

Replicerbarhed og replikation

Erik Gahner Larsen

Kausalanalyse i offentlig politik

Gæsteforelæsning

- ▶ Evaluering af 'Evalueringen af offentlige politikker'
- ▶ Feedback?

Eksamen og vejledning

- ▶ Dato for aflevering: 2. juni, klokken 12.00
- ▶ Vejledning I: Mandag 9. maj efter klokken 13
- ▶ Vejledning II: Mandag 23. maj

Evaluering

- ▶ Semesterets undervisningsevaluering
- ▶ Evalueringsperioden starter mandag d. 25. april (i dag)
- ▶ Løber frem til og med søndag d. 8. maj

Add on I: AGORA

- ▶ Tid: Onsdag d. 4. maj
 - ▶ Hele dagen fra kl. 08.45 (al undervisning er aflyst)
- ▶ Sted: SDU O100 - ved hovedindgangen
- ▶ Yderligere oplysninger på
<https://www.facebook.com/AGORASDU/>

Add on II: Introduktion til R

- ▶ Tirsdag d. 21. juni

Dagsorden

- ▶ Åben videnskab (*open science*)
- ▶ Reproducerbarhed
- ▶ Replikation
- ▶ Workflow

POLITICAL SCIENCE

When contact changes minds: An experiment on transmission of support for gay equality

Michael J. LaCour¹ and Donald P. Green²

Can a single conversation change minds on divisive social issues, such as same-sex marriage? A randomized placebo-controlled trial assessed whether gay ($n = 22$) or straight ($n = 19$) messengers were effective at encouraging voters ($n = 972$) to support same-sex marriage and whether attitude change persisted and spread to others in voters' social networks. The results, measured by an unrelated panel survey, show that both gay and straight canvassers produced large effects initially, but only gay canvassers' effects persisted in 3-week, 6-week, and 9-month follow-ups. We also find strong evidence of within-household transmission of opinion change, but only in the wake of conversations with gay canvassers. Contact with gay canvassers further caused substantial change in the ratings of gay men and lesbians more generally. These large, persistent, and contagious effects were confirmed by a follow-up experiment. Contact with minorities coupled with discussion of issues pertinent to them is capable of producing a cascade of opinion change.

Hvor stort er problemet? (John et al. 2012)

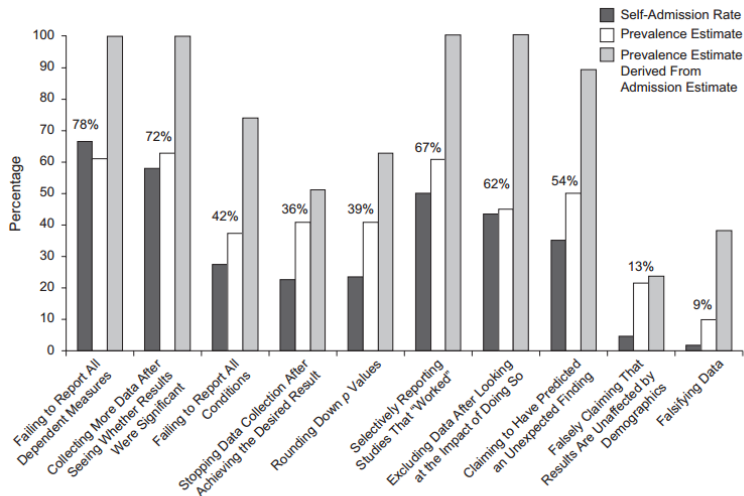


Fig. 1. Results of the Bayesian-truth-serum condition in the main study. For each of the 10 items, the graph shows the self-admission rate, prevalence estimate, prevalence estimate derived from the admission estimate (i.e., self-admission rate/admission estimate), and geometric mean of these three percentages (numbers above the bars). See Table 1 for the complete text of the items.

Udfordring

- ▶ Meget forskning kan ikke replikeres
 - ▶ Nogle mener sågar, at det meste forskning ikke kan replikeres
- ▶ Meget forskning kan ikke reproduceres
- ▶ Hvad er forskellen?

Reproduktion og replikation

- ▶ Reproducere: at skabe de *samme* resultater med det *samme* data
- ▶ Replikere: at gennemføre en undersøgelse med *andre* data (evt. nye data)
- ▶ Reproduktion og replikation bliver dog ofte anvendt som var de indbyrdes ombyttelige
 - ▶ Væsentlige forskelle på de to
- ▶ Herrnson (1995, 452): "Replication repeats an empirical study in its entirety, including independent data collection. It enables a researcher to comment on whether data used in an original study were collected properly or whether generalizations supported under one set of conditions are also supported under others. Replications increase the amount of information for an empirical research question and increase the level of confidence for a set of empirical generalizations."

Videnskab (lidt teori om viden, epistemologi)

- ▶ Videnskab er en kollektiv, åben praksis
 - ▶ Ikke individuel magi
- ▶ Uafhængigt af hvordan vi måler verden (observationer, interviews, etnografi etc.), skal andre kunne se, hvordan denne viden er indsamlet
- ▶ Videnskabelige resultater skal kunne reproduceres af andre
- ▶ To typer af transparens
 - ▶ Produktionstransparens
 - ▶ Analytisk transparens

Reproducerbarhedskontinuum (Peng 2011)

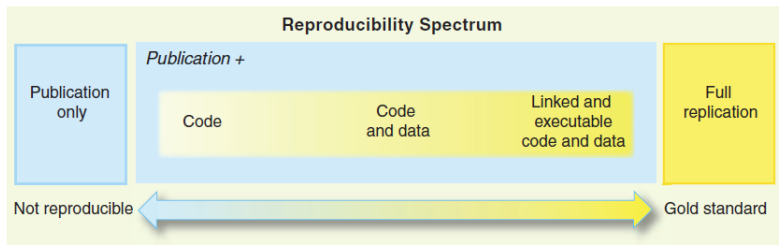


Fig. 1. The spectrum of reproducibility.

Publikationsbias

- ▶ Kun positive resultater bliver publiceret
 - ▶ 'file drawer problem' (Rosenthal 1979)
- ▶ Forskere bruger ensidede hypotesetests for at få signifikante resultater (Gerber et al. 2010)
- ▶ Implikation: Forskere skriver ikke deres negative fund op
 - ▶ Forkert viden/forkerte opfattelser bliver sjældent korrigeret

Publikationsbias: Hvorfor? (Franco et al. 2014)

Table 3. Cross-tabulation between statistical results of TESS studies and their publication status (column percentages reported). Pearson χ^2 test of independence: $\chi^2(6) = 80.3$, $P < 0.001$.

	Null	Mixed	Strong
Not written	64.6%	12.2%	4.4%
Written but not published	14.6	39.0	34.1
Published (non-top-tier)	10.4	37.8	38.5
Published (top-tier)	10.4	11.0	23.1
Total	100.0	100.0	100.0

Publikationsbias: Hvorfor? (Franco et al. 2015, 2016)

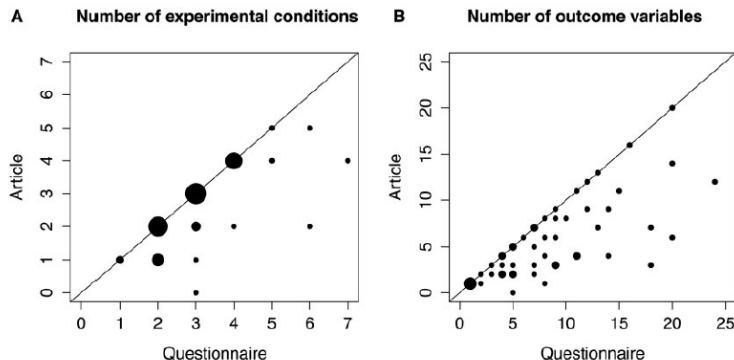


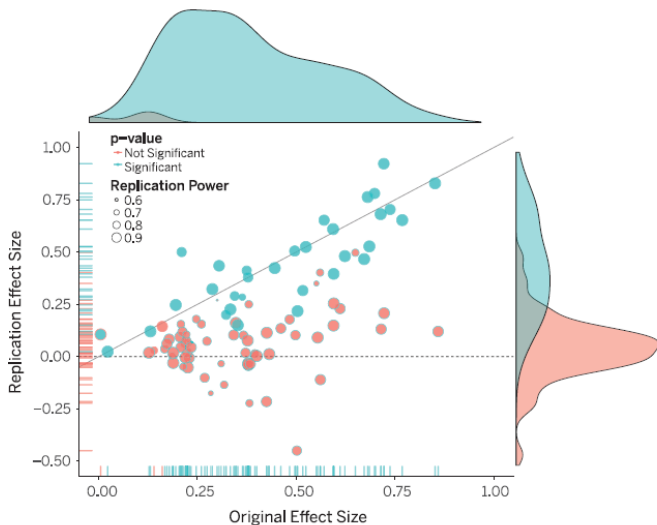
Fig. 1 Comparing design features in questionnaires and published results in TESS studies.

Note. Point size in each panel is proportional to the number of studies with a particular questionnaire-article value pair.

Hvorfor er det vigtigt?

- ▶ Meget forskning kan ikke replikeres
 - ▶ Igen: 'False positives' (Ioannidis 2005)
 - ▶ Tidsskrifter publicerer primært artikler der bekræfter en teori
- ▶ Penge brugt på forskning, der ikke kan replikeres
 - ▶ Alene US\$28,000,000,000 inden for præklinisk forskning i USA (Freedman et al. 2015)
- ▶ Et af de definerende karakteristika ved videnskab er, at det kan reproduceres.
 - ▶ Reproducerbar forskning: pleonasme
 - ▶ "the only way to understand and evaluate an empirical analysis is to know the exact process by which the data were generated and the analysis produced" (King 1995)

Hvorfor er det vigtigt? (Open Science Collaboration 2015)



Original study effect size versus replication effect size (correlation coefficients).

Hvorfor er det vigtigt i offentlig politik?

- ▶ Reproducerbarhed

- ▶ Vi skal vide hvordan studier kommer frem til de konklusioner, de gør
- ▶ Især for kritisk at kunne evaluere validiteten
 - ▶ Særinteresser

- ▶ Replikationer

- ▶ Svært at gennemføre replikationer
- ▶ Der er brug for gode replikationsstudier
- ▶ Ingen grund til at tro, at succesraten for replikationer i studiet af offentlig politik vil være højere end i psykologisk forskning

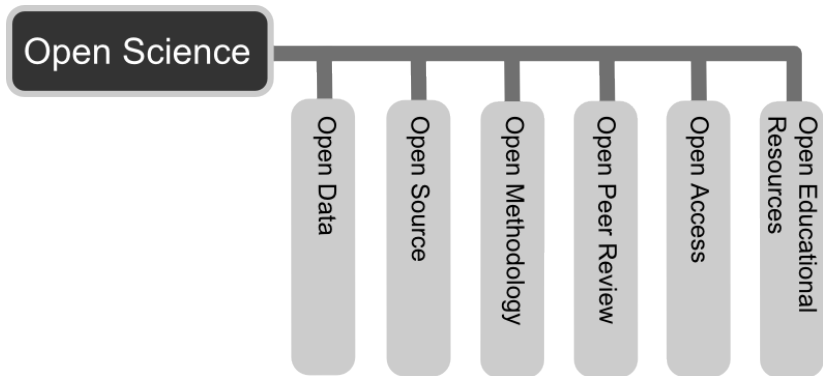
Løsninger på 'replikationskrisen' (Ioannidis 2014)

- ▶ Større samarbejde på tværs af forskningsgrupper
- ▶ Replikationskultur
- ▶ Preregistrering
- ▶ Deling (data og andet materiale)
- ▶ Reproducerbarhed
- ▶ osv.

Dimensioner af reproducerbar forskning

- ▶ Tilgængelighed
 - ▶ På kort og længere sigt, *INTET* “available upon request”
- ▶ Dokumentation
- ▶ Data
 - ▶ Jo mere, desto bedre
- ▶ Åbenhed
- ▶ Kode
- ▶ Software

Dimensioner af åben videnskab (Neuhold 2014)



Åben videnskab (Braun og Ong 2014)

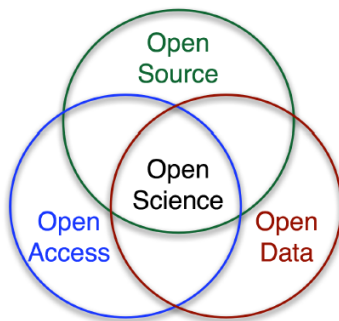


FIGURE 1.2

Open Science = Open Source Software + Open Access Papers + Open Data

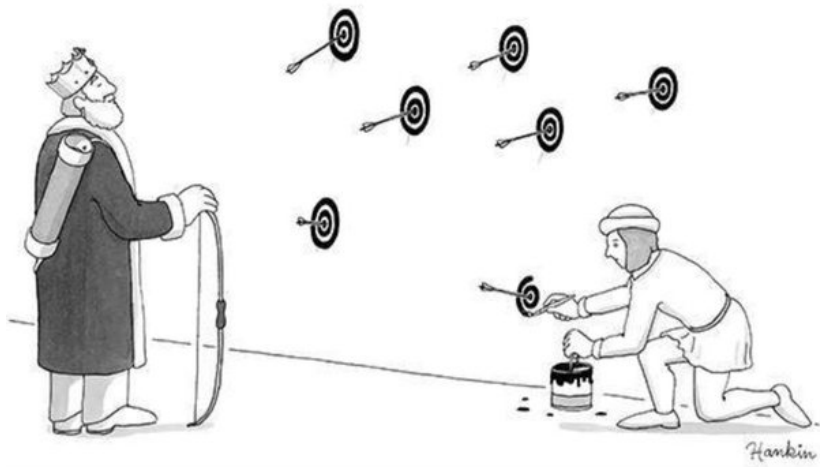
Åben videnskab: fordele

- ▶ Adgang til forskning (ikke bag betalingsmure)
- ▶ Faciliterer transparent og reproducerbar forskning
- ▶ Samarbejde med andre, bedre forskning (Kolata 2010)
- ▶ Sikrer adgang til data og andet materiale på længere sigt
- ▶ Effektivt
- ▶ Bedre forskning (Wicherts et al. 2011)
- ▶ Flere bruger ens viden (Piwowar et al. 2007, Piwowar og Vision 2013)

Mange måder at opnå signifikante resultater

- ▶ Lave hypoteser *efter* analyserne (HARKing, Kerr 1998)
- ▶ Kontrol over dataindsamlingen
 - ▶ Gennemfør analyserne løbende og stop når der er noget signifikant
- ▶ Analyser flere afhængige variable
- ▶ Vælg en anden kontrolgruppe
- ▶ Fjern ekstreme observationer
- ▶ Undersøg interaktionseffekter
- ▶ osv. osv. osv.

Mange måder at opnå signifikante resultater: *Texas sharpshooter fallacy*



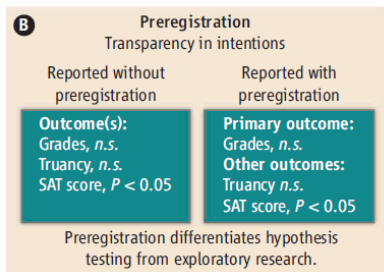
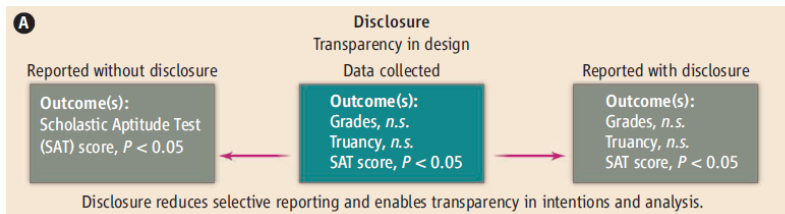
Mange måder at opnå signifikante resultater (Simmons et al. 2011)

Table 1. Likelihood of Obtaining a False-Positive Result

Researcher degrees of freedom	Significance level		
	$p < .1$	$p < .05$	$p < .01$
Situation A: two dependent variables ($r = .50$)	17.8%	9.5%	2.2%
Situation B: addition of 10 more observations per cell	14.5%	7.7%	1.6%
Situation C: controlling for gender or interaction of gender with treatment	21.6%	11.7%	2.7%
Situation D: dropping (or not dropping) one of three conditions	23.2%	12.6%	2.8%
Combine Situations A and B	26.0%	14.4%	3.3%
Combine Situations A, B, and C	50.9%	30.9%	8.4%
Combine Situations A, B, C, and D	81.5%	60.7%	21.5%

Note: The table reports the percentage of 15,000 simulated samples in which at least one of a set of analyses was significant. Observations were drawn independently from a normal distribution. Baseline is a two-condition design with 20 observations per cell. Results for Situation A were obtained by conducting three *t* tests, one on each of two dependent variables and a third on the average of these two variables. Results for Situation B were obtained by conducting one *t* test after collecting 20 observations per cell and another after collecting an additional 10 observations per cell. Results for Situation C were obtained by conducting a *t* test, an analysis of covariance with a gender main effect, and an analysis of covariance with a gender interaction (each observation was assigned a 50% probability of being female). We report a significant effect if the effect of condition was significant in any of these analyses or if the Gender \times Condition interaction was significant. Results for Situation D were obtained by conducting *t* tests for each of the three possible pairings of conditions and an ordinary least squares regression for the linear trend of all three conditions (coding: low = -1, medium = 0, high = 1).

Transparens i formidling (Miguel et al. 2014)



C

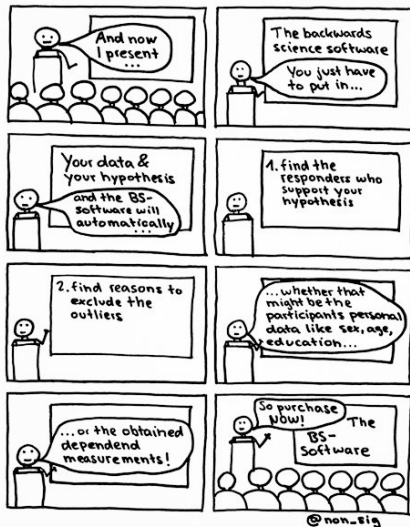
Open data and materials
Transparency in analysis

Summer break	Grades	Truancy	SAT score
Short	2.95	2%	1020
Short	3.30	0%	1360
Long	2.32	4%	9.80 ?
Long	3.87	0%	1450

Open data reduce errors and fraud and facilitate replication and extension.

Three mechanisms for increasing transparency in scientific reporting. Demonstrated with a research question: "Do shorter summer breaks improve educational outcomes?" *n.s.* denotes $P > 0.05$.

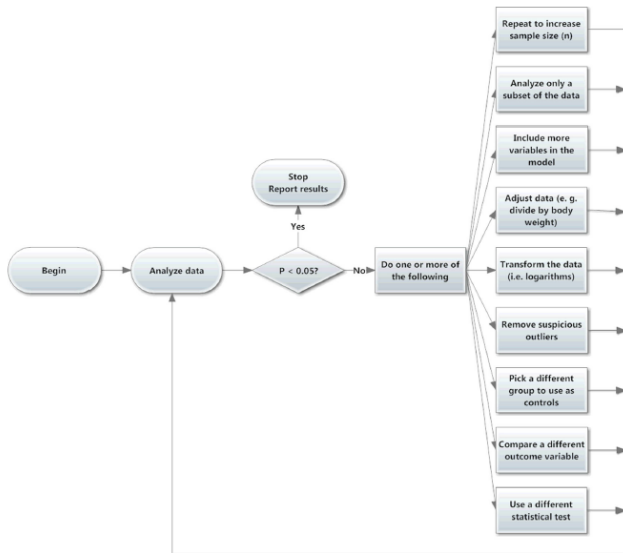
Mange måder at opnå signifikante resultater



Mange måder at opnå signifikante resultater

<u>P-VALUE</u>	<u>INTERPRETATION</u>
0.001	HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	
0.04	SIGNIFICANT
0.049	
0.050	OH CRAP. REDO CALCULATIONS.
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	
0.07	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE $P < 0.10$ LEVEL
0.08	
0.09	
0.099	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS
≥ 0.1	

Mange måder at opnå signifikante resultater (Motulsky 2014)



Eksempel: Crowdsourcing dataanalyse

- ▶ Der er mange forskellige måder at teste én hypotese på
- ▶ Selv med det *samme* datasæt kan forskellige forskere komme frem til forskellige konklusioner
- ▶ Giver fodbolddommere oftere rødt kort til mørke spillere? (Silberzahn et al. 2015)
 - ▶ “Twenty-nine teams involving 61 analysts used the same data set to address the same research questions: whether soccer referees are more likely to give red cards to dark skin toned players than light skin toned players and whether this relation is moderated by measures of explicit and implicit bias in the referees’ country of origin. [...] Twenty teams (69%) found a significant positive effect and nine teams (31%) observed a nonsignificant relationship.”

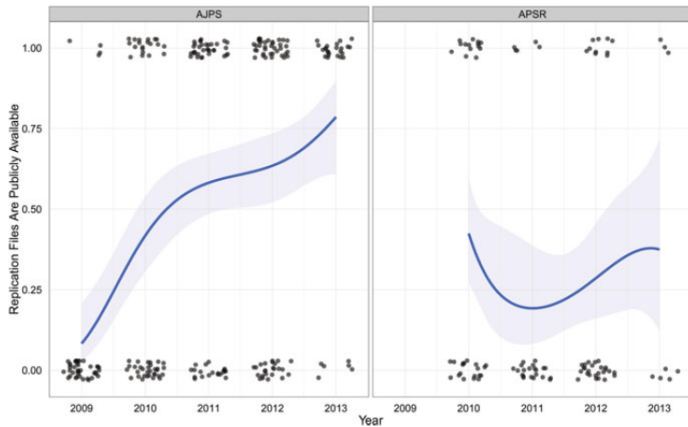
Hvor ofte er reproduktionsmaterialet tilgængeligt? (Key 2016)

Table 1

Replication Material Availability by Journal

Journal Name	Data and Code	Only Data	Only Code	Not Available	Total Articles
IO	81.6%	4.1%	4.1%	10.2%	49
AJPS	85.6%	3.4%	1.7%	9.3%	118
PA	94.4%	0%	3.7%	1.9%	54
BJPS	32.4%	16.2%	1.5%	50.0%	68
JOP	43.1%	3.6%	2.2%	51.1%	137
APSR	27.9%	4.4%	0%	67.6%	68

Hvor ofte er reproduktionsmaterialet tilgængeligt? (Dafoe 2014)



Reproduktionsmateriale

- ▶ Forskelligt materiale
 - ▶ Rå data
 - ▶ Metadata
 - ▶ Importerbar data
 - ▶ Stata do-fil/R-script
 - ▶ En readme

Reproduktionsmateriale: rå data

- ▶ Det fulde, rå datasæt
 - ▶ Data der *ikke* er behandlet på nogen måde
 - ▶ Evt. som de er indsamlet
 - ▶ Evt. som de er blevet tilsendt/downloadet
- ▶ Sikrer at der altid er originalt materiale, man kan gå tilbage til
 - ▶ Må *aldrig* overskrives/ændres

Reproduktionsmateriale: metadata

- ▶ Data omkring ens rå data
- ▶ Al information, der hjælper en med at forstå ens rå data
 - ▶ Kan være i én fil eller med flere supplerende filer
- ▶ I en spørgeskemaundersøgelse kan det være:
 - ▶ Spørgeskema
 - ▶ Kodebog
 - ▶ Samplinginformation
 - ▶ Og så videre
- ▶ Standarder for metadata:
<http://www.dcc.ac.uk/resources/metadata-standards>

Reproduktionsmateriale: importerbar data

- ▶ Det rå materiale er ikke altid optimalt i forhold til reproduktion
 - ▶ Datasæt der ikke er kompatible med forskellige statistikprogrammer (.xls, .sav, .dta)
- ▶ Få eksporteret datasættet til et åbent format
 - ▶ Gerne .csv (kommasepareret fil)
 - ▶ Lav så få ændringer som muligt
 - ▶ Dokumenter hvilke (og hvorfor)
- ▶ Brug gode standarder og gør det nemt for andre at bruge dine data (White et al. 2013)

Reproduktionsmateriale: Stata do-fil/R-script

- ▶ Alt der laves ved det importerbar data skal dokumenteres
 - ▶ Kommandoer, funktioner m.v.
 - ▶ Import af data, rekodninger, analyser, tabeller, figurer
 - ▶ Det kan være flere forskellige filer
- ▶ Automatisering
 - ▶ Alt der kan automatiseres, bør automatiseres
 - ▶ Kan køres fra start til slut
- ▶ Tekstfiler til det software, man anvender
 - ▶ Syntax-fil til SPSS, do-fil til Stata, R-script til R
- ▶ Lav kommentarer, der beskriver hvad der foregår
 - ▶ Men undlad at bruge output fra analyserne som kommentarer!!!

Reproduktionsmateriale: Stata do-fil/R-script, gode råd (Wilson et al. 2014)

- ▶ Skriv til mennesker (ikke computere)
 - ▶ Overskueligt, meningsfulde navne, konsistent opbygning
- ▶ Lad computeren gøre arbejdet (automatiser procedurer)
- ▶ Undgå gentagelser
- ▶ Lav inkrementelle forandringer
 - ▶ Undgå store ændringer, tjek om tingene fungerer
 - ▶ Dokumenter ændringer (version control)
- ▶ Forvent fejl
- ▶ Se også: Ten Simple Rules for Reproducible Computational Research (Sandve et al. 2013)

Reproduktionsmateriale: readme

- ▶ Overordnet dokumentation
- ▶ Oversigt over filer
 - ▶ Indhold og format
 - ▶ Rækkefølge, organisering af filer i undermapper
- ▶ Detajleret beskrivelse af procedure, der reproducerer ens resultater
- ▶ Referencer og links til data og andet materiale, der anvendes

Reproduktionsmateriale: data

- ▶ “Data management should be woven into every course in science.” (Nature 2009)
- ▶ Sørg for at dine data ikke går tabt!
 - ▶ Bedste løsning: upload dem til permanente dataarkiver
- ▶ Backup
- ▶ Hvis en artikel bliver publiceret og data går tabt (og ikke kan genskabes), bør artiklen trækkes tilbage
 - ▶ Derfor: Sørg for at data ikke går tabt
- ▶ Der er simple guidelines til, hvordan man gør det nemt for sig selv og andre at vedligeholde og bruge data (Goodman et al. 2014)

Data og anonymitet

- ▶ I politologisk forskning bruger vi ofte følsomme data
 - ▶ Især når det vedrører holdningsspørgsmål
- ▶ Personer deltager i spørgeskemaer med forsikringen om, at de ikke kan identificeres
- ▶ Sørg *altid* for, at data er anonymiseret
- ▶ Gerne k-anonymitet (Sweeney 2002)
- ▶ Hvilke data skal du have adgang til, før du kan identificere personer, du kender?
 - ▶ Postnummer, fødselsdagsdato og køn?

Workflow: et par råd

- ▶ Overvej hvordan du bygger projekter op
- ▶ Ikke noget der er ret meget fokus på
 - ▶ Én mappe
 - ▶ Talrige forskellige filer med samme indhold
 - ▶ Ingen kontrol med hvad der ændres
- ▶ Sørg for at systematisere et workflow, så det er nemt at dokumentere, hvad der er sket
 - ▶ Med andre ord: reproducerbar forskning

Anbefaling: versionskontrol

- ▶ Undlad intetsigende titler på dokumenter
 - ▶ `dok1.docx`, `final1.docx`, `final2.docx`, `finalfinal.docx`
- ▶ Dokumenter udviklingen i dit projekt
 - ▶ Overvej Git/GitHub: Git/GitHub, Transparency, and Legitimacy in Quantitative Research
 - ▶ Gør det nemt for dig selv at finde tidligere versioner af dit arbejde
- ▶ Overvej plain text
 - ▶ Word er WYSIWYG
 - ▶ Arbejd med LaTeX eller Markdown

Anbefaling: gode formater

Type	Anbefaling	I stedet for
Data/tabel	csv	sav, dta, xls
Tekst	tex, md	docx
Billeder	png, tiff	jpg, gif
Endelig version	pdf	docx

Gennemførelsen af en replikation

- ▶ To studier er aldrig 100% identiske
 - ▶ Som minimum forskel i tid
- ▶ 36 spørgsmål, der bør besvares (Brandt et al. 2014)
- ▶ Beskrivelse af effekt (9 spørgsmål)
 - ▶ Hvad vil man gerne replikere? Hvordan blev det pågældende studie gennemført?
- ▶ Design af replikationsstudie (7 spørgsmål)
- ▶ Forskelle på originale studie og replikationsstudie (9 spørgsmål)
 - ▶ Forskelle i procedure, mål, stimuli etc.
- ▶ Analyse af replikation (3 spørgsmål)
- ▶ Registrering (1 spørgsmål)
- ▶ Rapportering (7 spørgsmål)

Data Access and Research Transparency (DA-RT)

- ▶ Data skal være tilgængelige og proceduren skal være transparent
- ▶ Har ikke været *status quo* i den politologiske litteratur, men tiderne skifter. . .
 - ▶ Forudsætning for at kunne publicere hos blandt andre: American Journal of Political Science, American Political Science Review, American Politics Research, British Journal of Political Science, Comparative Political Studies, European Journal of Political Research, European Union Politics, International Security, Journal of Conflict Resolution, Journal of Experimental Political Science, Journal of European Public Policy, Journal of Peace Research, Journal of Theoretical Politics, Quarterly Journal of Political Science, Party Politics, Political Analysis, Political Behavior, Political Communication, Political Science Research and Methods, Research and Politics, Journal of Politics
- ▶ Mere info: dartstatement.org

Måder at maksimere sandsynligheden for at ens studie kan replikeres

- ▶ Sørg for at det kan reproduceres
- ▶ Større stikprøve
- ▶ Gode mål (høj reliabilitet)
- ▶ Undgå at teste for mange hypoteser på det samme data
- ▶ Test samme hypotese flere gange
 - ▶ Søg flere empiriske implikationer af hypotesen

Gennemfør replikationer!

- ▶ Hvis du/I stadig mangler en idé til en seminaropgave: gennemfør et replikationsstudie
- ▶ Det er vigtigt, at der gennemføres replikationsstudier
- ▶ Find et studie og se, om du/I kan replikere det
- ▶ Forslag:
 - ▶ Find amerikansk studie, repliker i Danmark
 - ▶ Find studie der tester generel hypotese på gamle data og se om det kan replikeres på nye data
 - ▶ Evt. studier der bruger første runde af European Social Survey

Flere og flere replikationsstudier (Makel et al. 2012)

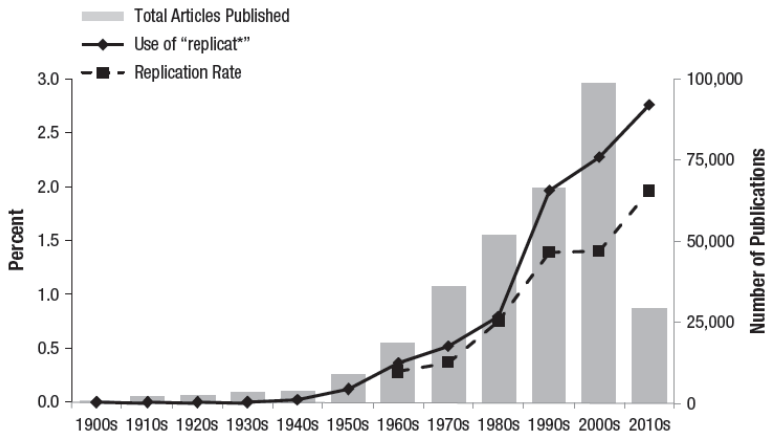


Fig. 1. Replication rate in the top 100 psychology journals. The solid line represents the percentage of publications (from 100 journals with the highest 2010 5-year impact factor) that used the term "replicat*." The dashed line reports the replication rate based on the percentage of articles using the term "replicat*" that were actual replications. The bars represent the total number of articles published in that decade. The 2010s bar is truncated because data from only 2.5 years of the current decade were available.

Dagens øvelse: Reproducere/replikere

- ▶ Diskuter med sidemanden
- ▶ Overvej hvilke studier der vil være gode at replikere
 - ▶ Hvorfor?
 - ▶ Hvilke udfordringer vil der være forbundet hermed?
- ▶ Gerne relateret til det område din/jeres seminaropgave omhandler
- ▶ 5 minutter

Næste gang

- ▶ Præsentation af seminaropgave, part 2
- ▶ Send gerne et par sider ud inden weekenden
- ▶ Mandag, kl. 10-12 i U31A