) durchzuführen

Vermeidung von sowohl unter- als auch überpowererten Studien



AUSSAGEKRÄFTIGE BEFUNDE

Informativ mit Blick auf ein bestimmtes inferenzstatistisches Ziel (Hedberg, 2018; Lakens, 2022)

KORREKTE BEFUNDE

Größe und Richtung des Effekts (Gelman & Carlin, 2014; Ioannidis, 2005, 2008; Sims et al., 2022)

REPRODUZIERBARKEIT & REPLIZIERBARKEIT

(Open Science Collaboration, 2015)

RESSOURCENSCHONUNG

Finanziell, materiell, personell (Bausell & Li, 2002; Halpern et al., 2002; Lenth, 2001)

FORSCHUNGSANTRÄGE

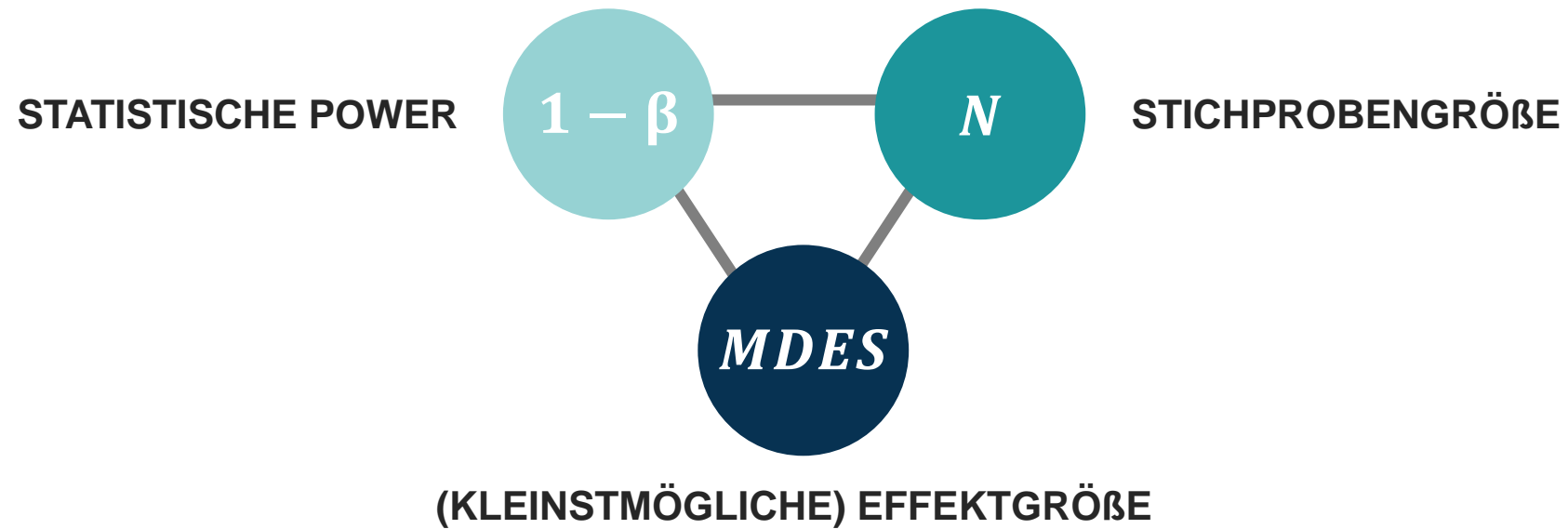
(Education Endowment Foundation, 2022; German Research Foundation, 2022; Institute of Education Sciences, 2023)

STANDARDS GUTER WISSENSCHAFTLICHER PRAXIS

“basic expectations for quantitative research reporting” (JARS-Quant; American Psychological Association, 2020, p. 77)

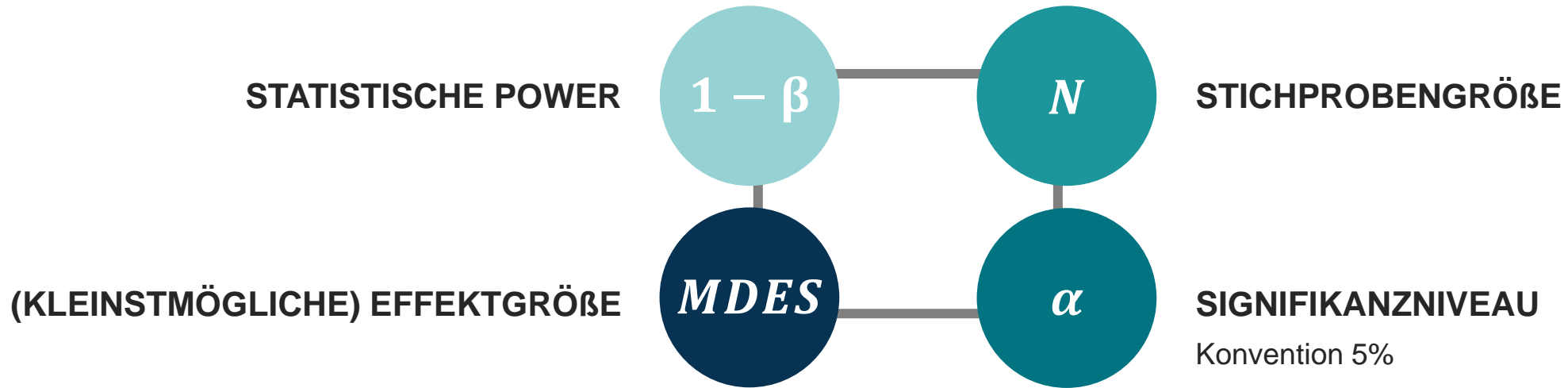
Poweranalyse

Outputs



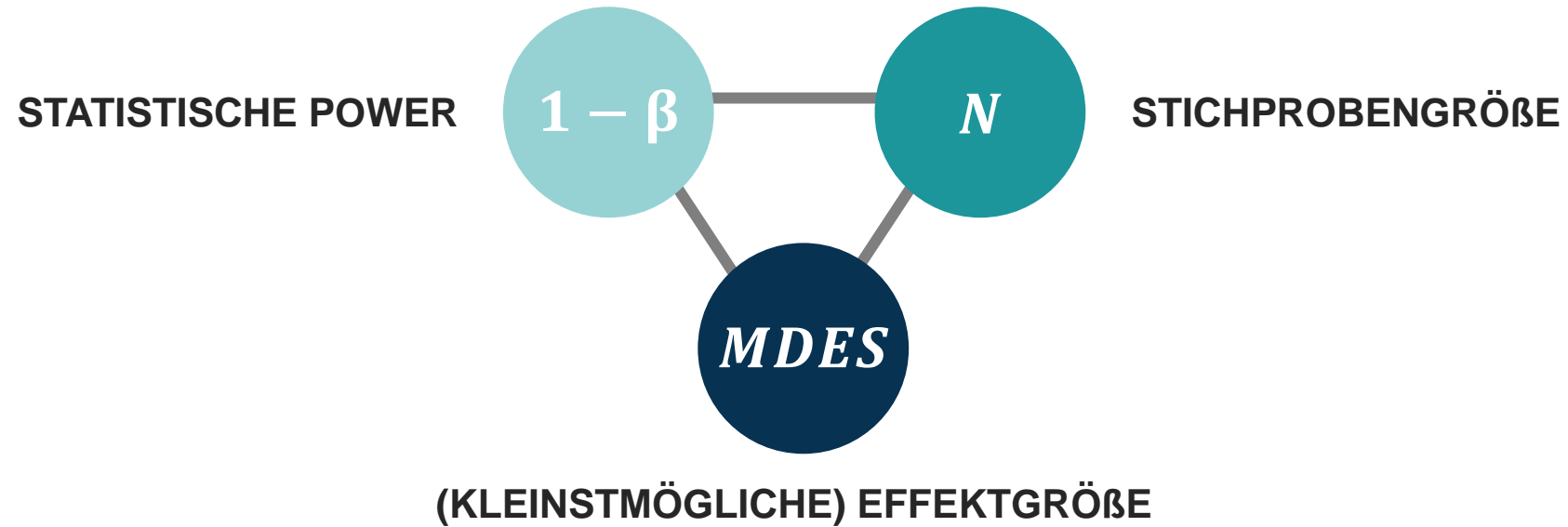
Poweranalyse

Outputs



Poweranalyse

Outputs



Die Effektgröße – „Der problematische Parameter“

Rationalen zur Identifikation der *MDES* (Bloom, 2006; Brunner et al., 2018; Schochet, 2008)

ÖKONOMISCH / KOSTEN-NUTZEN

Welcher Effekt wiegt die Kosten der Intervention auf?

→ Kosten der Intervention müssen a priori abgeschätzt werden

PROGRAMMATISCH

Welcher Effekt ist erreichbar gegeben des Kontexts der Interventionsstudie?

→ Orientierung an vorheriger Forschung möglich, ABER Vergleichbarkeit muss geprüft werden

POLITISCH

Welcher Effekt genügt den (politischen) Entscheidungsträgern?

→ Empirisch etablierte Benchmarks werden benötigt

Benchmarks für leistungsbezogene Interventionseffekte



Journal of Research on Educational Effectiveness

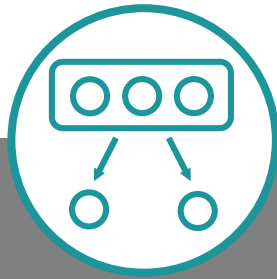
 Routledge
Taylor & Francis Group

ISSN: (Print) (Online) Journal homepage: <https://www.tandfonline.com/loi/uree20>

Empirical Benchmarks to Interpret Intervention Effects on Student Achievement in Elementary and Secondary School: Meta-Analytic Results from Germany

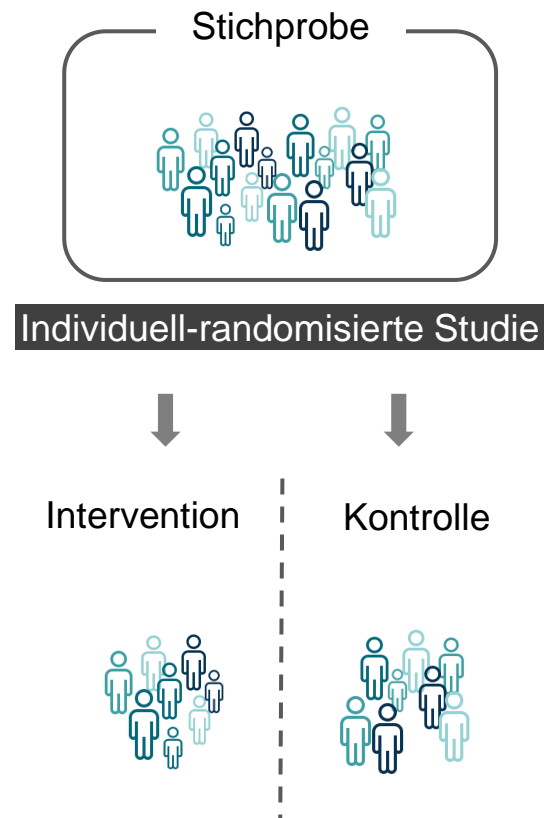
Martin Brunner, Sophie E. Stallasch & Oliver Lüdtke

Poweranalyse für experimentelle Designs in R



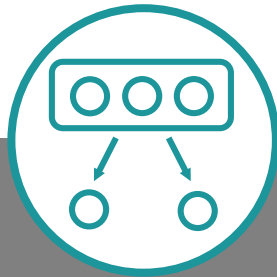
Teil I: Ein-Ebenen-Designs

Randomisierte Interventionsstudien: Ein-Ebenen-Design



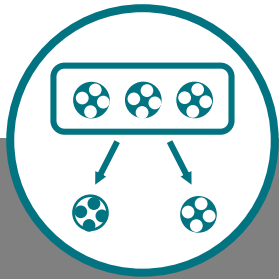
<https://sophiestallasch.github.io/2024-workshop-GEBCF-power>

HANDS-ON!



Teil I: Ein-Ebenen-Designs

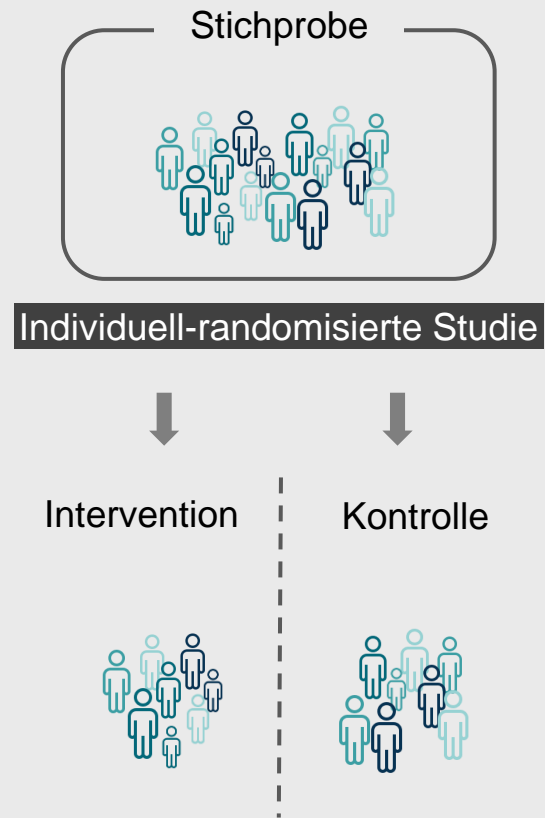
Poweranalyse für experimentelle Designs in R



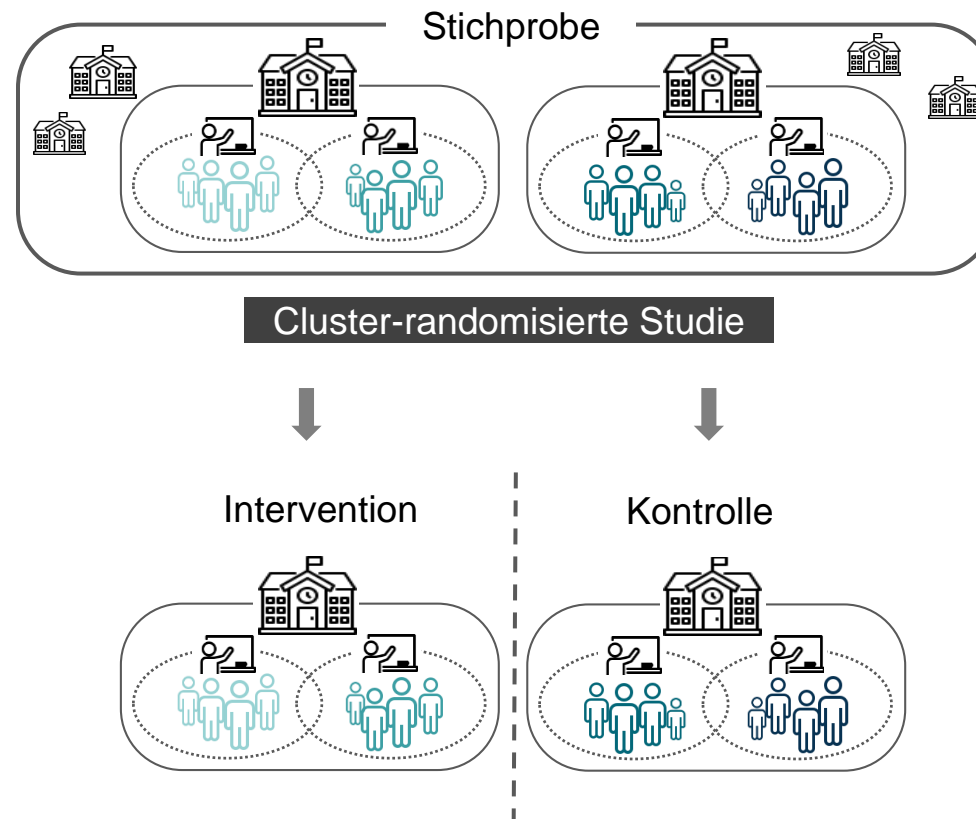
Teil II: Mehrebenen-Designs

Randomisierte Interventionsstudien

Ein-Ebenen-Design

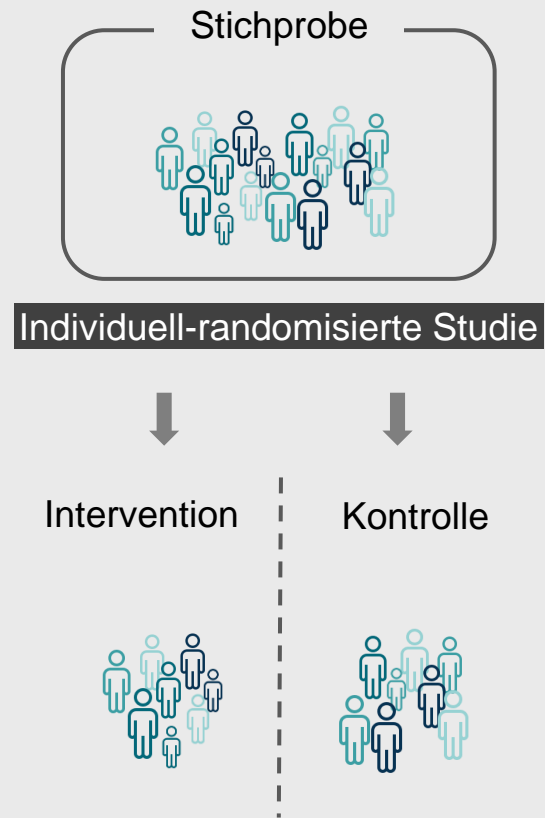


Mehrebenen-Design

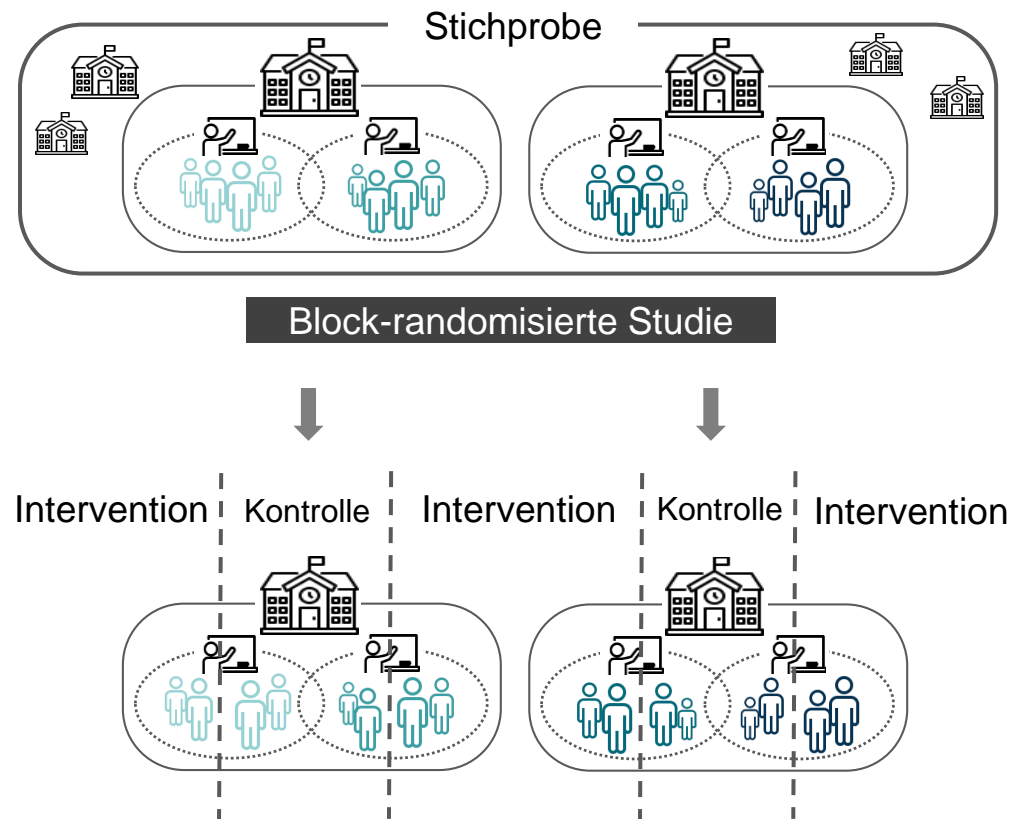


Randomisierte Interventionsstudien

Ein-Ebenen-Design



Mehrebenen-Design



STATISTISCHE POWER

Wahrscheinlichkeit, einen Effekt zu finden, wenn dieser in der Population existiert

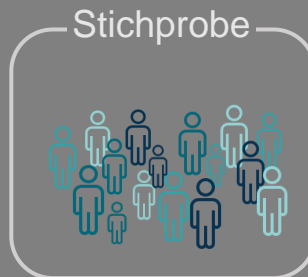
Designsensitivität

SCHÄTZGENAUIGKEIT

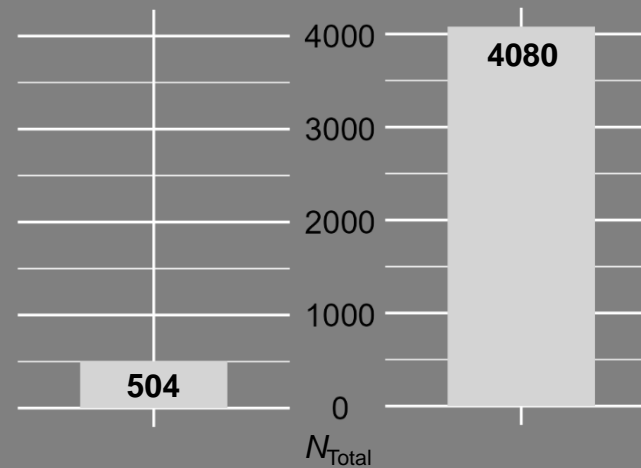
Standardfehler des Interventionseffekts

STICHPROBENGROÖE

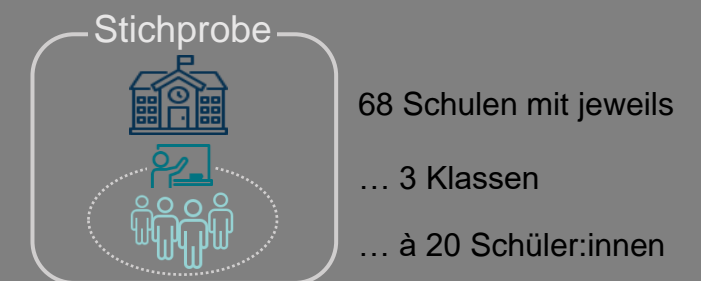
Individuell-randomisierte Studie



Beispiel: Matheleistung 4. Klasse
 $\delta = .25$ (Power 80%, $\alpha = .05$, zweiseitig)



Cluster-randomisierte Studie

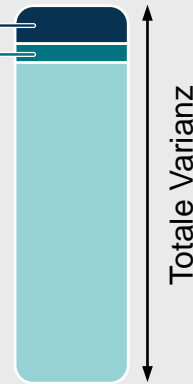


Designparameter

INTRAKLASSENKORRELATION (ICC; ρ)

Unterschiede zwischen Schulen $\rho_{L3} = .10$

Unterschiede zwischen Klassen $\rho_{L2} = .05$

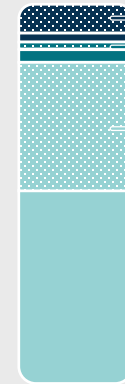


ERKLÄRTE VARIANZ DURCH KOVARIATEN (R^2)

R^2_{L3} Erklärte Varianz auf Schulebene

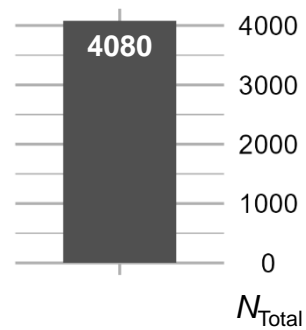
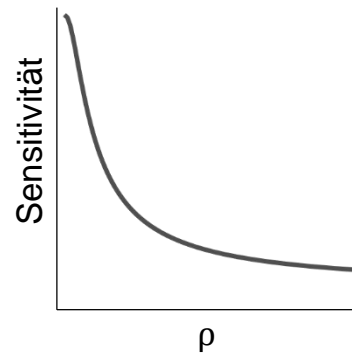
R^2_{L2} Erklärte Varianz auf Klassenebene

R^2_{L1} Erklärte Varianz auf Individualebene



Beispiel: Matheleistung 4. Klasse

$\delta = .25$ (Power 80%, $\alpha = .05$, zweiseitig)

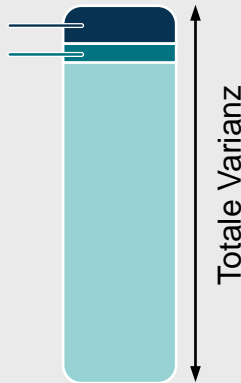


Designparameter

INTRAKLASSENKORRELATION (ICC; ρ)

Unterschiede zwischen Schulen $\rho_{L3} = .10$

Unterschiede zwischen Klassen $\rho_{L2} = .05$



ERKLÄRTE VARIANZ DURCH KOVARIATEN (R^2)

R_{L3}^2 Erklärte Varianz auf Schulebene

R_{L2}^2 Erklärte Varianz auf Klassenebene

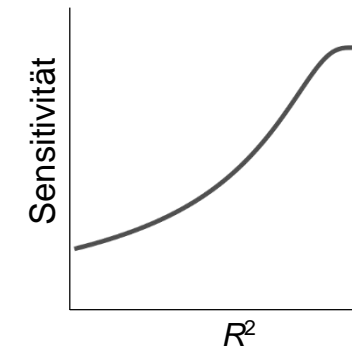
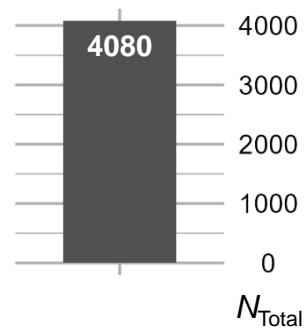
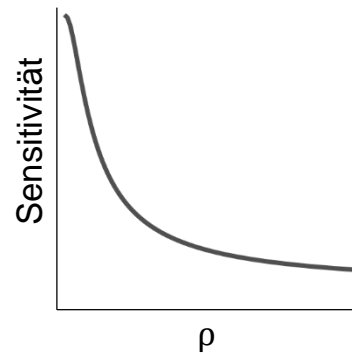
R_{L1}^2 Erklärte Varianz auf Individualebene



Ein-Ebenen-Design (über alle Schüler:innen)

R_T^2 = Erklärte Varianz total

Beispiel: Matheleistung 4. Klasse
 $\delta = .25$ (Power 80%, $\alpha = 05$, zweiseitig)

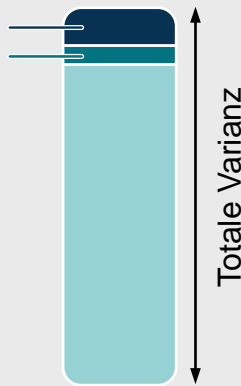


Designparameter

INTRAKLASSENKORRELATION (ICC; ρ)

Unterschiede zwischen Schulen $\rho_{L3} = .10$

Unterschiede zwischen Klassen $\rho_{L2} = .05$



ERKLÄRTE VARIANZ DURCH KOVARIATEN (R^2)

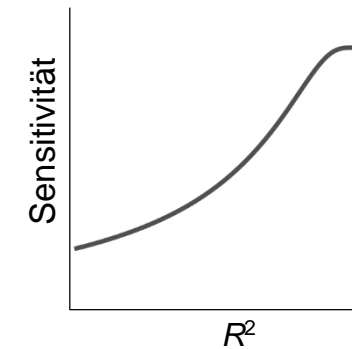
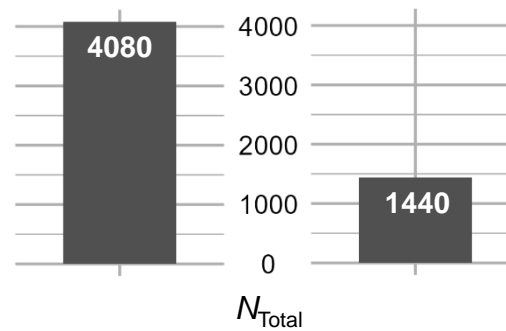
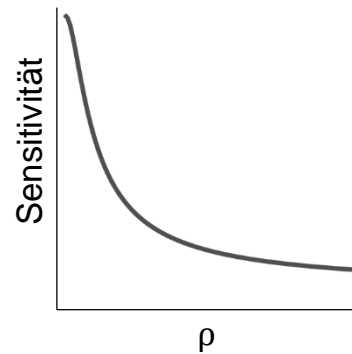
$R^2_{L3} = .76$ Erklärte Varianz auf Schulebene

$R^2_{L2} = .35$ Erklärte Varianz auf Klassenebene

$R^2_{L1} = .40$ Erklärte Varianz auf Individualebene

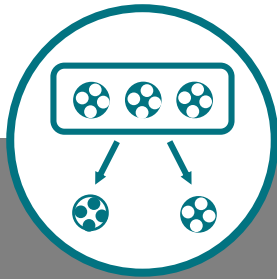


Beispiel: Matheleistung 4. Klasse
 $\delta = .25$ (Power 80%, $\alpha = 05$, zweiseitig)



<https://sophiestallasch.github.io/2024-workshop-GEBCF-power>

HANDS-ON!



Teil II: Mehrebenen-Designs

Poweranalyse für experimentelle Designs in R



Fragen & Diskussion

Referenzen

- American Psychological Association (Ed.). (2020). *Publication manual of the American Psychological Association* (7th ed.). American Psychological Association.
- Bausell, R. B., & Li, Y.-F. (2002). *Power analysis for experimental research: A practical guide for the biological, medical and social sciences*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511541933>
- Bloom, H. S. (1995). Minimum detectable effects: A simple way to report the statistical power of experimental designs. *Evaluation Review*, 19(5), 547–556. <https://doi.org/10.1177/0193841X9501900504>
- Bloom, H. S. (2005). Randomizing groups to evaluate place-based programs. In H. S. Bloom (Ed.), *Learning More From Social Experiments. Evolving analytic approaches* (pp. 115–172). Russell Sage Foundation.
- Bloom, H. S. (2006). *The core analytics of randomized experiments for social research*. MDRC Working Papers on Research Methodology. http://www.mdrc.org/sites/default/files/full_533.pdf
- Brunner, M., Keller, U., Wenger, M., Fischbach, A., & Lüdtke, O. (2018). Between-school variation in students' achievement, motivation, affect, and learning strategies: Results from 81 countries for planning group-randomized trials in education. *Journal of Research on Educational Effectiveness*, 11(3), 452–478. <https://doi.org/10.1080/19345747.2017.1375584>
- Brunner, M., Stallasch, S. E., & Lüdtke, O. (2023). Empirical benchmarks to interpret intervention effects on student achievement in elementary and secondary school: Meta-analytic results from Germany. *Journal of Research on Educational Effectiveness*, 1–39. <https://doi.org/10.1080/19345747.2023.2175753>
- Education Endowment Foundation. (2022). *Statistical analysis guidance for EEF evaluations*. <https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/evaluation-design/EEF-Analysis-Guidance-Website-Version-2022.14.11.pdf?v=1698086955>

Referenzen

- Erdfelder E. (2022). Teststärke. In M. A. Wirtz (Hrsg.), *Dorsch. Lexikon der Psychologie*. <https://dorsch.hogrefe.com/stichwort/teststaerke>
- Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science*, 9(6), 641–651. <https://doi.org/10.1177/1745691614551642>
- German Research Foundation (Ed.). (2022). *Proposal preparation instructions. Project proposals*. https://www.dfg.de/formulare/54_01/54_01_en.pdf
- Halpern, S. D., Karlawish, J. H. T., & Berlin, J. A. (2002). The continuing unethical conduct of underpowered clinical trials. *JAMA*, 288(3), 358. <https://doi.org/10.1001/jama.288.3.358>
- Hedberg, E. C. (2018). *Introduction to power analysis: Two-group studies*. SAGE Publications, Inc. <https://doi.org/10.4135/9781506343105>
- Institute of Education Sciences. (2023). *Education research grants program. Request for applications*. (ALN: 84.305A). https://ies.ed.gov/funding/pdf/2021_84305A.pdf
- Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, 2(8), e124. <https://doi.org/10.1371/journal.pmed.0020124>
- Ioannidis, J. P. A. (2008). Why Most Discovered True Associations Are Inflated. *Epidemiology*, 19(5), 640–648. <https://doi.org/10.1097/EDE.0b013e31818131e7>
- Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, 8(1), 1–28. <https://doi.org/10.1525/collabra.33267>
- Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, 55(3), 187–193. <https://doi.org/10.1198/000313001317098149>
- Lortie-Forgues, H., & Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educational Researcher*, 48(3), 158–166. <https://doi.org/10.3102/0013189X19832850>

Referenzen

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.
<https://doi.org/10.1126/science.aac4716>

Schochet, P. Z. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics*, 33(1), 62–87. <https://doi.org/10.3102/1076998607302714>

Schönbrodt, F.D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, 47, 609–612.
<http://dx.doi.org/10.1016/j.jrp.2013.05.009>

Sims, S., Anders, J., Inglis, M., & Lortie-Forgues, H. (2022). Quantifying “promising trials bias” in randomized controlled trials in education. *Journal of Research on Educational Effectiveness*, 16(4), 1–18. <https://doi.org/10.1080/19345747.2022.2090470>