KØBENHAVNS UNIVERSITET

Hvordan ved vi om vi hjælper?
Effekt- og virkningsevaluering af udviklingsbistand

Eva Broegaard
Ninja Klejnstrup

Særlig tak til:

Ole Winckler Andersen, Danida/OECD
Henrik Hansen, Institut for fødevare- og ressourceøkonomi
Jens Anders Kovsted, Ph.d
Beate Bull, Rådgiver, EVAL-Norad

Hvordan ved vi om vi hjælper?

Fokus på de aktuelle tendenser for effekt- og virkningsevaluering – hvor er debatten og hvorfor?

- **Udfordre særstatus** for eksperimenteter/RCT
  - Men ingen RCT-bashing her!
  - Kvantitative og kvalitative muligheder

De klassiske spørgsmål:

- Virker det vi gør – har det en effekt?
- Gør vi det rigtige – og gør vi det rigtigt?
- Kan vores viden bruges fremadrettet?
- **Metodisk fokus** – så bredere relevans

Nogle **nye analyser** – og nogle **evigtunge debatter**

Ikke metodisk i dybden – mere med **den brede pensel**

Baggrund – lidt om os

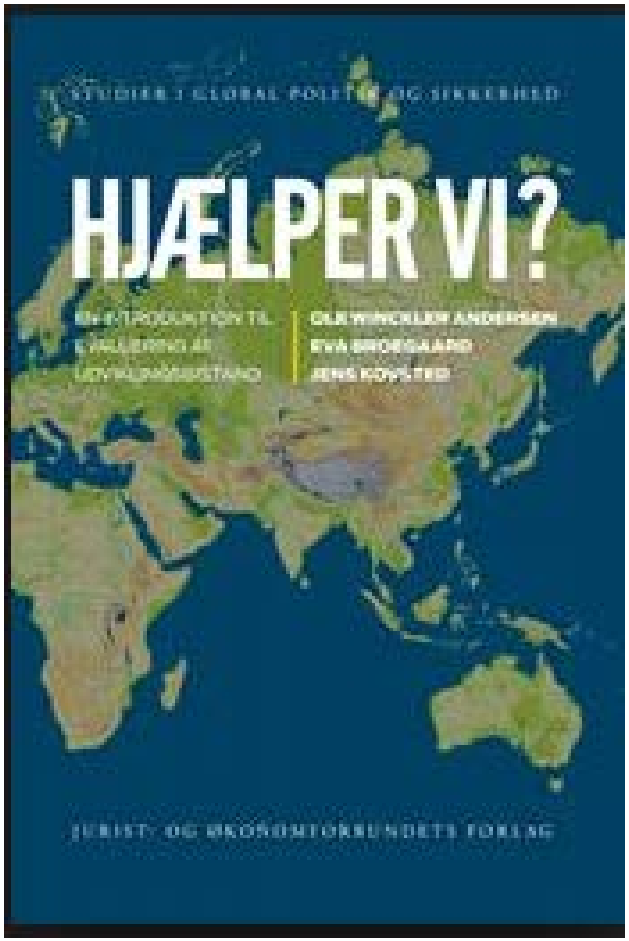
Evalueringsnørder – Økonom og Scient.pol.

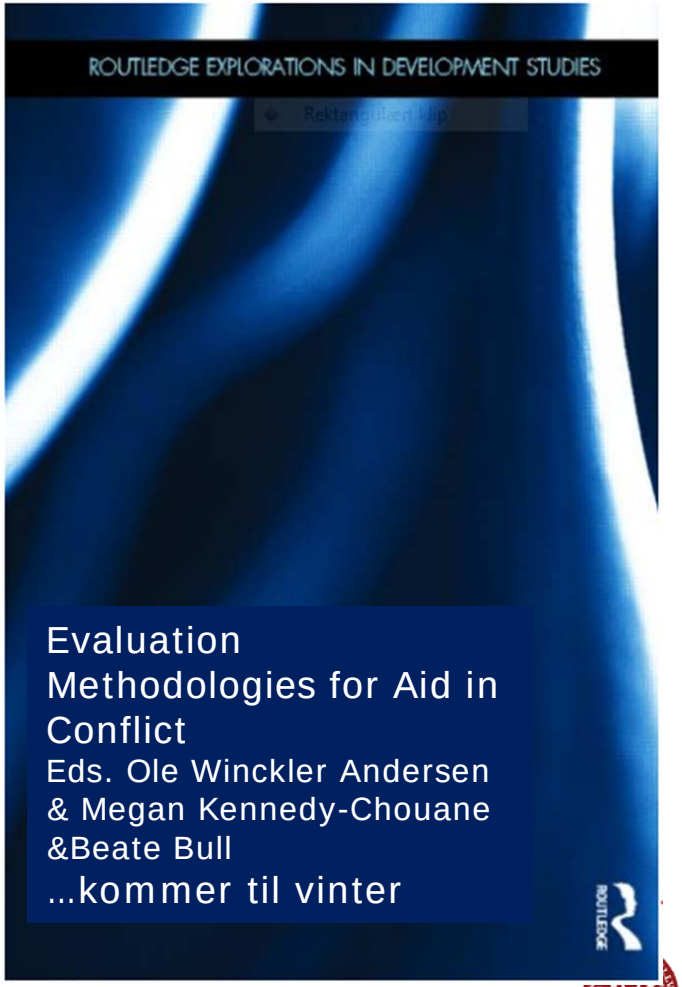
- Tidligere Danida's Evalueringskontor
- Fokus på effektevaluering som "det muliges kunst"
  - Krav om evidens – for et uhyre diverst felt
- International udveksling af erfaringer:
  - Eksperimentelle metoder er stærke og vigtige
  - Men langt fra nok! Andre kvantitative + kvalitative i spil
  - Debatter om metoder og tilgange: Hvad kan vi så gøre

NB! Det er ikke Danida der står her i dag

- Personlige erfaringer, publiceret materiale, den bredere internationale debat om udviklingsbistand og evaluering…
  - Så "**the usual disclaimers apply":** Alle synspunkter er udtryk for oplægsholdernes personlige vurderinger og erfaringer osv…

Lidt af det vi trækker på:



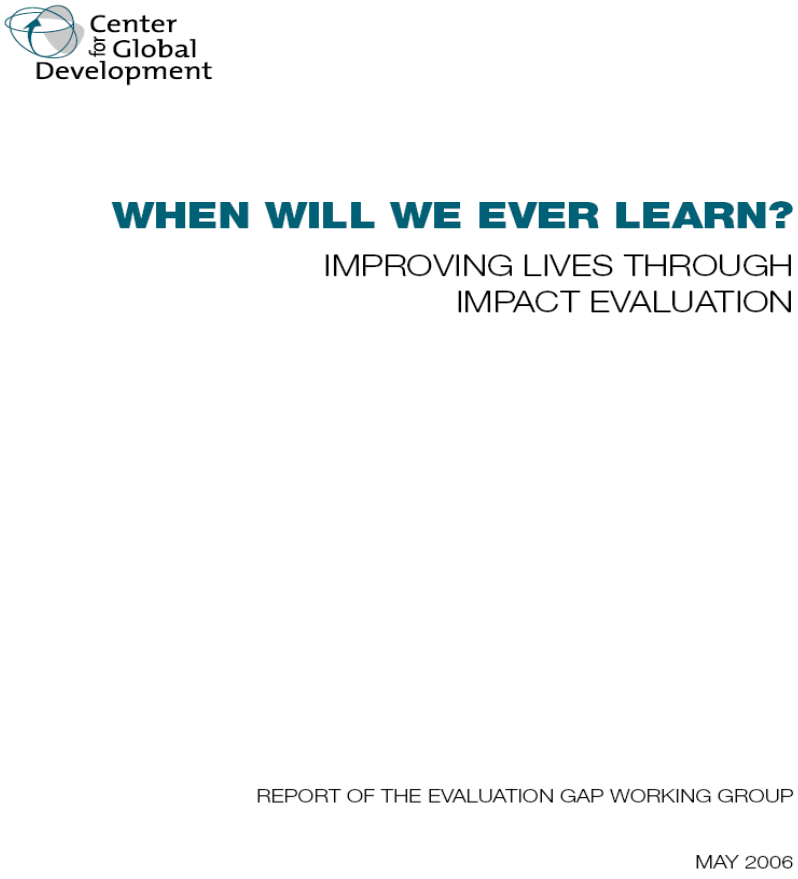


Påstand/håb: Metodekrigen er slut…?

Men selvom vi er færdige med at råbe er vi ikke færdige med at lære af hinanden?

- Fra Flyvbjerg (2011) til praktisk syn på f.eks.:
- **”Repræsentativitet”, ”cases” og ”generalisering”**
- Stadig **praktisk potentiale** for evalueringsfeltet

Effektevaluering– den internationale arena



"Yet after decades in which development agencies have disbursed billions of dollars for social programs, and developing country governments and nongovernmental organizations (NGOs) have spent hundreds of billions more, it is deeply disappointing to recognize that we know relatively little about the net impact of most the social programs."

U

Effektevaluering - terminologi

The international Initiative for Impact Evaluation 3ie

- "Rigorous impact evaluations studies are analysis that measure the **net change in outcome**s for a particular groups of people that can be **attributed** to a specific program using the best methodology available, feasible and appropriate….
- [….]
- "the difference in the indicator of interest (Y) with the intervention Y1) and without the intervention Y0). That is, **impact = Y1 – Y0.** An impact evaluation is a study which tackles the issue of attribution by identifying the **counterfactual** value of Y (y=) in a rigorous manner (White 2010).

Den internationale arena

Skabte en stærk trend – som i andre sektorer:

3ie - Improving lives through impact evaluation:

- Budget: 2011 - 40 mio US
- Klart fokus på RCT (men ikke udelukkende) og systematiske reviews ud fra et klassisk evidens-hierarki

Jpal poverty lab: Conducting Rigorous IE:

- J-PAL researchers conduct randomized evaluations to test and improve the effectiveness of programs and policies aimed at reducing poverty.
- Bannerjee, Duflo m.fl.
- "2013: 83 affiliate professors. 424 ongoing or completed evaluations in 53 countries. 1512 people trained".

DIME: Verdensbankens " Development Impact Evaluation"

- Vægt på eksperimentielle metoder –summativt og formativt!
- 2010: IE for ca. 13% af aktiviteterne; budget: ca. 14 mio US

DK: Udviklingsbistand og evaluering

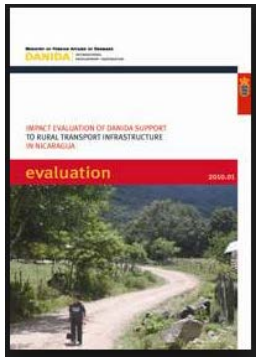
Danidas Evalueringskontor; evalueringer siden 1982;

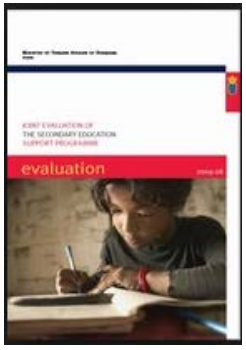
- Også stærk interesse for effektevaluering

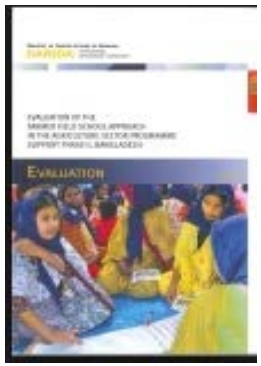
RCT har ikke været anvendt – (endnu?)

Andre slags kvantitativ og/eller kontrafaktisk analyse mulig

- Men krævede kreativitet og en del held – timing/data!
- Kombineret med teoribaseret analyse og kvalitative metoder…







IE Praksis – Hurtig status

- Der bliver genereret en stadig større mængde evidens om effekt, der gentages forsøg, lavet metastudier osv.
  - Men der er en lang række indsatser/situationer, hvor RCT ikke er anvendelig
- Dækningsgrad; relevans for feltet?
  - Varierer, men estimeres typisk 5-15%
- Ud fra OECD-DAC database: Fortsat stor andel
  - Helt eller delvist kvalitative; uden kontrafaktisk analyse;
  - Mange er transparente, grundige og solide – andre med ”plads til forbedring”…
- Interesse: Hvor sikre kan vi blive – og hvordan?
  - Hvordan kan vi se på hvornår og hvorfor det virker?
- Flere spor i debatten – både om kvalitative, blandede metoder og **kvantitative alternativer**

Kvantitative effektevalueringer/IE

- Hansen, Klejnstrup and Andersen (2013):
  - A Comparison of Model-Based and Design-Based Impact Evaluations of Interventions in Developing Countries, American Journal of Evaluation, 34: 320-338

- Den kvantitative effektevaluering og

  Den **kontrafaktiske tilgang**

- **Lodtrækningsforsøget** som en løsning på det fundamentale problem ved kausal inferens

- Diskussionen af fordele og ulemper for RCT/lodtrækningsforsøg
  - Ofte særstatus, men….

- **Kvantitative alternativer** i praksis

Kvantitative IE - udgangspunktet

- En kvantiativ **effektevaluering** søger at identificere et projekt, program eller en politiks **kausale bidrag** til et udfald *hos modtagerne.*

- **Effekten** af en intervention defineres som forskellen mellem det udfald en deltager **faktisk** oplever, og det udfald hun ville have oplevet i den **kontrafaktiske** situation, hvor hun ikke deltog i programmet – alt andet lige.

$$\delta(u) = Y_1(u) - Y_0(u)$$

$$Y(u) = \begin{cases} Y_1(u) \text{ hvis } u \text{ deltager} \\ Y_0(u) \text{hvis } u \text{ ikke deltager} \end{cases}$$

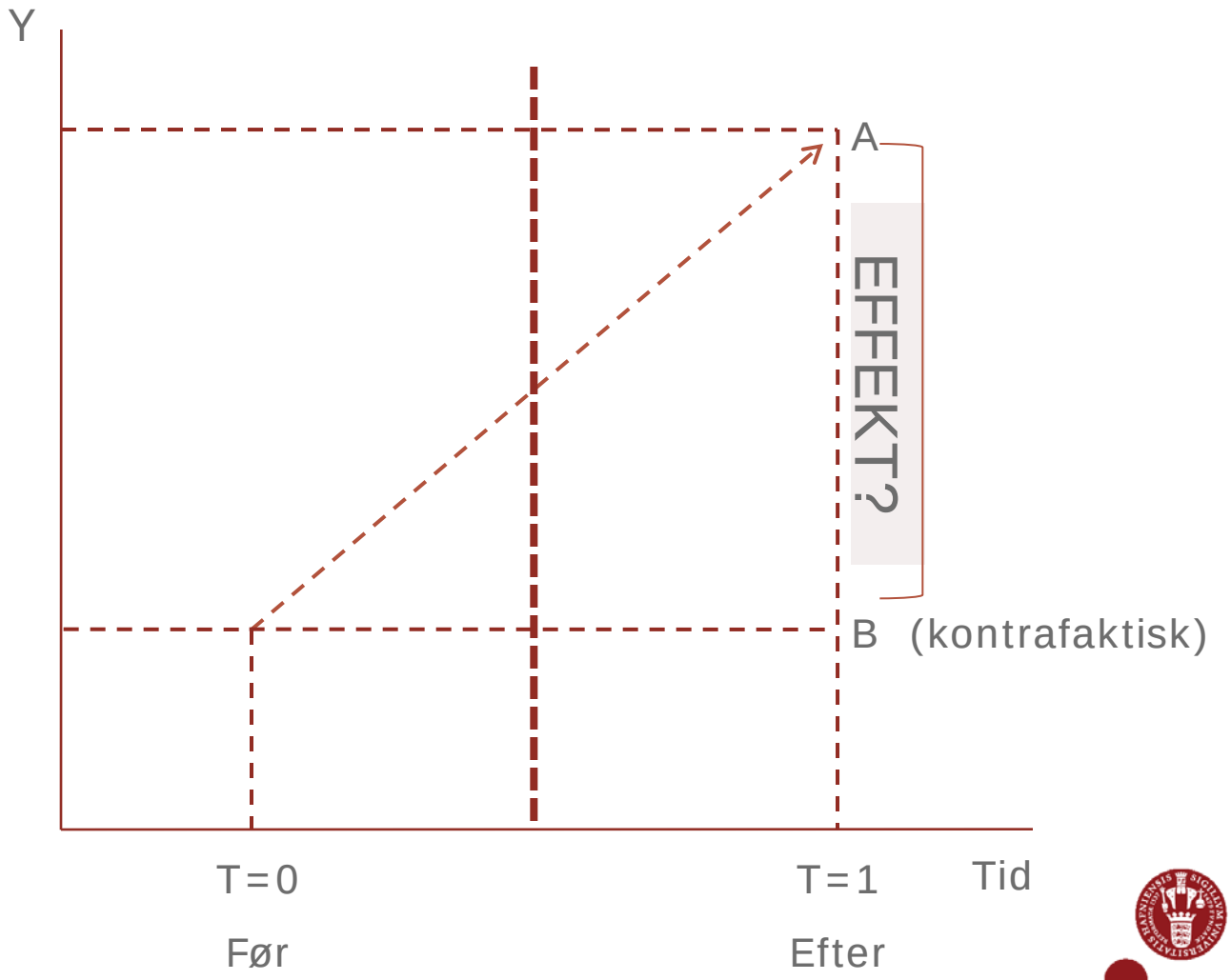
- Det **fundamentale problem** ved kausal inferens: det kontrafaktiske observeres ikke

En bud på det kontrafaktiske udfald: Den oprindelige tilstand

Efter programmet er udfaldet A

Før programmet var udfaldet B

Effekten er A-B



En evalueringsløsning

- Vi kan ikke bestemme den kausale effekt på individ niveau, så vi søger den gennemsnitlige effekt i befolkningen.
- Vi finder den gennemsnitlige effekt som forskellen mellem det gennemsnitlige udfald for deltagerne og det gennemsnitlige udfald for en gruppe af ikke-deltagere

$$\bar{\delta} = E_U[Y(u)|D = 1] - E_U[Y(u)|D = 0]$$

$$D = \begin{cases} 1 \text{ hvis } u \text{ deltager} \\ 0 \text{ hvis } u \text{ ikke deltager} \end{cases}$$

- Dvs. vi erstatter det ikke observerede *kontrafaktiske* udfald med et estimat, der er baseret på det gennemsnitlige udfald blandt ikke-deltagerne.
- Er det et problem?

Fordele og ulemper

- **Hvorfor vælge lodtrækningsforsøg?**
  - De sikrer at vi kan arbejde ud fra "alt andet lige" i sammenligningen
  - De er robuste i den forstand at de ikke forudsætter en økonomisk eller statistisk model

  Men
  - De er ikke særligt informative da vi ikke har en økonomisk eller statistisk model **(formaliseret teori)**
  - Vi ved ikke noget om den eksterne validitet
  - Kan være etisk/politisk omkostningsfulde - forudsætter f.eks. at tildeling/implementering af indsatser sker ud fra evalueringens præmisser

- **Alternativet: Statistik modellering**
  - Teoretisk: kan give samme resultat som lodtrækningsforsøget
  - Er informativ om en statistiske eller økonomiske model
  - Vi kan sige noget om ekstern validitet
  - Kan basere sig på eksisterende data – fleksibilitet og ressourceeffektivitet

  Men
  - Resultaterne afhænger af **modellens antagelser** – hvis antagelserne er forkerte for vi skæve resultater (bias).
  - Risiko for "**uobserverede"** forhold af betydning?
  - Anses derfor i nogle sammenhænge som "inferiør" – ikke tilstrækkelig pålidelig.

Direkte sammenligninger

- Sammenligner design- og model-baserede estimater:
  - **Samme deltagere**
  - **Kontrol/sammenligningsgruppen afviger**
  - **Hvis effekt-estimaterne afviger signifikant er den model-baserede estimator biased**

- Mange direkte sammenligninger fandt skæve model-baserede estimater
  - Primært **amerikanske arbejdsmarkedsinterventioner**
  - Model-baseret tilgang er videreudviklet siden da!

- 4 nyere direkte sammenligninger i sat i udviklingslande
  - **Model-baserede estimater matcher design-baserede** når udvælgelsen til programmet er velbeskrevet.
  - Dvs. ”leverer” på lige fod med RCT

De 4 studier

- Lodtrækningsforsøg/RCT sammenlignet med:
  - **Regression Discontinuity Design** (RDD)
  - **Ordinary Least Square / Lineær Regression** (OLS)
  - **Propensity Score Matching (PSM)** – med forskellige matching-strategier
  - **Propensity Score Matching** og **instrument-variable**.
- Tegner det samme billede
- Vi ser på ét eksempel

PROGRESA

- Mexicansk **Conditional Cash Transfer (CCT) Programme** fra 1997
  - Målrettet sundhed og uddannelse af børn i fattige husstande
  - Kontanter mod fremøde i skole og ved sundhedsundersøgelser.

- **Klart beskrevet udvælgelseskriterier**
  1. Lokaliteter udvalgt på baggrund af marginalitetsindex
  2. Husstande udvalgt på baggrund af ”discriminant score”
  3. Ratificering af liste over støtteberettigede på landsbymøder (få ændringer)

- I alt 506 lokaliteter
- 186 tilfældigt udvalgt til at udgøre kontrolgruppe

Hvilke konklusioner tegner sig?

- **Den modelbaserede tilgang kan føre til gode estimater**
  - Ikke signifikant forskellige fra RCT baserede estimater
- Men kræver en model – som kræver en udfoldet **programteori/indsatsteori**
- Ideelt set bør evaluatoren formulere en **eksplicit model** for det specifikke program, som indeholder en beskrivelse af forhold at betydning for både effekten og udvælgelsen
- Nationale spørgeskemaundersøgelser *kan* bruges til at danne sammenligningsgrupper, men **forskelle i survey instrumenter** kan føre til signifikante biases.
  - Simple udfaldsvariable giver bedste resultater.
- **Jo mere vi kan bruges vores viden om program, kontekst-betydning, udvælgelse, jo bedre**
- Cook, Shadish &Wong (2008) har en tilsvarende konklusion…

Debatten om virkningsevaluering:

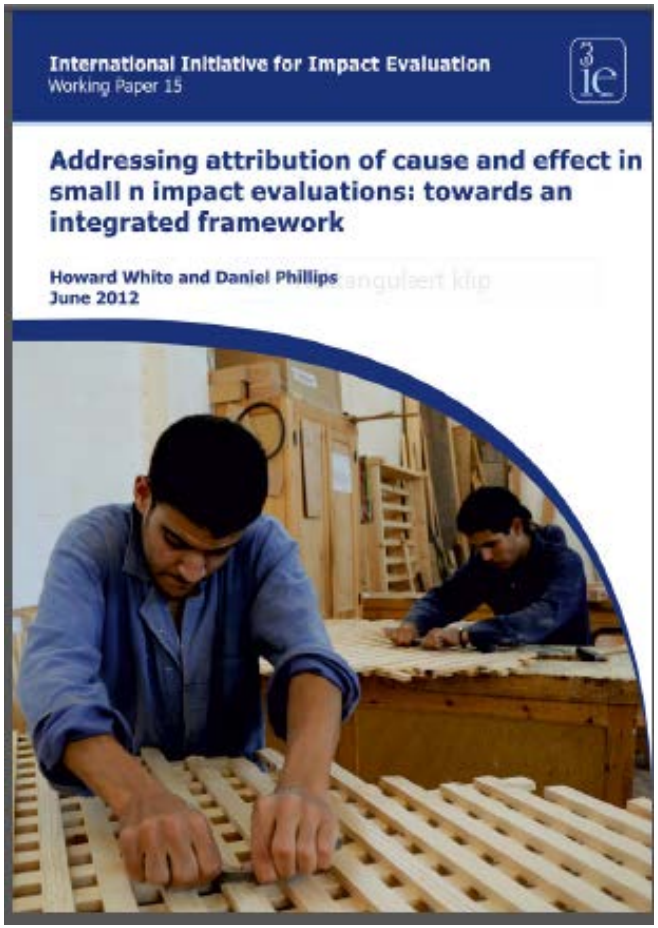
Så: kvantitative alternativer, men stadig begrænset dækning!

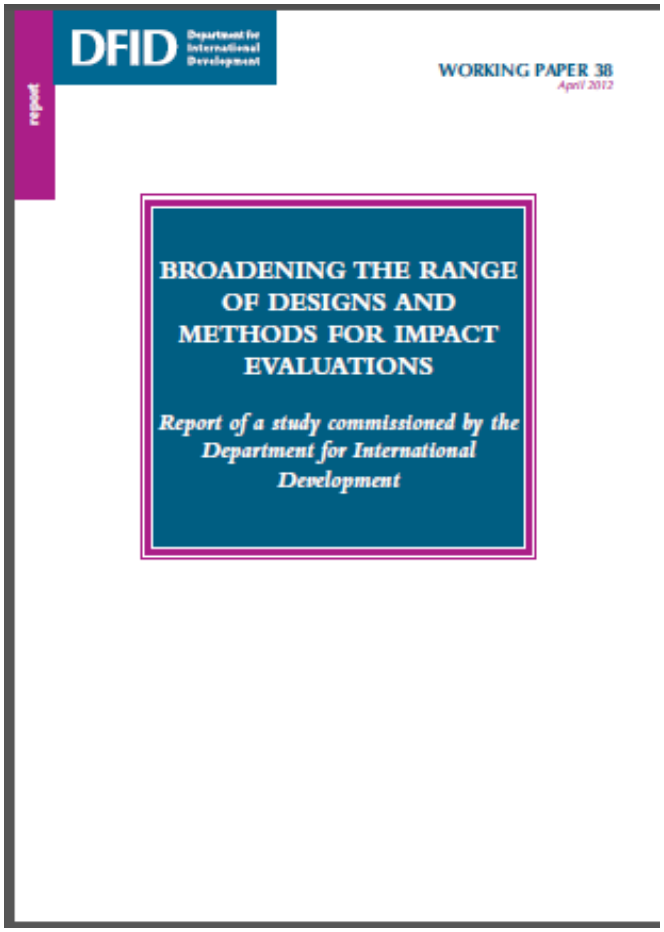
- Problemstillinger:
  - Forudsætter stadig "med" og "uden"
  - Præcision i forhold "hvem" og "hvad" vi sammenligner. Men:

  - Ikke altid muligt at afgrænse behandlings- og kontrolgruppen klart nok
    *Videnssprednin*g; *trickle down/ringe i vandet-indsatser*
  - Indsatsen er baseret på tilpasning til kontekst/modtagergruppe
    Komplicerede og komplekse indsatser/situationer
    Reformer, Kapacitetsudvikling, en del sociale indsatser
  - Når indsats og resultater er på system-niveau, og kompleksitet kun dårligt kan reduceres; helhedssyn og udforudsigelighed
    Indsatser i fht. fred og konflikt:
    Interdependens; kritisk masse, emergens

  - Der ud over: Også interesse for at se på, hvordan "hvorfor" og "hvordan" kan belyses – ikke kun om. Supplement eller selvstændig analyse.

Systematisk teoribaseret evaluering - variationer over tema

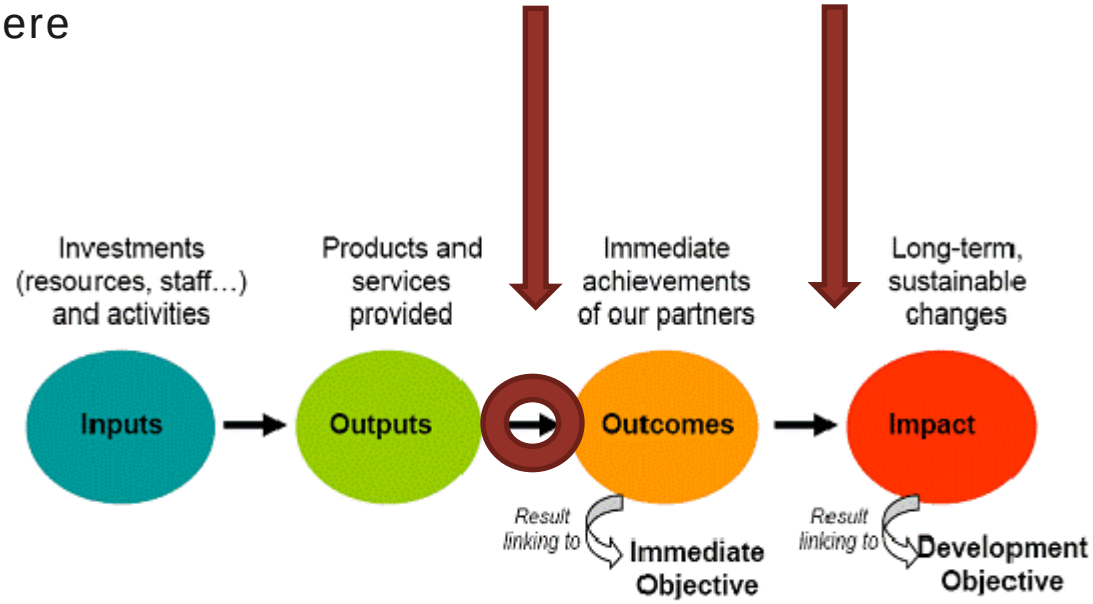




Udgangspunkt – Stern et al

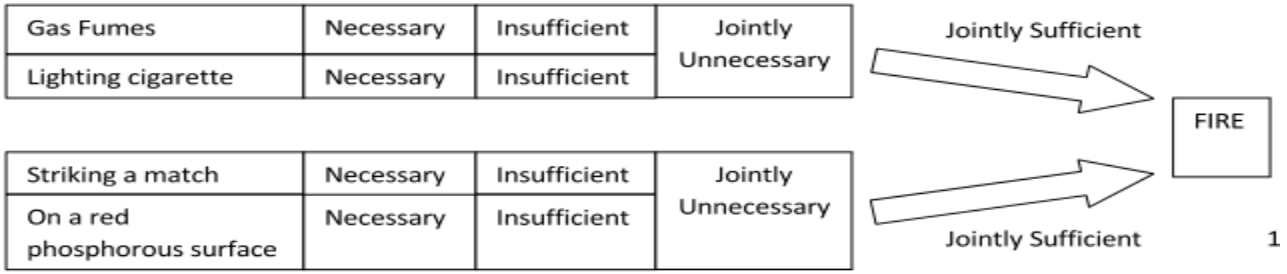
- Når sammenligning med en "alt andet lige" præmis ikke giver mening
- F.eks. når de kausale mekanismer ikke er helt lige til
- Når konteksten – og dens betydning – varierer.
- Når vi (også) vil vide mere om hvorfor og hvordan.

  - I bistandsverdenen: Logiske modeller udbredte – men programteori går videre



Kausale modeller:

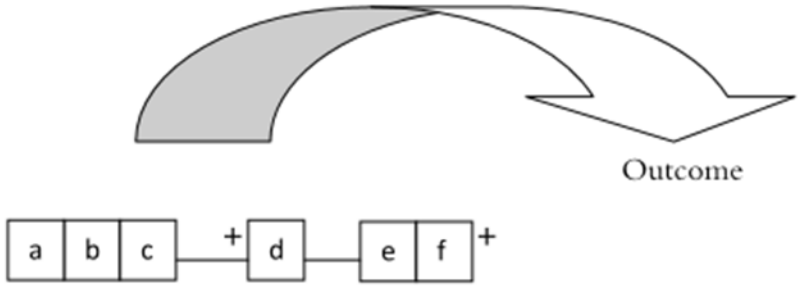
- A **forårsager** B – enkel, tilstrækkelig årsag
  - Vaccination giver immunitet
  - Myggenet reducerer risiko for malaria
    - Men allerede her begynder vi at **forudsætte/antage** en masse om både intervention og kontekst
- A kan være en **medvirkende faktor** til B, hvis…
  - A i sig selv hverken er tilstrækkelig eller nødvendig – men del af en "kausal pakke"



- Leder frem til udfoldet model/programteori

Konfigurationer og Kausalitet

Figure 2: Configurational causation

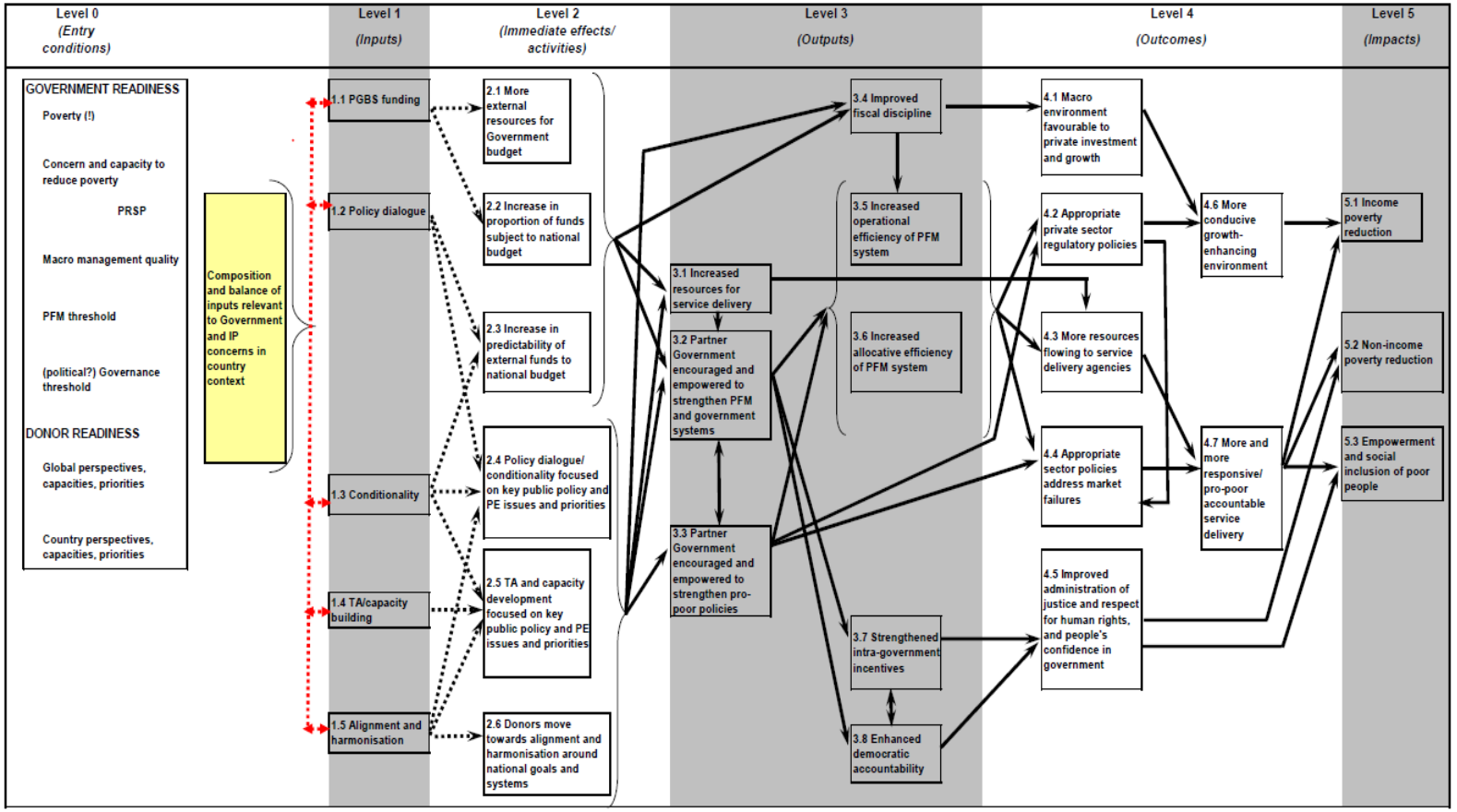


Books	Good teachers	Incentives	Attendance	Performance
YES	YES	YES	NO	LOW
YES	YES	YES	YES	HIGH
YES	NO	YES	YES	LOW
YES	YES	YES	YES	HIGH
NO	YES	YES	YES	LOW
YES	YES	YES	YES	HIGH
YES	YES	NO	YES	LOW
YES	YES	YES	YES	HIGH

Kilde: Bafani i Stern et al (2012).

Lad os lige se et virkeligt eksempel… Budgetstøtte – Uganda case study

Figure A1.1: Causality Map for the Enhanced Evaluation Framework



Risici for bias? White et al.

- Udgangspunkt: RCT'er og statistiske analyser med "**large n**" kan ikke levere det hele…
- Mener samtidig at der i praksis ofte er plads til forbedring i **"small** n" analyser; f.eks. Single-Case-studier:
  - Projektbias, deltager-perception, H-effekter,
    - For lidt blik for andre forklaringer, risiko for overvurdering af egen betydning og effekt.

- Her ses de teoribaserede tilgange som en mulighed for at:
  - Undersøge **mekanisme**r; ikke antage virkning, men kritisk analyse, hypoteser osv.
  - Specifikt inddrage "**hvad skete der ellers**"; kontekstændringer, andre aktører, tilpasninger osv.
  - Ofte **blandede metoder**, kval og kvant dækning osv.

Nogle forskellige tilgange

- Tydeliggøre antagelser om sammenhænge – hvad gør indsatsen, hvilke mekanismer skal i spil og hvad betydning har forhold i omgivelserne , for at resultater opnås

- **Kontekst – Mekanisme – Virkning / CMO konfigurationer**
- **Kontrafaktisk tænkning – ikke ren med og uden**

- Realist evaluation; Realist synthesis (Pawson and Tilley)
  - Udvikle/opstille ”mid-range programteori” og undersøge hvad virker for hvem, hvordan, under hvilke omstændigheder…

- **General Elimination Methodology**/Modus Operandi Method (Scriven)
  - Identificere, undersøge, udelukke forskellige kausale forklaringer.

- **Contribution Analysis** (Mayne):
  - Opstille og undersøge programteorien – og andre evt. plausible forklaringer; opbygge” credible contribution story””

- **Qualitative Comparative Analysis** (Rihoux; Befani og andre)
  - Fyldig operationalisering af programteori
  - Systematisk sammenligning på tværs af en række forskellige konfigurationer
  - Ikke kun ”small n”.

Så – Potentiale og relevans? Ja!

Vi kan estimere effekter solidt – også når vi ikke kan bruge RCT

- Velovervejet, model/teoribaseret brug af eksisterende data *kan* muliggøre dif-in-dif/kontrafaktisk kvantitativ effekt-estimering

Teoribaseret evaluering kan styrke vores viden om virkning

- Også når "ren" kontrafaktisk sammenligning ikke er mulig
- Når det er en pointe at alt IKKE er lige
- Vi får ikke et "gennemsnits effektestimat" eller **** signifikans
  - Men viden om, hvornår og hvordan det virker
- Potentiale for styrket anvendelighed/transfer!
- Bidrager til styrkelse af case-baseret "small n" evaluering

Styrket brug af eksisterende data + bred metodebevidsthed muliggør frugtbart blandede metoder

”The Peace Dividend”

Metodekrigen er ovre

- Så skal vi bare ha’ plukket fredens frugter!

Litteraturliste

Andersen, Ole Winckler, Eva Broegaard and Jens Anders Kovsted (2012): Hjælper vi? En introduktion til evaluering af udviklingsbistand. DJOEF.

Andersen, Ole Winckler, Beate Bull and Megan Kennedy-Chouane (eds) (2014, forthcoming): Evaluation methodologies for aid in conflict. Rourledge.

Andersen, O. W., Eva Broegaard & Ninja Kleinstrup: En effektiv udviklingsbistand: Om evidens, "randomistas" og relevans. I Samfundsøkonomen, april 2011 – no 2.

Befani, Barbara et al. (2007): Realistic Evaluation and QCA: Conceptual Parallels and an Empirical Application. *Evaluation* ,April 2007.13(2),.

Broegaard, Eva; Ted Freeman & Carsten Schwensen: Experience from a Phased Mixed-Methods Approach to Impact Evaluation of Danida Support to Rural Transport Infrastructure in Nicaragua. In *Journal of Development Effectiveness, vol 3, no. 1.*Francis & Taylor, 2011.

Support to Rural Transport Infrastructure in Nicaragua. In Hansen, Henrik, O. W. Andersen & H. White (red): *Impact Evaluation of Infrastructure Interventions.* Routledge.

Centre for Global Development (2006): *When will we ever learn? Improving lives through impact evaluation.* Report, Centre for Global Development, Washington.

Cook, T.D.; Shadish, W.T. and Wong, V.C, (2008): Three conditions under which experiment and observational studies produce comparable causal estimates: new findings from within study comparisons. *Journal of policy analysis and management*, 27 (4).

Flyvbjerg, B (2011): Case Study. In Denzin, NK and Lincoln, Y.S. (eds): The Sage Handbook of Qualitative Research. Sage, Thousand Oaks.

Forss, K.; Marra, M. & Schwartz, R (eds) (2011): "*Evaluating the Complex: Attribution, Contribution, and Beyond*". Comparative Policy Evaluation, Vol. 18., Transaction Publishers, New Jersey.

Hansen, Henrik, Ole Winckler Andersen and Howard White (eds) (2012). *Impact Evaluation of Infrastructure Interventions*. Routledge.

Hansen, Klejnstrup and Andersen (2013) A Comparison of Model-Based and Design-Based Impact Evaluations of Interventions in Developing Countries, American Journal of Evaluation, 34: 320-338

Mayne, J. (2011) "Contribution analysis: Addressing cause and effect", chapter 3 in: " Forss, K.; Marra, M. & Schwartz, R (eds) "*Evaluating the Complex: Attribution, Contribution, and Beyond*", Comparative Policy Evaluation, Vol. 18., Transaction Publishers, New Jersey

Pawson, R. and N. Tilley (1997) "*Realistic Evaluation*", Sage, London.

Rodrik, D (2008): *The New Development Economics: We Shall Experiment, But How Shall We Learn*? Harvard.

Rogers (2008): Using Programme Theory to Evaluate Complicated and Complex Aspects of Interventions, *Evaluation,* 14(1): 29-48.

Rogers, P.J. (2009) 'Learning from the evidence about evidence-based policy'. In Productivity Commission (ed), *Strengthening Evidence-Based Policy in the Australian Federation*.

Rogers and Funnell (2011): *Purposeful Program Theory: Effective use of theories of change and logic models.* John Wiley & sons.

Stern, E. et al (2012).*Broadening the Range of Designs for Impact Evaluation*. Working Paper 38, DFID, London.

Vogel, Isobel (2012*): Review of the use of "Theory of Change" in international development*. Review Report for DFID, UK, April 2012. ,

White and Phillips (2012): Addressing attribution of cause and effect in small n impact evaluations. Towards and integrated framework. *Working paper 15. International Initiative for Impact Evaluation*. New Delhi.

Også når kontrafaktisk sammenligning ikke er mulig