

Vorlesung: Empirische Bildungsforschung

Sitzung 7: Effekte der Klassengröße

Dr. Edgar Treischl
Last update: 2022-04-25

This presentation is licensed under a CC-BY-NC 4.0 license. You may copy, distribute, and use the slides in your own work, as long as you give attribution to the original author on each slide that you use. Commercial use of the contents of these slides is not allowed.



Ablauf und Lernziele

1. Stellschrauben im Bildungssystem (u.a. die Klassengröße)
2. Effekt der Klassengröße auf die individuelle Schulleistung
3. Damit verbunden: Vertiefung Kausalität
4. Experimentelle/Quasi-experimentelle Befunde

1. Stellschrauben im Bildungssystem

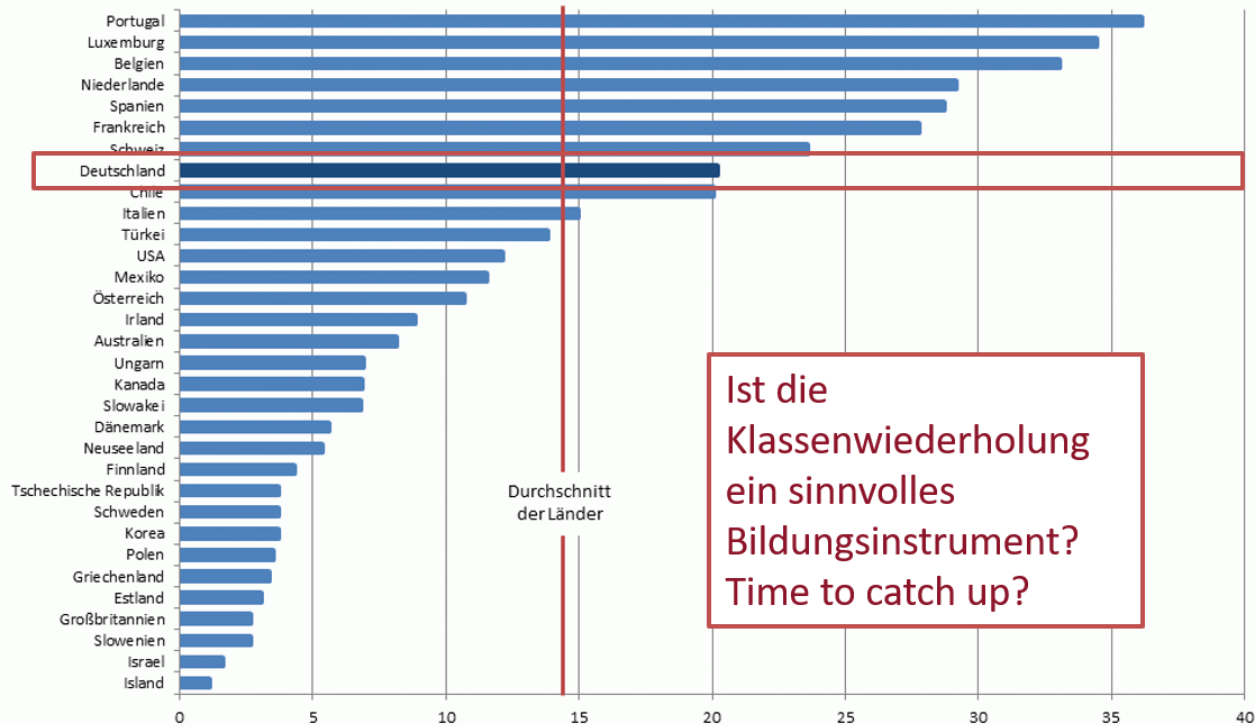
School Tracking



Created with mapchart.net ©

Klassenwiederholung

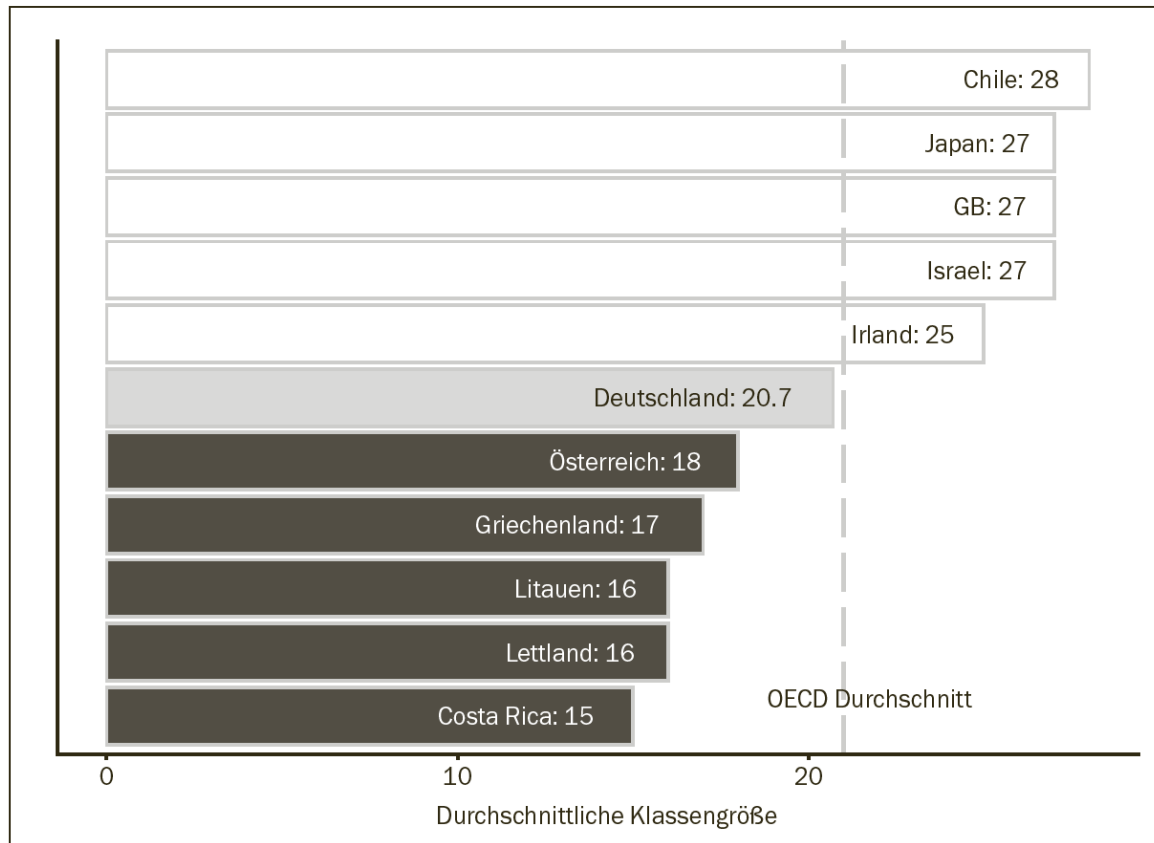
OECD: Anteil der 15-Jährigen, die mindestens eine Klasse wiederholt haben
in Prozent, 2012



Ist die
Klassenwiederholung
ein sinnvolles
Bildungsinstrument?
Time to catch up?

Quelle: PISA 2012

Klassengröße



Anmerkung: Dargestellt ist die durchschnittliche Anzahl an Grundschulkindern pro Schulklassen im Jahr 2018 basierend auf den Angaben der OECD Statistics Plattform (2018).

2. Kausaler Effekt der Klassengröße?

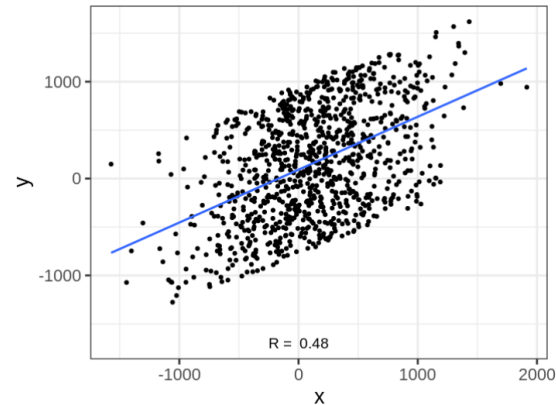
Power Analysis ▶ Start  Effect Size  Power  Estimate  Low Power

Why do we need power analysis?

Students presenting a study in seminar and other courses often criticize if a study relies on a small sample. What does this mean? Can you tell me in which case a sample is too small? And should me always try to maximize the sample size? Power analysis tells which sample size is needed to have the statistical power to detect an effect.

In order to run a power analysis, we have to provide several information: First, we have to make an assumption about the effect we want to estimate. On the left side you can see scatter plot with simulated data. Go on and adjust the effect size of X on Y. Like the correlation between X and Y, we have to assume how both variables are related. Second, holding other aspects like effect size constant, we have to specify how much statistical power do we want to detect an effect of X on Y.

This app illustrates both aspects and shows you how a power analysis helps researchers to determine the "right" sample size and how you can get an idea whether a study has (not) enough power to detect an effect.



Correlation:

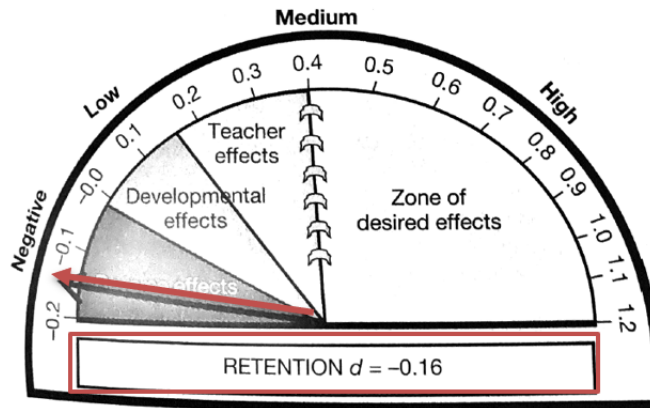


Meta-Analysen in der EBF: John Hattie



Effektgröße

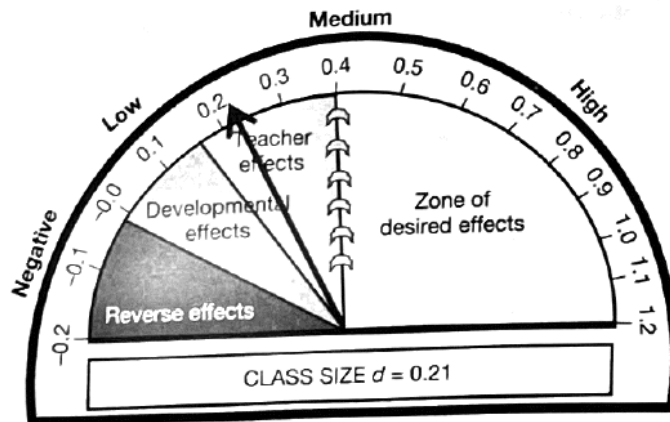
- Hattie fasst viele Studien und Meta-Analysen für den Bildungssektor zusammen: Hier das **Hauptergebnis für die Klassenwiederholung**:



KEY	
Standard error	na
Rank	136th
Number of meta-analyses	7
Number of studies	207
Number of effects	2,675
Number of people (2)	13,938

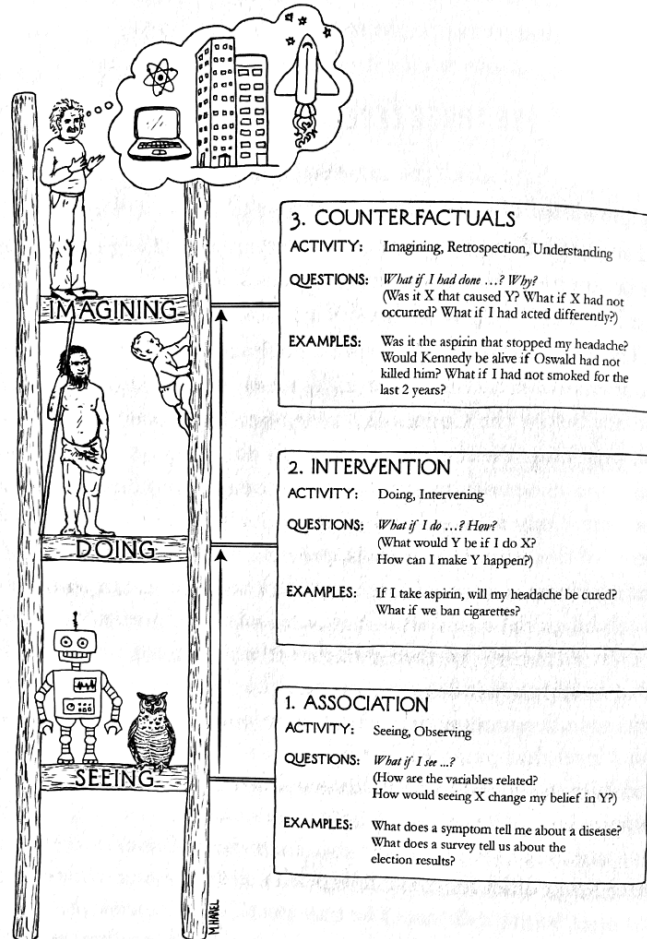
Effektstärken sind abhängig von der Skalierung der Variablen. Zur Vergleichbarkeit können Effekte standardisiert werden. Beispiel Cohens D: Cohen teilt die Effektgröße d in drei Stufen ein: Kleine Effekte ($d = 0,2$); mittlere Effekte ($d = 0,5$); sowie starke Effekte ($d = 0,8$).

Ein Blick in Hattie (Visible Learning) zeigt, bisherige Forschung kommt zu dem Ergebnis, dass die Klassengröße einen sehr kleinen Effekt ($d=0,2$) auf die individuelle Schulleistung hat.



KEY	
Standard error	
Rank	na
Number of meta-analyses	106th
Number of studies	3
Number of effects	96
Number of people (x)	785
	550,339

Ladder of Causation von Judea Pearl (The Book of Why)

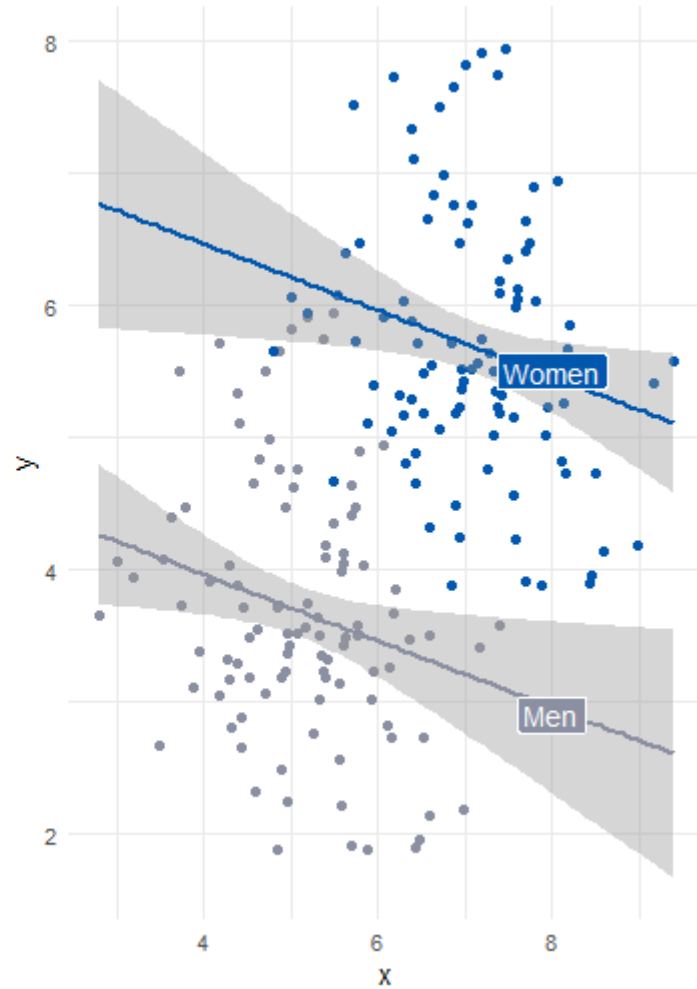
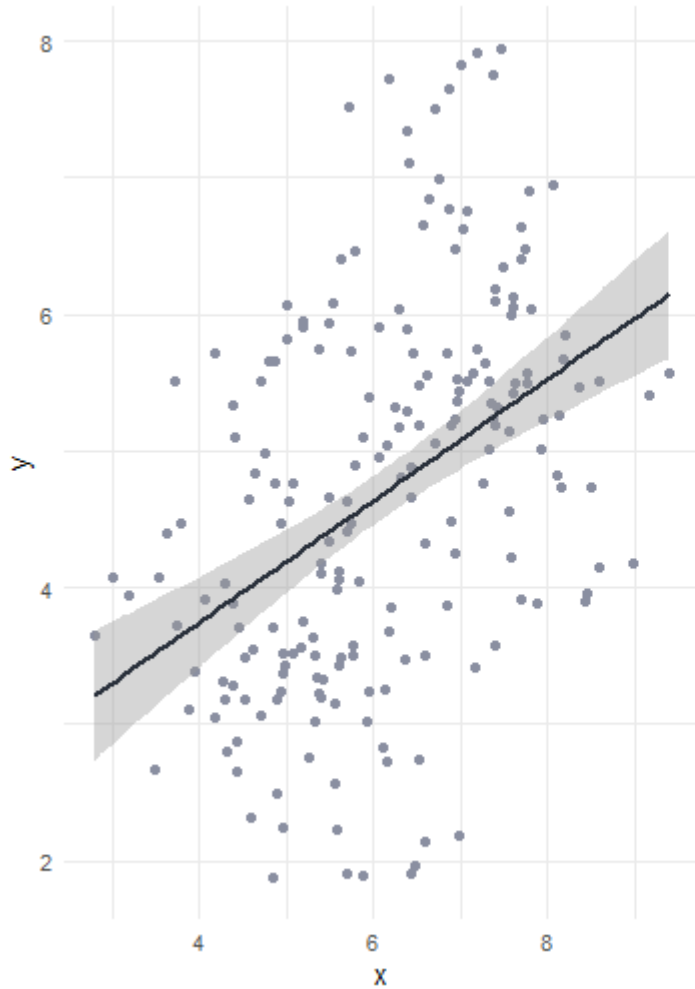


Fallstricke von Querschnittsanalysen

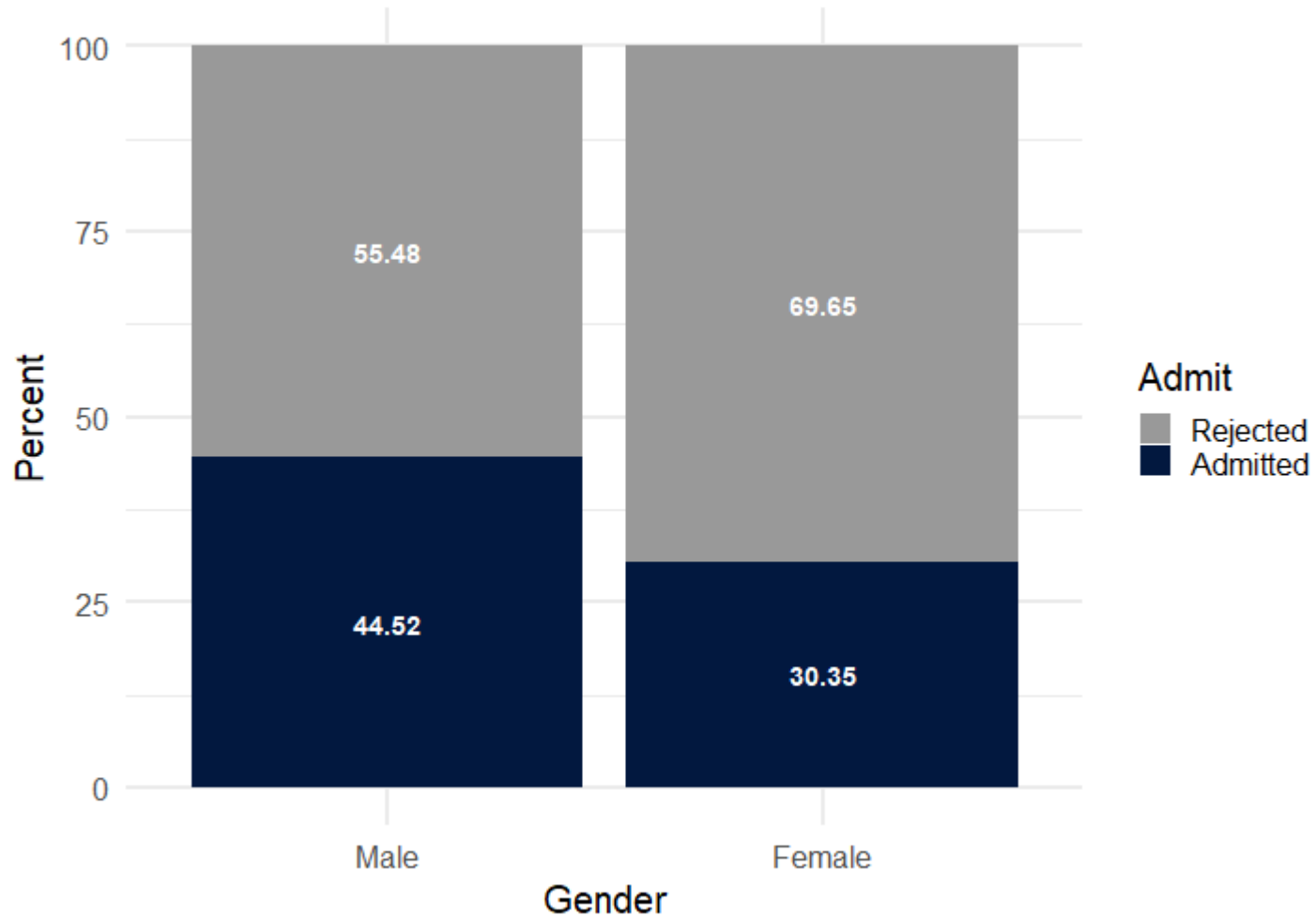
(Siehe Treischl & Wolbring 2020: 111-119)

- Zeitliche Abfolge von Ursache und Wirkung
- Reifung, intervenierende Ereignisse und das aktuelle Zeitgeschehen
- Selbst- und Fremdselektion
- Drittvariablenkontrolle

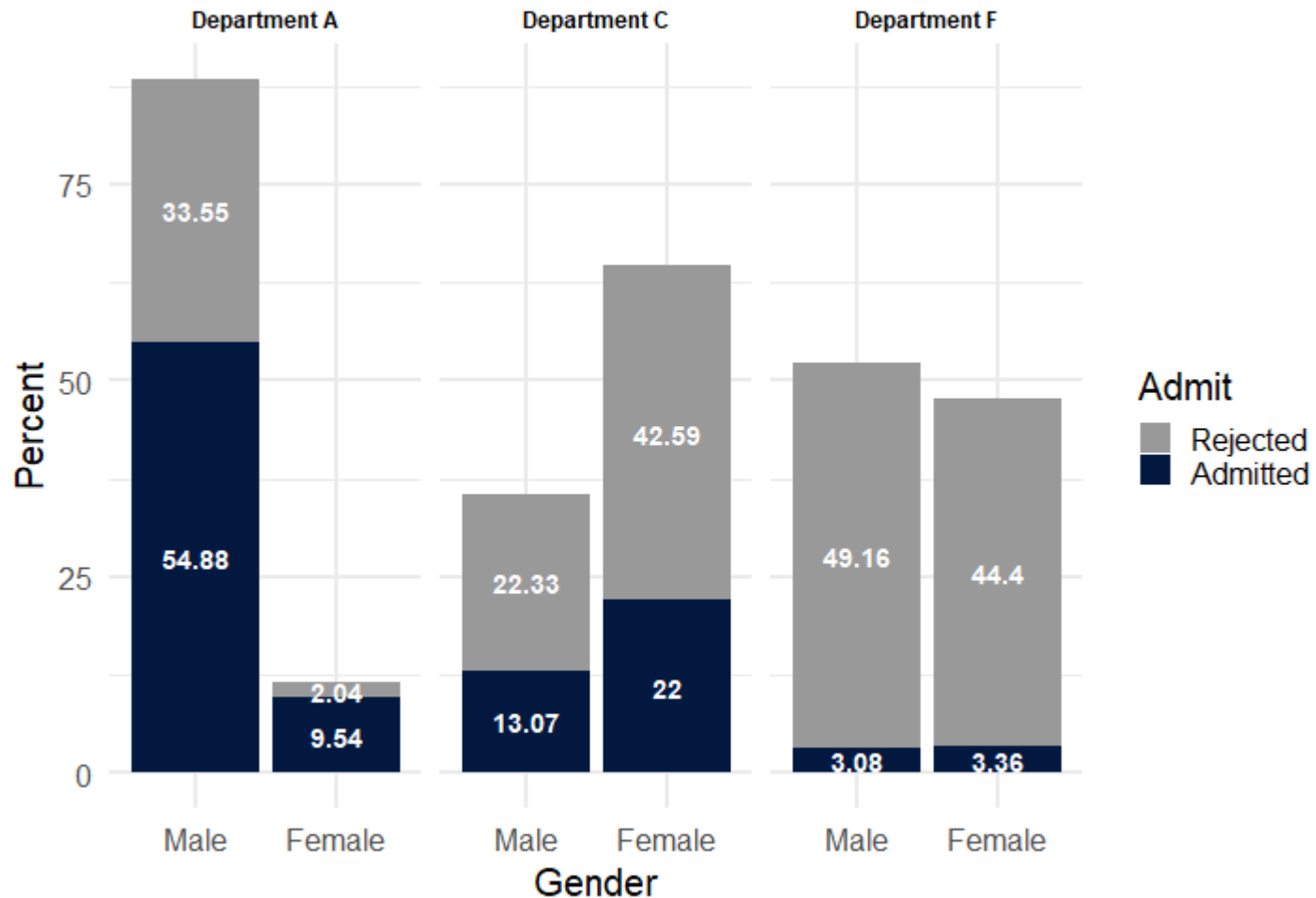
Simpson's Paradox



Beispiel: Zulassung Berkeley



Beispiel: Zulassung Berkeley II



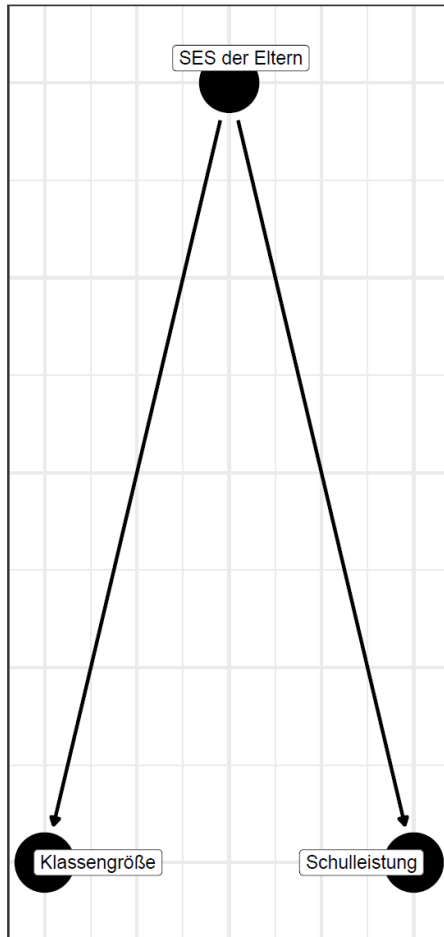
3. Vertiefung Kausalität

- mit Hilfe der Regression: Berechnung eines Effekts durch Kontrolle beobachteter Variablen, d.h. Kontrolle aller Variablen die sowohl den Treatmentstatus beeinflussen und einen Einfluss auf die abhängige Variable haben
- Hoffnung: Nach Kontrolle der Variablen ist die Zuweisung des Treatmentstatus Klassengröße „so gut wie zufällig“ (CIA)

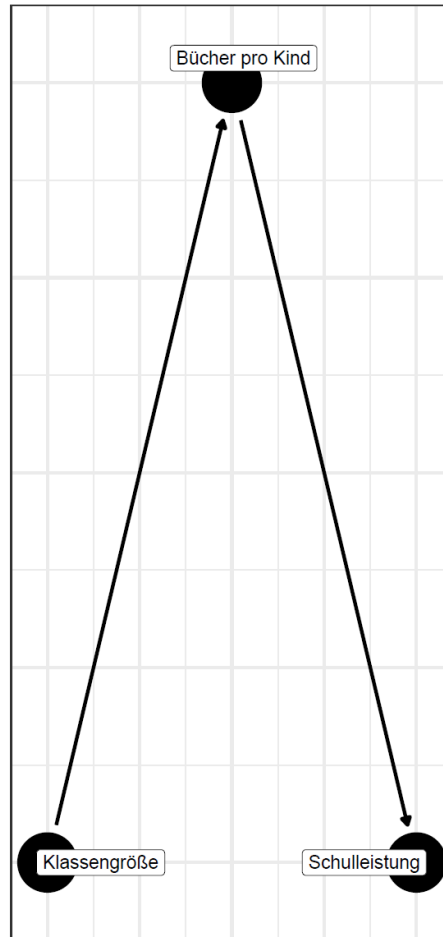
Merke: Konditionale Unabhängigkeit (conditional independence assumption, CIA): Bei der Regression muss die Annahme getroffen werden, dass nach der Kontrolle der anderen unabhängigen Variablen die Verteilung der Einheiten über die Treatment- und Kontrollgruppe in Hinblick auf die abhängige Variable so gut wie zufällig erfolgt.

Kausale Struktur

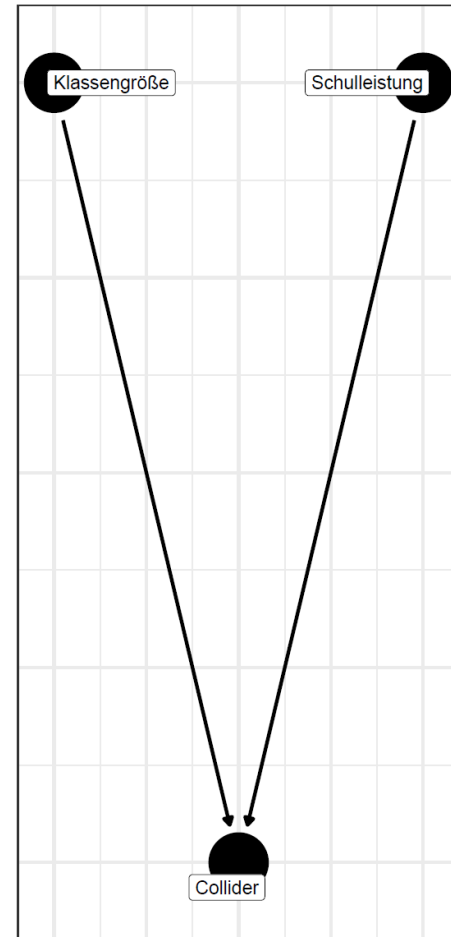
The Confounder



The Mediator



The Collider



Simulation von Daten mit R

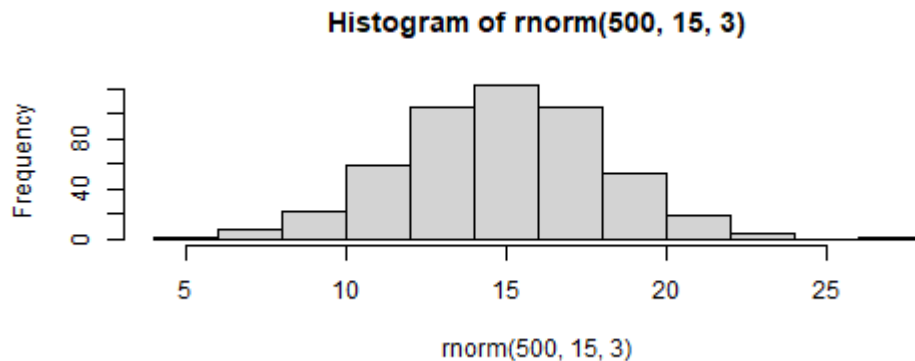
```
#rep: repeat (x, n)  
rep(0, 5)
```

```
## [1] 0 0 0 0 0
```

```
#rnorm: Normalverteilung (n, mean, sd)  
rnorm(5, 15, 3)
```

```
## [1] 16.228206 20.066620 19.759765 14.007277 8.144293
```

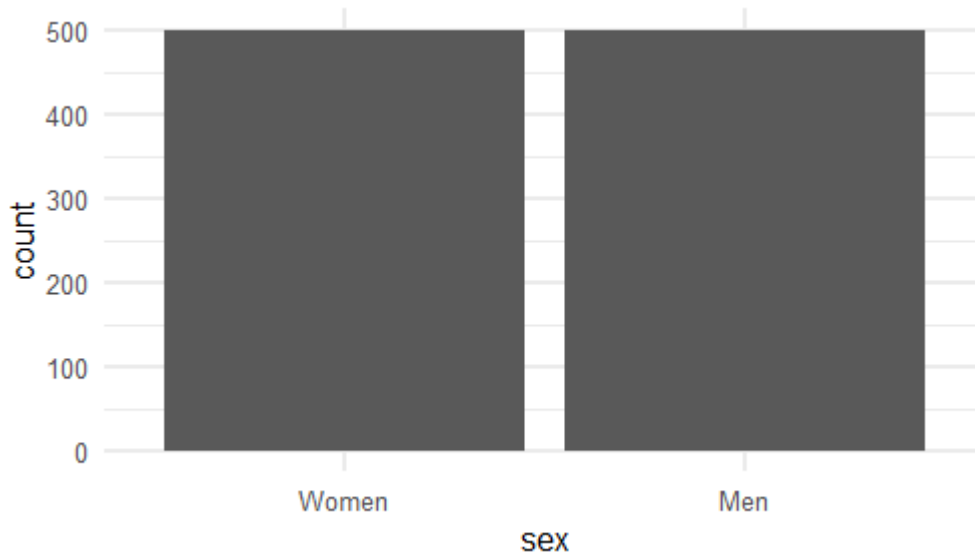
```
hist(rnorm(500, 15, 3))
```



Setup the confounder: Sex

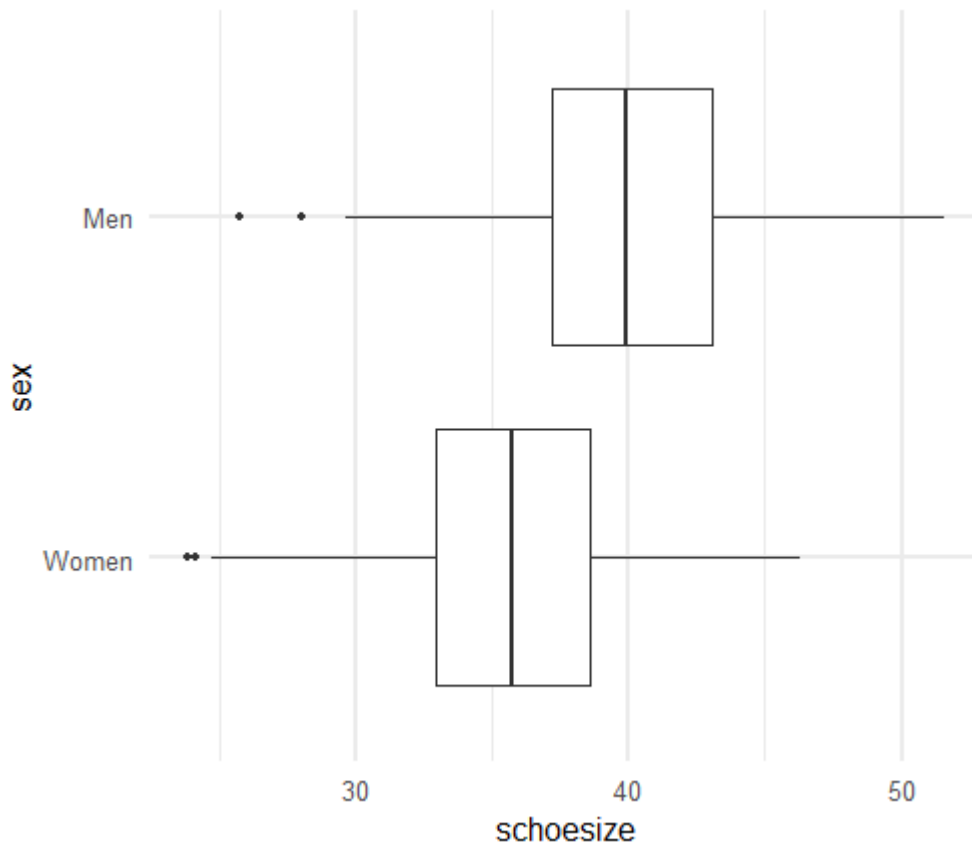
Shoesize \leftarrow Sex \rightarrow Income

```
#Generate the sex variable: 500 Men and 500 Women  
sex <- rep(c(0,1), 500)  
sex.label <- c("Women", "Men")
```



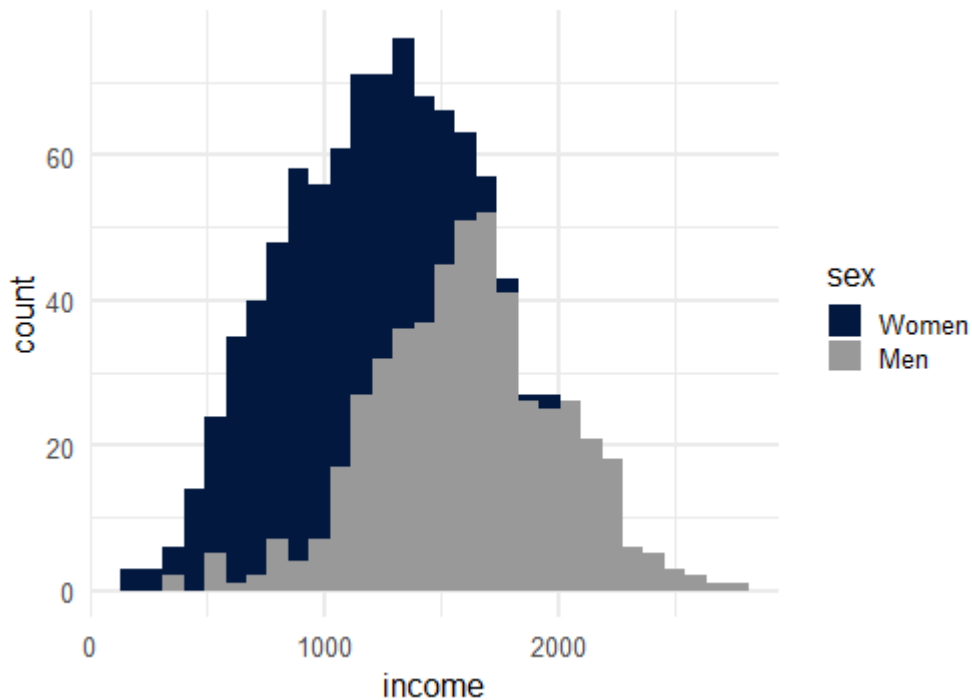
Setup the confounder: Shoe size

```
#Simulate shoe sizes, bigger ones for men!  
shoesize <- rnorm(1000, 36, 4) + sex*rnorm(1000, 4, 2)
```



Setup the confounder: Income

```
#Simulate income that depends on the confounder  
#But, by definition, income does not depend on X!  
income <- rnorm(1000, 1000, 333) + sex*rnorm(1000, 600, 250)
```



Results: Shoesize → Income (Confounderbias)

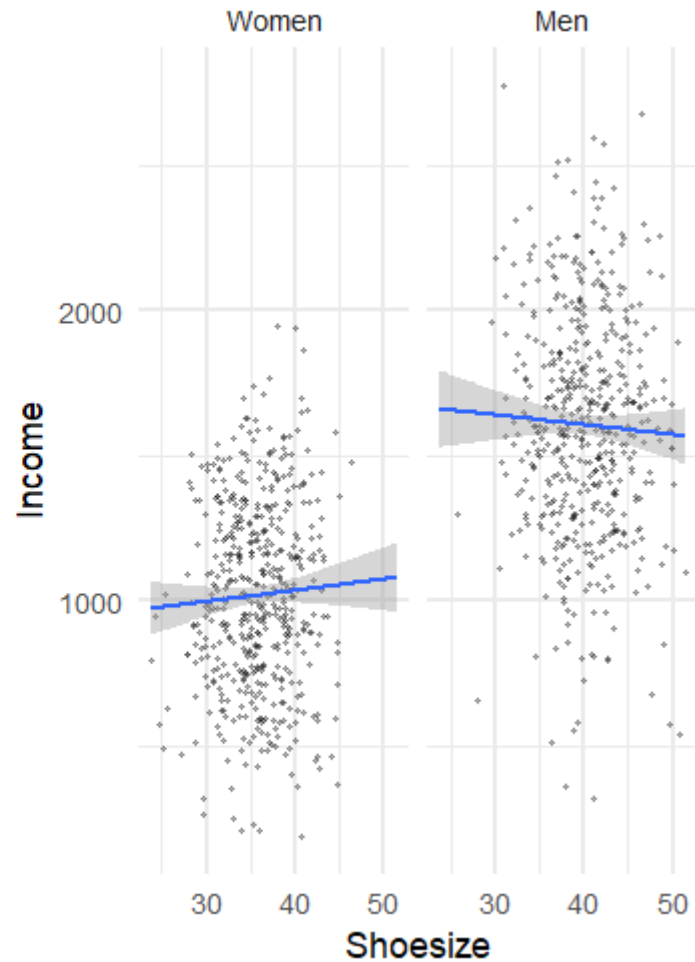
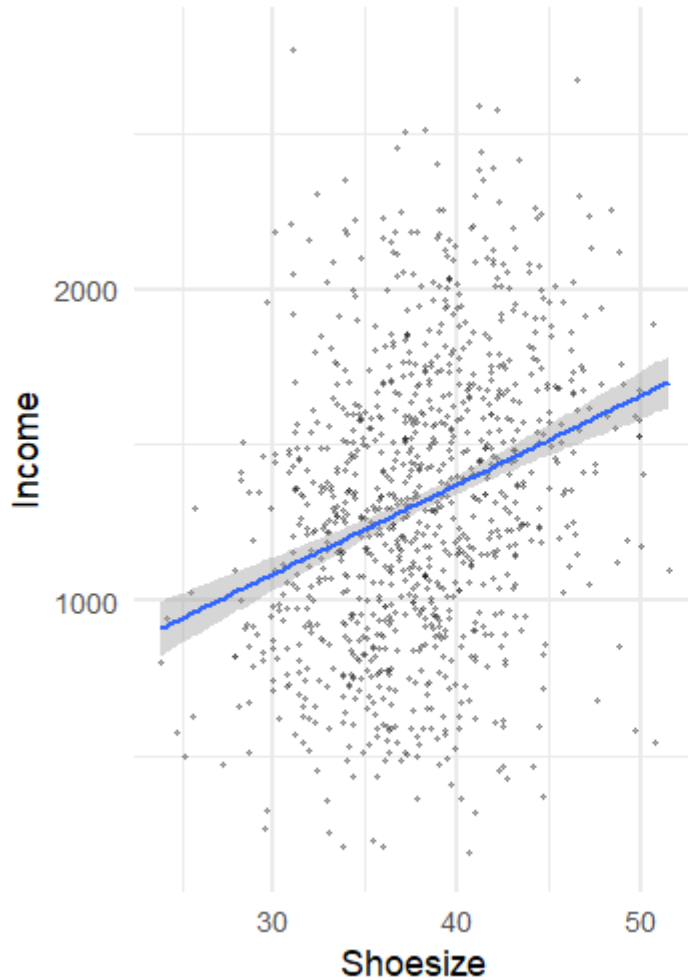
```
m1 <- lm(income ~ schoesize, data = data)
summary(m1)
```

```
##
## Call:
## lm(formula = income ~ schoesize, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1203.08  -312.17   -3.07   297.19  1661.27
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  224.729    113.959   1.972  0.0489 *
## schoesize    28.702     2.983   9.620 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 445.1 on 998 degrees of freedom
## Multiple R-squared:  0.08486,    Adjusted R-squared:  0.08395
## F-statistic: 92.55 on 1 and 998 DF,  p-value: < 2.2e-16
```

Results: Shoesize \leftarrow Sex \rightarrow Income

	Model 1	Model 2
(Intercept)	224.73 *	1024.43 ***
	(113.96)	(99.00)
schoesize	28.70 ***	-0.13
	(2.98)	(2.73)
sexMen		586.50 ***
		(25.80)
N	1000	1000
R2	0.08	0.40
*** p < 0.001; ** p < 0.01; * p < 0.05.		

Konfundierung grafisch



Setup the mediator: $X \rightarrow \text{Mediator} \rightarrow Y$

```
#X Variable
x <- rnorm(1000, 10, 3)

#Some random noise
random1<- runif(1000,min=min(x),max=max(x))

#Important Step 1: The mediator = 0,5*x + 0,5*random noise
mediator <- x*0.5+ random1*0.5

#Correct?
cor(x, mediator)
```

```
## [1] 0.4637501
```

```
#A second random variable
random2<- runif(1000,min=min(mediator),max=max(mediator))

#Important Step 2: y = 0,5*mediator + 0,5*random variable
y <- 0.5*mediator + 0.5*random2
```


Results: $X \rightarrow \text{Mediator} \rightarrow Y$

	Model 1	Model 2
(Intercept)	6.98 *** (0.32)	4.48 *** (0.31)
x	0.27 *** (0.03)	0.02 (0.03)
mediator		0.51 *** (0.03)
N	1000	1000
R2	0.07	0.30
*** p < 0.001; ** p < 0.01; * p < 0.05.		

Setup the collider

$X \rightarrow \text{Collider} \leftarrow Y$

```
#Some X  
x <- 1 + rnorm(1000, 5, 3)  
#Some Y  
y <- rnorm(1000, 10, 1)  
  
#The collider  
collider <- 1*x + 1*y + rnorm(1000, 1, 1)
```

Results: $X \rightarrow \text{Collider} \leftarrow Y$

	Model 1	Model 2
(Intercept)	10.06 *** (0.07)	4.38 *** (0.18)
x	-0.01 (0.01)	-0.52 *** (0.02)
collider		0.51 *** (0.02)
N	1000	1000
R2	0.00	0.52
*** p < 0.001; ** p < 0.01; * p < 0.05.		

Remember:

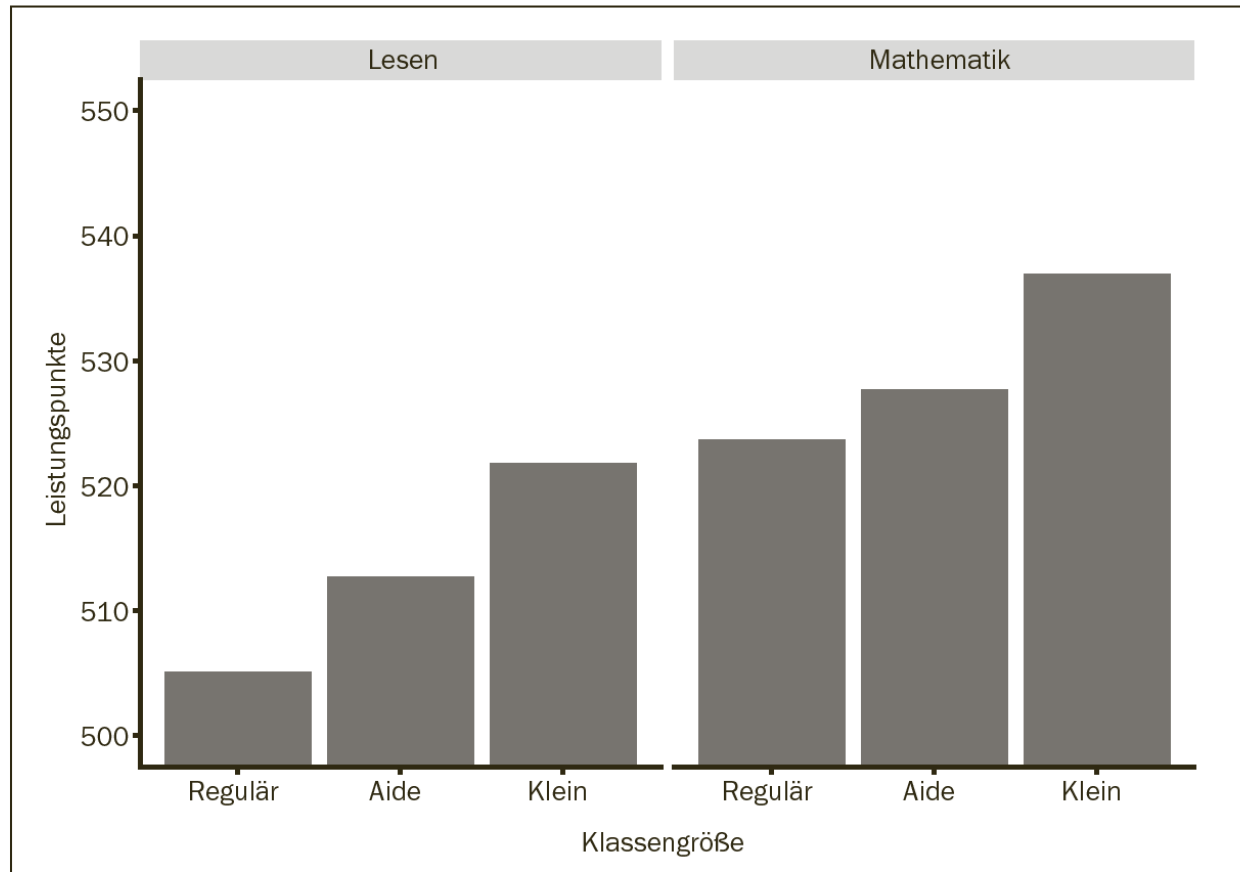
- Die Nicht-Kontrolle eines *Confounders* kann eine verzerrte Assoziation zwischen X und Y hervorrufen
- Die Kontrolle eines *Colliders* kann eine verzerrte Assoziation zwischen X und Y hervorrufen
- Die Kontrolle eines *Mediators* reduziert die wahre Effektgröße von X auf Y

4. Experimentelle und Quasi-experimentelle Befunde

STAR Experiment

- Feldexperiment in den USA
- STAR (Student-Teacher Achievement Ratio) Experiment/Project
- Randomisierung von LK und SuS
- 11.600 SuS in drei Versuchsgruppen: Regulär (22-25 SuS), Aide (22-25 SuS + Unterstützungskraft), Klein (13-17 SuS)
- Hauptbefund: Substantieller Treatmenteffekt

Ergebnisse des STAR Experiments



Quasi-experimentelle Studien: Kurzer Auszug aus Bach und Sievert 2018

Quasi-experimentelle Studien nutzen eine Form der *exogenen Variation* zur Schätzung des kausalen Effekts der Klassengröße.

D.h. es wird nicht der Effekt der Variation der Klassengröße untersucht, sondern bei Bach und Sievert 2018 wird bspw. die durchschnittliche Klassengrößen aufeinanderfolgender Jahrgänge innerhalb von Schulen untersucht. Dazu schreiben sie:

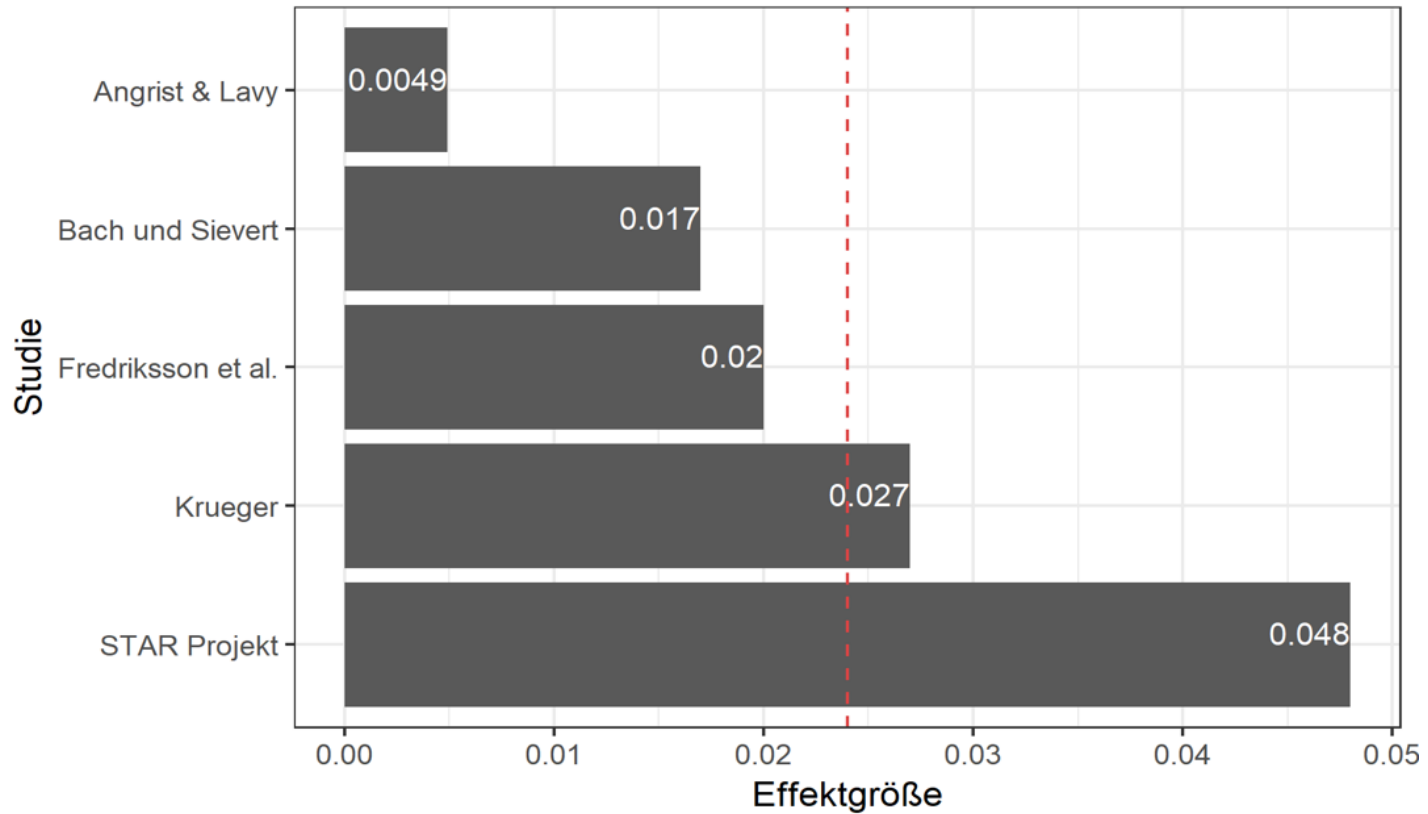
“Dahinter steckt die Idee, dass die durchschnittliche Klassengröße direkt von der Gesamtgröße des Jahrgangs einer Schule abhängt, letzterer aber rein zufällig bedingt manchmal größer und manchmal kleiner ausfällt. Der Hauptgrund hierfür ist, dass Geburten innerhalb von Schuleinzugsgebieten nicht gleichmäßig über das Jahr verteilt sind.”

Quasi-experimentelle Studien II

“Folglich wird es in manchen Jahren mehr Anmeldungen an einer Schule geben als in anderen. Kinder, die vergleichsweise kleinen Jahrgängen angehören, werden bei gegebener Anzahl an Parallelklassen so während ihrer gesamten Grundschulzeit in kleineren Klassenverbänden unterrichtet als andere Kinder – ohne dass SchulleiterInnen, LehrerInnen oder Eltern dies beeinflussen können.”

“Bei anderen quasi-experimentellen Studien wurde andere, exogene Variationen identifiziert. Allen gemeinsam ist jedoch die Idee der exogenen Variation, also eine Variation die nicht durch konfundierenden Faktoren beeinflusst wird und somit auch das Problem der Selbstselektion im Idealfall mindert.”

Befunde quasi-experimenteller Studien



Zusammenfassung

- Stellschrauben im Bildungssystem
- Beobachtungsdaten und Querschnittsanalysen zur Evaluation von Maßnahmen
- Fallstricke bei Querschnittsanalyse
- Drittvariablenkontrolle: Confounder, Mediator, Collider
- Kausaler Effekt der Klassengröße auf die individuelle Leistung (eher gering?)