UiO : Institutt for spesialpedagogikk
Det utdanningsvitenskapelige fakultet

, Datainnsamling: Utvalgsmetoder. Ytre validitet.

Astrid Marie Jorde Sandsør

Mål

- Gjennomgå det dere har lest
- Supplere det dere har lest
- Gi perspektiver og innsikt
- Gi redskaper til å tenke

I dag

- Hvilke trusler finnes? Hva kan testes? Hva kan bare tenkes?
- Utvalgsmetoder
- Replikasjon
- Effektstørrelser

- Slides fra SPED 4010, (Zachrisson, 2020)
- Lund et al (2002), Kapittel 4

Validitet

«Validitet, eller gyldighet, betyr i hvilken grad man ut fra resultatene av et forsøk eller en studie kan trekke gyldige slutninger om det man har satt seg som formål å undersøke» (Store norske leksikon, SNL)

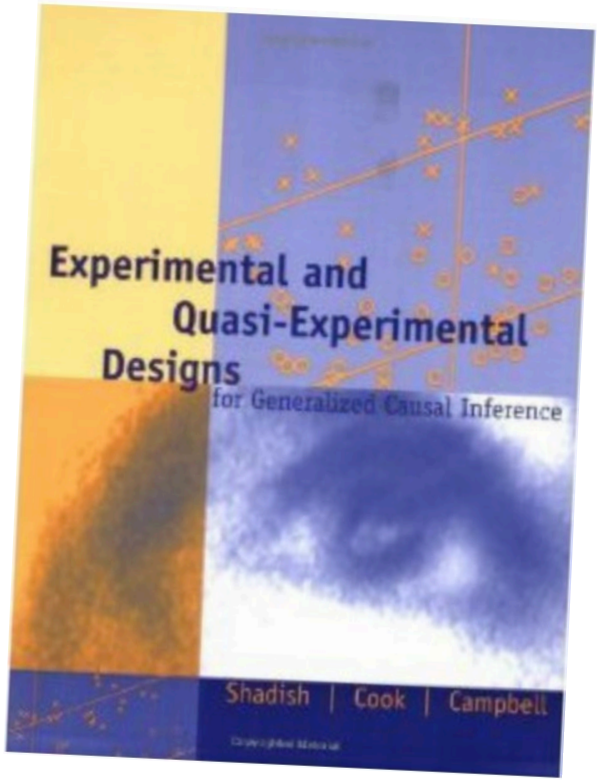
«Validity refers to the approximate truth of an inference» (Shadish et al)

- den omtrentlige sannheten ved en slutning (konklusjon)

«Tror vi at det denne studien viser er sant?»

Validitet

- Shadish, Cook, & Campell
  - Vadiditetssystem
  - Utformet for eksperimenter-brukes mer generelt
- Lund et al., 2002, Innføring i forskningsmetodologi. Oslo: Fagbokforlaget



Validitet

- Ytre validitet: Generalisering
- Indre validitet: Årsak
- Statistisk validitet: Er statistikken god (Thanassi)
- Begrepsvaliditet: måler vi det vi tror vi måler på en god måte? (Christian)

Ytre validitet

- Hvor generaliserbart er vår tolkning av en effekt?
- Grunner til feilslutninger
  - Mulige (possible)
  - Sannsynlige (probable)

Et eksempel: Greenberg & Schroder, 1997

- Jobbtrening for 18-40-åringer med psykisk utviklingshemming
  - Kontrollert eksperiment: Forsøksgruppe får treningsprogram, sammenligningsgruppe fortsetter som før
  - Resultat: Positiv effekt på tilknytning til arbeidsmarkedet og inntekt (indre validitet)
  - MEN:
    - Større effekt hvis IQ>=40
    - Hvis IQ<40; liten/ingen effekt
    - Implementert 12 steder, men ingen i sydlige USA
    - 5% av dem som ble invitert, deltok
    - 2/3 ble screenet ut (pga. atferdsvansker m.m.)
    - deltagerne var mer villige til å flytte
  - I hvilken grad kan resultatene generaliseres? (ytre validitet)

Et eksempel: 1+1 prosjektet & Two teachers

- 1+1 prosjektet: En ekstra lærer brukes til smågruppeundervisning i matematikk på barnetrinnet
  - Randomisering på skolenivå: halvparten forsøksskoler, halvparten sammenligningsskoler
- Two Teachers: En ekstra lærer hjelper klassen i leseopplæringen på barnetrinnet
  - Randomisering på klassenivå: noen klasser trukket ut til å få ekstra lærer, andre klasser er sammenligningsklasser

- Hvis vi finner effekter av tiltakene, hvor sikre kan vi være på at de samme effektene ville ha oppstått om eksperimentet gjennomføres i et annet land eller om alle på barnetrinnet i Norge får en ekstra lærer?

Generaliseringer innad i eksperimentet?

- Personer (for noen men ikke andre)?
- Steder (samme effekt alle steder)?
- Programmer (fikk alle samme «behandling»)?
- Utfall (andre typer utfall (feks livskvalitet)?

Generaliseringer utenfor eksperimentet?

- Personer (for noen men ikke andre)?
- Steder (samme effekt alle steder)?
- Programmer (fikk alle samme «behandling»)?
- Utfall (andre typer utfall (feks livskvalitet)?

Type generaliseringer

- Snever til vid
  - Fra deltagerne til populasjonen
- Vid til snever
  - Fra gruppe til subgruppe eller individ
- Samme nivå
  - Fra ett utvalg til et annet (feks sted)

Hva gjør vi?

- Identifiserer trusler mot ytre validitet
- Noen kan (kanskje) testes statistisk
- Noen krever at vi tenker

Trusler mot ytre validitet

1. Forskjeller mellom grupper/personer
2. Forskjeller over variasjon i behandling
3. Forskjeller mellom utfall
4. Forskjeller mellom steder

Trusler mot ytre validitet som *potensielt* kan testes

- «Heterogenitet i effekt»
- (Større) effekt for noen enn andre?
  - «noen»=feks kjønn, skoler, byer, kliniske subgrupper, utfall (feks type vansker)
  - Teste forskjeller mellom subgrupper
  - Lav styrke er en risiko for type II (konkluderer med at det ikke er en forskjell når det egentlig er det pga. store standardfeil)

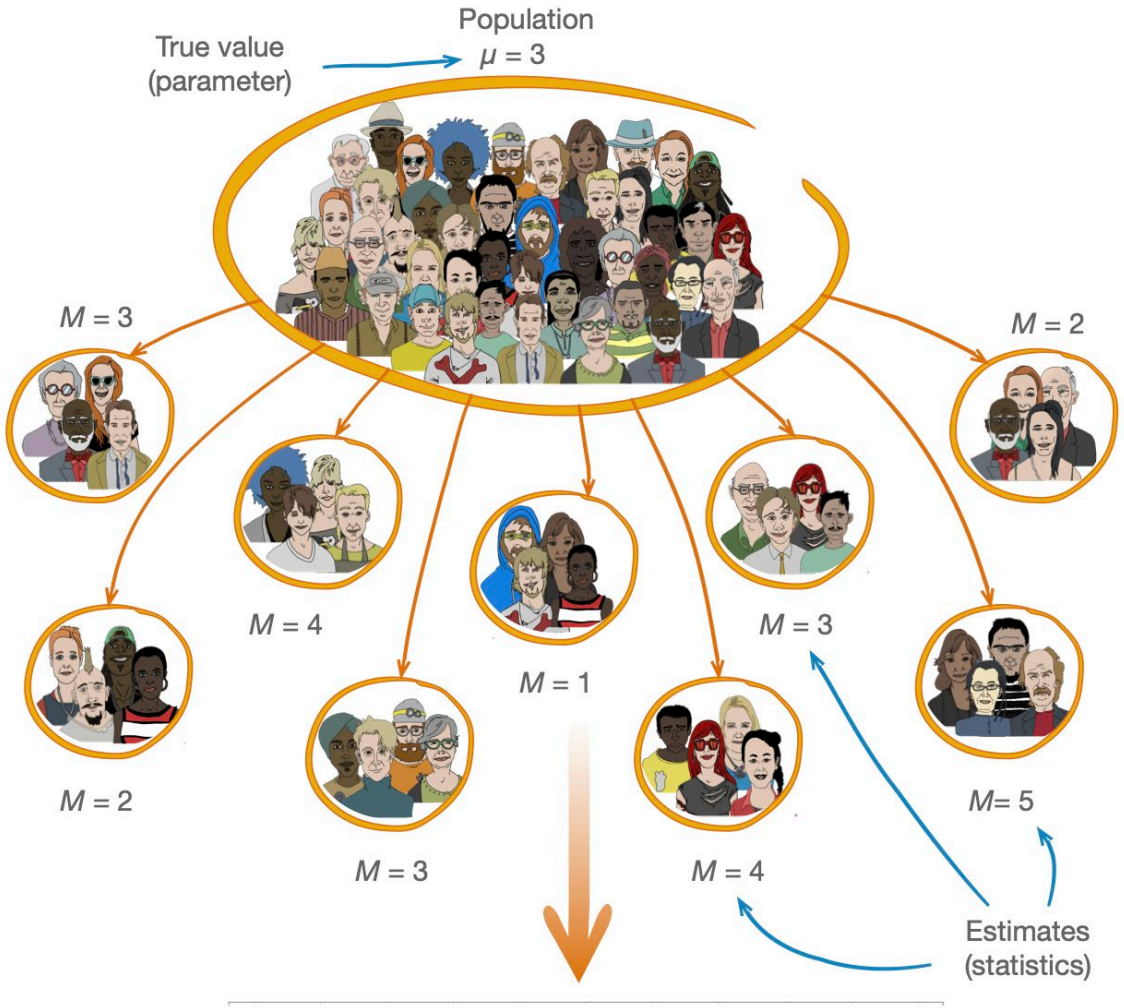
Effektstørrelse vs. retning på effekt

- Er variasjon i effektstørrelse mellom grupper indikasjon på svak ytre validitet?
- jenter r=.3; gutter r=.5
- Løst definert: Nei
- Hvorfor ikke?
  - De går i samme retning
  - Hypoteser definerer sjeldent effektstørrelser
  - Praktisk betydning er kun delvis assosiert med effektstørrelser

Trusler mot ytre validitet som krever tenking

- Kan resultatet generaliseres?
  - Andre tidspunkter?
  - Andre populasjoner?
  - Andre situasjoner?
- Utvalg (ikke-representativitet)
- Historie
- Setting/kontekst

Utvalg



Field, 2018

Utvalg

- Hvordan påvirker utvalget tolkning av resultatet?
  - Sannsynlighetsutvalg
  - Bekvemmelighetsutvalg

Enkelt sannsynlighetsutvalg

- Alle i populasjonen har lik sjanse for å bli samplet
- (i prinsippet) Representativt for populasjonen = generaliserbart
  - Krever at populasjonen er vel definert
  - Untaket i vårt felt; (Barn i Bergen)

Stratifisert sannsynlighetsutvalg

- Definere subgrupper av interesse (etnisk minoritet)
- Alle i subgruppen har lik sjanse for å bli samplet
  - resultater vektes tilbake til populasjonen = generaliserbart
- Sikrer at små grupper er representert
  - Noen eksempler: ECLS i USA, Trygg i Trondheim

Enhetsutvalg (cluster sampling)

- Velger enhet (feks skoler) tilfeldig
  - kan sample tilfeldig innad i skoler
    - klasser
    - individer
  - generaliserbart
  - eksempel: PISA, TIMMS/PIRLS

Sannsynlighetsutvalg

- (potensielle) muligheter
  - Effekter i populasjonen (hvis kausalt design)
  - Forekomst i populasjonen
  - Sammenhenger i populasjonen
- (potensielle) trussler
  - Selektiv deltagelse
    - Hvor representativ er faktisk utvalget
    - Hvem deltar når de blir utvalgt?
    - Frafall
  - Løsning: sammenligne utvalg og populasjon

Bekvemmelighetsutvalg

- Ikke tilfeldig: «rett sted til rett tid»
  - Mest vanlig i vårt felt
  - Det dere (mest sannsynlig) vil gjøre hvis dere samler data
  - Krever at vi tenker enda mer:
    - Hva er truslene mot generaliserbarhet?

Strategisk utvalg

- Definere tid og sted for utvalg
- Inkludere de som kommer

Eksempel:

- Barns Sosiale Utvikling (NUBU)
  - 5 kommuner
    - Nær Oslo
    - «representativ» demografi
  - 3 kohorter
  - Alle som kom på helsestasjon ble invitert
    - 60% deltok

Strategisk bekvemmelighetsutvalg

- (potensielle) muligheter
  - Effekter i populasjonen (hvis kausalt design)
  - Sammenhenger i populasjonen (hvis vi kan sannsynliggjøre noen grad av representativitet)
- (potensielle) trussler
  - Selektiv deltagelse
    - Hvor representativ er faktisk utvalget
    - Hvem deltar når de blir utvalgt?
    - Frafall
  - Løsning: sammenligne utvalg og populasjon

Trusler mot ytre validitet som krever tenking

- Utvalg (ikke-representativitet)
- Historie
- Setting/kontekst

Historie

- Er tolkningen generaliserbar til andre tidsrom?
  - Effekt av barnehage i 1975-1980
  - Kan dette informere oss om barnehagen i dag?
- Kohorteffekter (variasjon mellom årskull)
- Endringer i kontekst, f.eks.
  - Policy
  - Sosialt miljø
  - Segregering/fattigdom

Setting/kontekst

- Land/region, f.eks.
  - Er resultater fra Norge relevante i Hellas?
  - Er resultater fra Oslo relevante i Narvik?

Hvordan kan vi tenke rundt dette?

- Kjenne/undersøke relevante forhold
  - Ekstern informasjon
- Kritisk vurdering av forskjeller/likheter

Replikasjon (kan det reproduseres)

- Studier gjøres på nytt/gjentas
  - Kunnskap er kumulativ
    - Kunnskap er foreløpig
    - Popper/falsifikasjonisme
- Støtter/svekker generaliserbarhet

En studie er ikke nok

- En studie kan ha tilfeldige resultater
  - Sampling error (utvalget er ikke representativt)
  - Tilfeldige resultater
  - Problemer med design/gjennomføring
- Publication bias

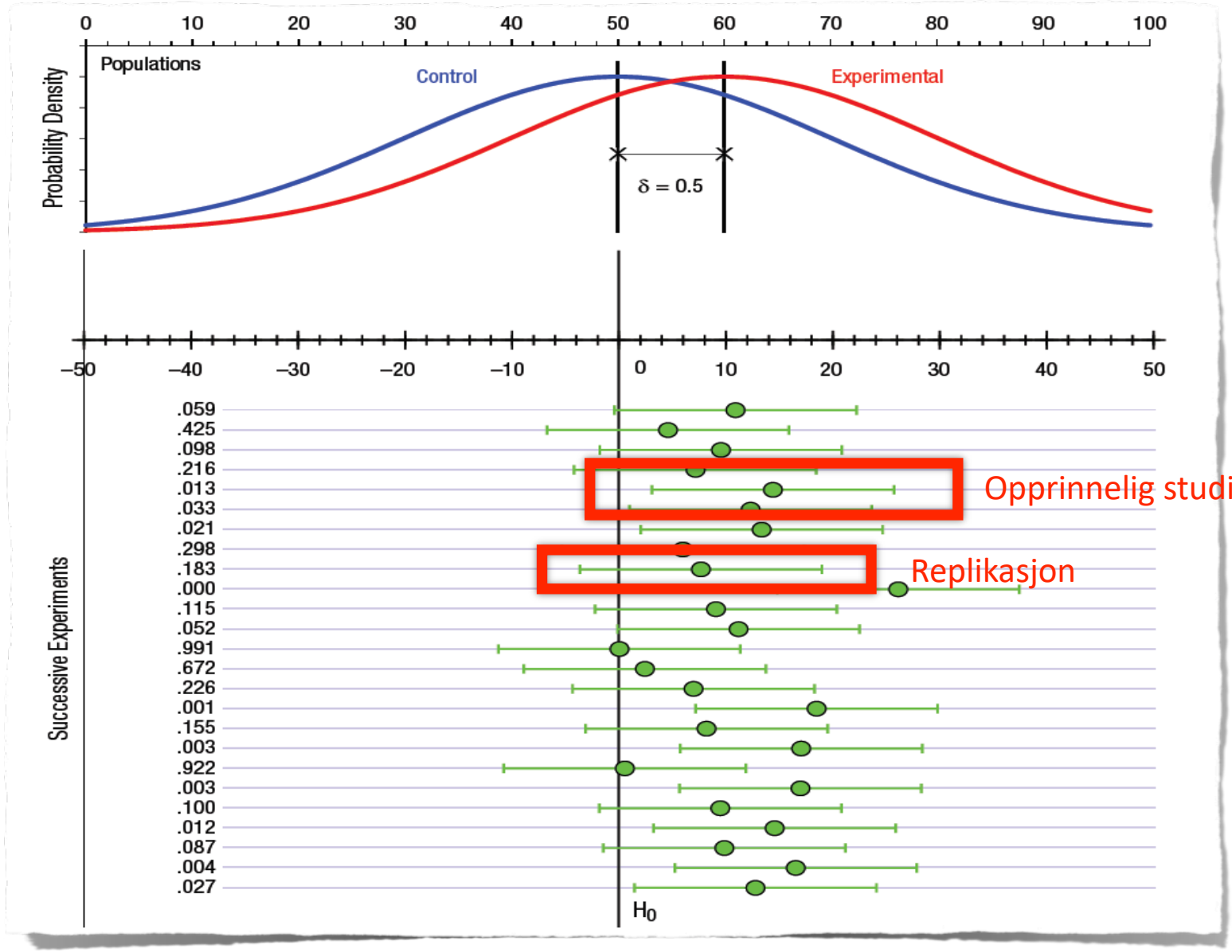
To typer replikasjoner

- Konseptuelle
  - Ulik populasjon eller gjennomføring
- Eksakt
  - Samme populasjon eller gjennomføring

To typer replikasjoner

- Hva vil det si at et resultat har blitt replikert?
  - Samme p-verdi?
  - Estimat innenfor 95% konfidensintervall?
  - Samme effektstørrelse?
  - (Vurdering, ikke eksakt svar)

Replikasjon



Men replikasjon er ikke enkelt

RESEARCH

RESEARCH ARTICLE

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*†

Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. Replication effects were half the magnitude of original effects, representing a substantial decline. Ninety-seven percent of original studies had statistically significant results. Thirty-six percent of replications had statistically significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of effects were subjectively rated to have replicated the original result; and if no bias in original results is assumed, combining original and replication results left 68% with statistically significant effects. Correlational tests suggest that replication success was better predicted by the strength of original evidence than by characteristics of the original and replication teams.

	Replications $P < 0.05$ in original direction	Percent
Overall	35/97	36
JPSP, social	7/31	23
JEP:LMC, cognitive	13/27	48
PSCI, social	7/24	29
PSCI, cognitive	8/15	53

Hovedfunn

- 36% av resultatene kunne replikeres
- 47% av replikasjonene hadde effekstørrelser innen 95% konfidensintervall
- 39% av replikasjonene ble betraktet som faktiske replikasjoner
- Effekstørrelse og signifikans i originalstudie var viktigere for replikasjon enn erfaringen til forskergruppen som replikerte

Hovedfunn

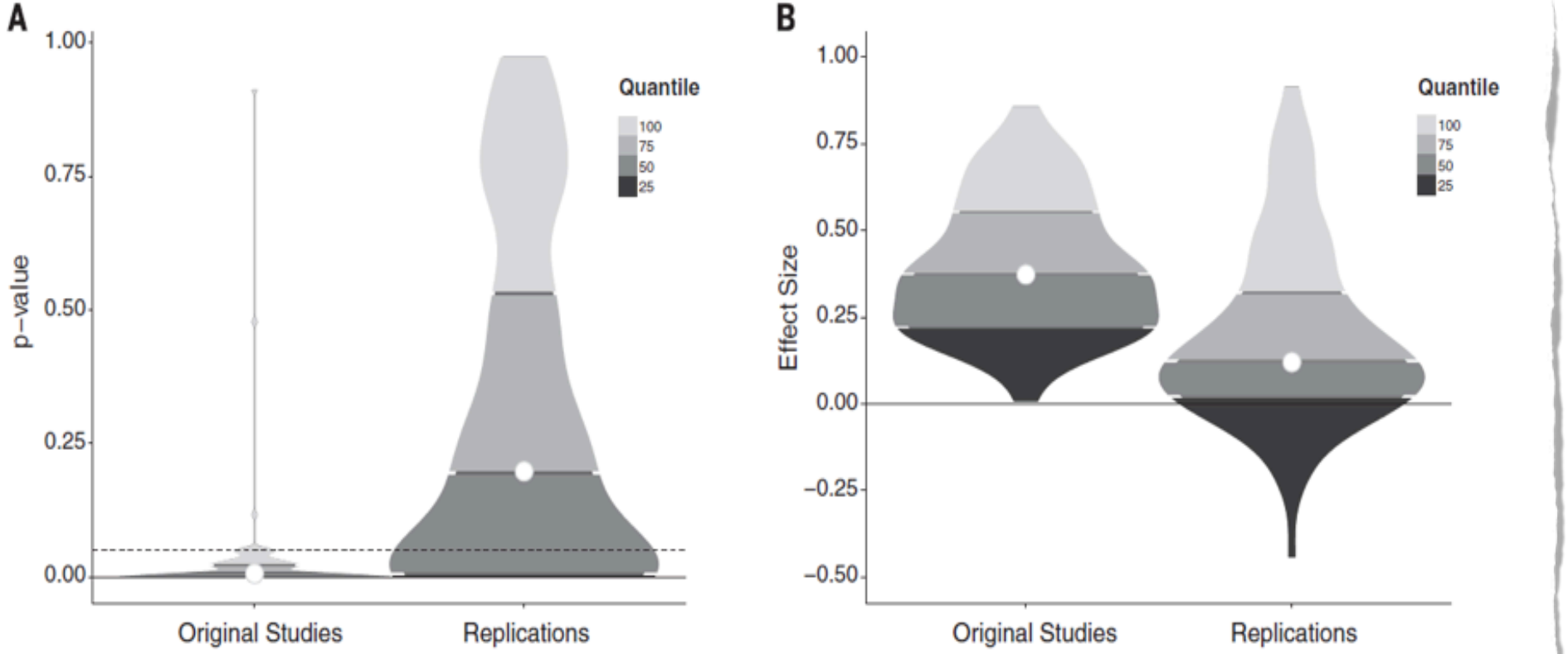


Fig. 1. Density plots of original and replication *P* values and effect sizes. (**A**) *P* values. (**B**) Effect sizes (correlation coefficients). Lowest quantiles for *P* values are not visible because they are clustered near zero.

Hva er problemet?

- Vær varsom dersom du ser studier med:
  - Lav statistisk kraft (statistical power)
  - Overraskende resultater
  - P-verdier nærme 0.05
- Risikerer at det er lav sannsynlighet for at de kan replikeres

Vurdering av ytre validitet

- Som annen validitet
  - Ikke et ja/nei spørsmål
- Vurdering av summen av argumenter
  - I lys av annen forskning
- Kontinuum fra sterk til svak
- Replikasjon er en del av denne vurderingen

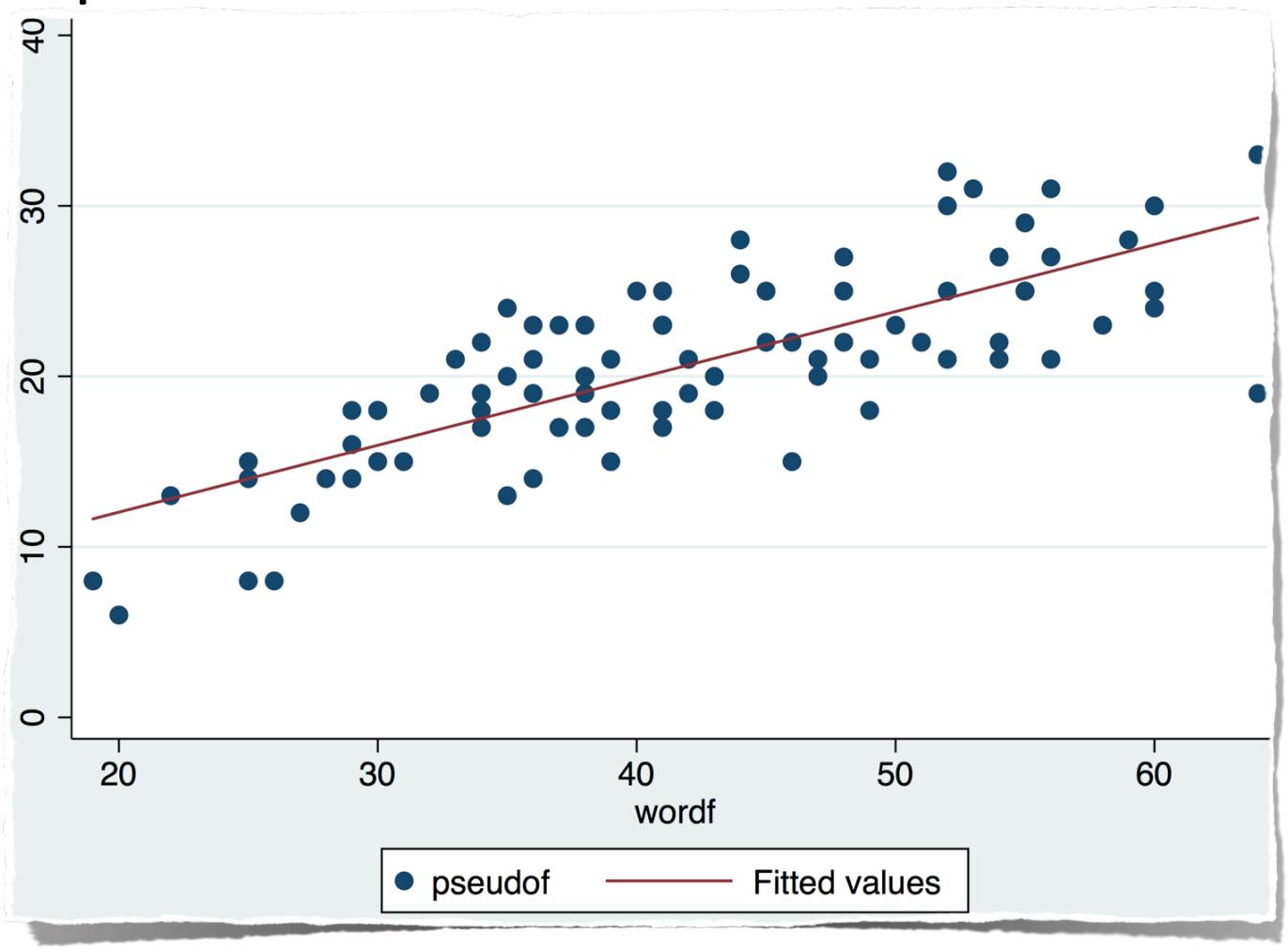
Hva er en effektstørrelse (ES)

- Størrelsen på en effekt vs. signifikans
  - (test statistikk) = (utvalgstørrelse) x (ES) - Cohen (1977)
- En (hvilkensomhelst) index for sammenheng eller forskjeller
- Assosiasjon, forskjell, andel
  - f.eks. korrelasjon (0-1) eller forskjell mellom grupper (i SD)

Vanlige effektstørrelser

- rn’B - familien (sammenheng)
  - r & partial r
  - standardisert Beta
- d - familien (forskjell)
  - Cohen’s d
  - Hedge’s g
- Standarisert ES gjør det lettere å forstå og å sammenligne

Eksempel r



Eksempel forskjell i standardavvik (SD)

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
F	45	43.42222	1.606329	10.77558	40.18488	46.65956
M	45	39.75556	1.537463	10.31362	36.657	42.85411
combined	90	41.58889	1.122452	10.64852	39.3586	43.81918
diff		3.666667	2.22353		-.7521318	8.085465

diff = mean(F) - mean(M) t = 1.6490
Ho: diff = 0 degrees of freedom = 88

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0
Pr(T < t) = 0.9486 Pr(|T| > |t|) = 0.1027 Pr(T > t) = 0.0514

Cohen's d: 3.67/10.65= 0.34

Finnes det standarder?

differences. Effect sizes (Cohen's *d*), were calculated as indicators of the magnitude of differences between groups and interpreted as small (.20 to .49), moderate (.50 to .79), and large (≥ .80). Second, Confirmatory Factor Analyses (CFA) was conducted in

- Cohen's (1977) book on statistical power
  - rules of the thumb for power calculations «Cohen's standards» - only to be applied if no other info was available
- «The folklore of social research» (Hedges, 2008)
  - Decontextualized criteria (…) are not helpful

Finnes det standarder?

The Earth Is Round ($p < .05$)

Jacob Cohen

American Psychologist
December 1990 Vol. 45, No. 12, 1304-1312

Things I Have Learned (So Far)

Jacob Cohen
New York University

ABSTRACT

This is an account of what I have learned (so far) about the application of statistics to psychology and the other sociobiomedical sciences. It includes the principles "less is more" (fewer variables, more highly targeted issues, sharp rounding off), "simple is better" (graphic representation, unit weighting for linear composites), and "some things you learn aren't so." I have learned to avoid the many misconceptions that surround Fisherian null hypothesis testing. I have also learned the importance of power analysis and the determination of just how big (rather than how statistically significant) are the effects that we study. Finally, I have learned that there is no royal road to statistical induction, that the informed judgment of the investigator is the crucial element in the interpretation of data, and that things take time.

Så hva betyr effekstørrelser?

Child Development, January/February 2000, Volume 71, Number 1, Pages 173–180

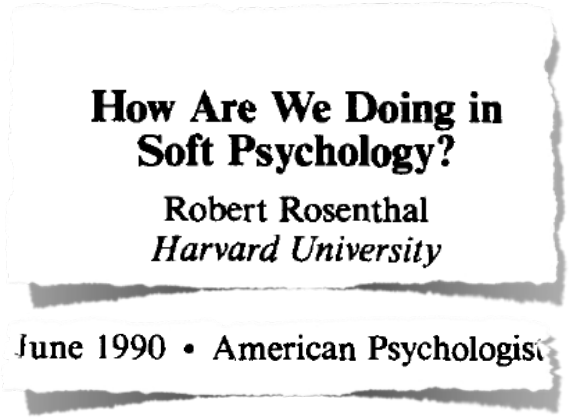
Effect Size, Practical Importance, and Social Policy for Children

Kathleen McCartney and Robert Rosenthal

«Just as children are best understood in context, so are effect sizes»

Er 0.03 en stor effektstørrelse?

- Hindrer aspirin hjerteinfarkt? Rosenthal (1990)
  - Ja, 0.9 % risiko i forsøksgruppe, 1.7% risiko i sammenligningsgruppe
- Hva gjør at effekten er stor?
  - Type utfall (død)
  - Kostnad av tiltak (nesten gratis)
  - Ingen kjente bivirkninger
  - Høy kvalitet på studien (Eksperiment med placebo, n=22000)
  - (Kjent kasual sti)


How Are We Doing in Soft Psychology?

Robert Rosenthal
Harvard University

June 1990 • American Psychologist

Praktisk betydning avhengig av konteksten

Child-Care Effect Sizes for the NICHD Study of Early Child Care and Youth Development

NICHD Early Child Care Research Network

February–March 2006 • American Psychologist
 0003-066X/06/$12.00
Vol. 61, No. 2, 99–116 DOI: 10.1037/0003-066X.61.2.99

Table 5
Child-Care and Parenting Effect Sizes

Variable	Maternal care *d*	Child-care quality *r/R*	Child-care quality *r*	Child-care quality r_p	Child-care quality *d*	Parenting *r/R*	Parenting *r*	Parenting r_p	Parenting *d*
Cognitive development									
School readiness	.01	.44	.27***	.12***	.38***	82	.50***	.24***	.89***
Receptive language	−.03	.43	.27***	.12***	.39***	.86	.54***	.22***	.85***
Expressive language	−.04	.43	.19***	.08*	.34**	.85	.37***	.17***	.62***

- Sammenligne ES med andre studier

«Oversette» effekter

- Måneders læring
  - f.eks. gitt fravær av intervensjon
- Andre relevante sammenligninger
  - grupper, kontekster