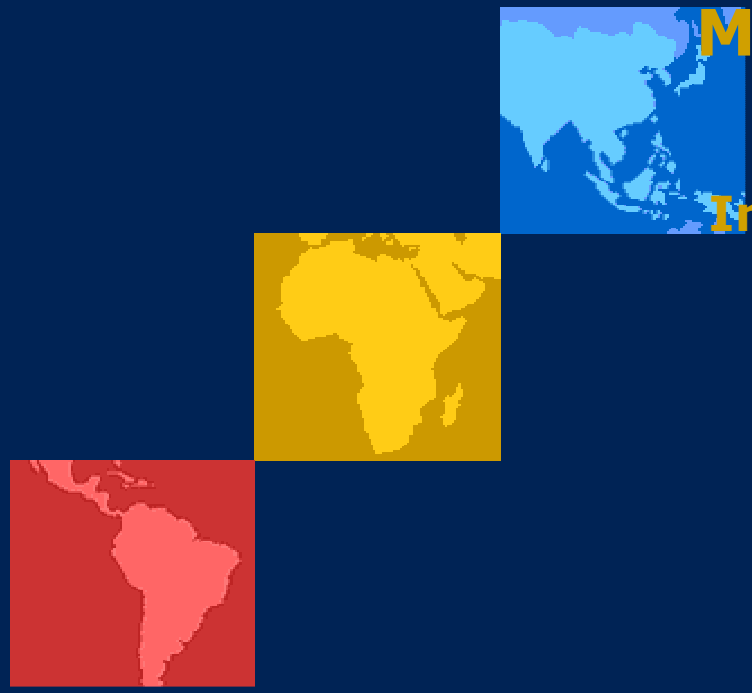
IOB
INSTITUTE OF DEVELOPMENT POLICY AND MANAGEMENT

University of Antwerp



Impactevaluatie. Mogelijkheden en aandachtspunten

Nathalie Holvoet
Instituut voor Ontwikkelingsbeleid en -beheer

Vlaams Evaluatie Platform
sub-groep Evaluatie in Ontwikkelingssamenwerking

1 april 2010

Overzicht

1. Inleiding
2. '**Impact**'evaluatie: what's in a name?
3. Experimentele benadering
4. Quasi-experimentele benaderingen
5. Niet-experimentele benaderingen
6. Bezint vooraleer je begint
7. Referenties

1. Inleiding (1)

- toenemend belang/**vraag** M&E in OS (cf. NAA)
  - 'evidence-based' beleid
  - resultaatsgericht beheer en budgettering
  - 'downward' accountability
- **aanbod** M&E is laag
  - monitoring > evaluatie
    - ✓descriptief > analytisch
    - ✓definitie 'evaluatie'
    - →'zo *systematisch* en *objectief* mogelijk'
    - →gebruik bestaande methodologie uit sociale wetenschappen

1. Inleiding (2)

- (impact)evaluatie: 'publiek goed' kenmerken
- tekort aan **'valide'** (impact)evaluatie
  - ✓ teleurstellende resultaten van 'meta-evaluaties' (zie Center for Global Development, 2006)
  - ✓ potentieel negatieve effecten van niet-valide impactevaluaties

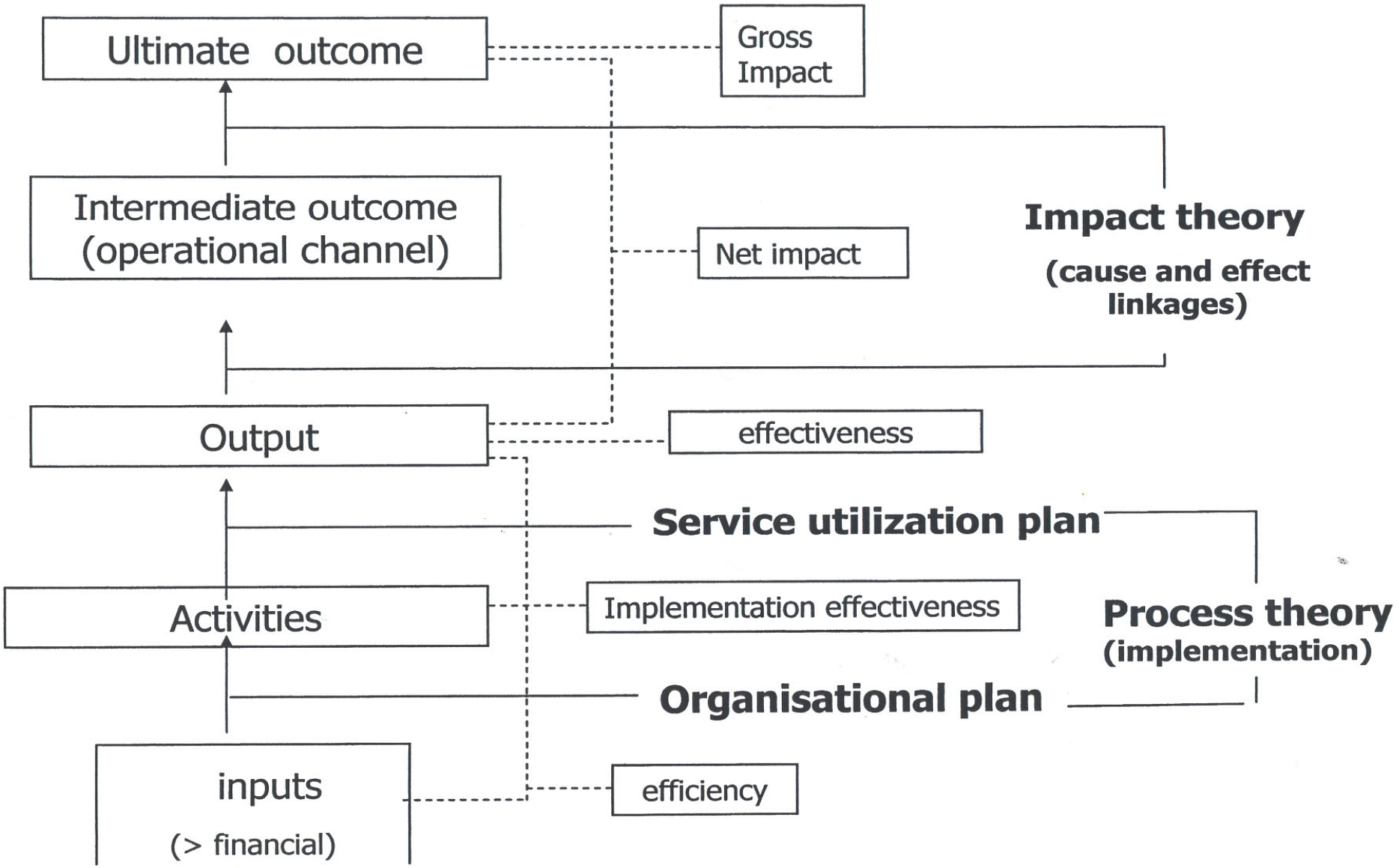
2. 'Impact'evaluatie: what's in a name? (1)

- **impact**evaluatie ≠ **proces**evaluatie (see slide 6)
- **voorbeeld**

Je bezoekt een microkredietproject en op basis van interviews blijkt dat het gemiddelde maandinkomen van de vrouwen die een jaar geleden krediet ontvingen, gedaald is van 1000 shilling naar 900 shilling. Uit interviews met een groep vrouwen die net krediet hebben ontvangen blijkt dat hun gemiddeld maandinkomen over dezelfde periode gedaald is van 1100 shilling naar 950 shilling.
Wat is de impact van het kredietprogramma op het gemiddelde maandinkomen van de participanten?

a. -100
b. +50
c. -50
d. +900

program theory evaluation and evaluation criteria



2. impactevaluatie: what's in a name? (2)

	'interventie' groep	'controle' groep
Voor interventie	E_1	C_1
Na interventie	E_2	C_2

$E_2 - E_1$ = bruto-impact

$(E_2 - E_1) - (C_2 - C_1)$= (netto) impact (1) (eenvoudig)

$(E_2 - E_1)/E_1 - (C_2 - C_1)/C_1$ = (netto) impact (2) (proportioneel)

2. impactevaluatie: what's in a name? (3)

- '**causaliteit**' → **interne validiteit** (Shadish, Cook and Campbell, 2002)

  - **interne validiteit**

    betrouwbaarheid van een conclusie/gevolgtrekking over het verschil die een interventie maakt in een specifieke context

  - **externe validiteit**

    betrouwbaarheid van een conclusie/gevolgtrekking over de veralgemening (generalisering) van dat effect tot andere populaties, contexten, andere operationaliseringen van de onafhankelijke en afhankelijke variabelen

2. Impactevaluatie: what's in a name? (4)

Statistische conclusie-validiteit (*statistical conclusion validity*)	validiteit van conclusie over de **correlatie** tussen een interventie (onafhankelijke variabele) en een resultaat (afhankelijke variabele)
Interne validiteit (enge zin)	validiteit van de conclusie over **causaliteit** tussen een interventie (onafhankelijke variabele) en een resultaat (afhankelijke variabele)
Constructie-validiteit	validiteit van de conclusie over relatie tussen een operationele variable en de 'constructie' (**representatie**)
Externe validiteit (enge zin)	validiteit van de conclusie of een causale relatie ook blijft gelden bij variatie over personen, settings, etc. (**extrapolatie**)

2. Impactevaluatie: what's in a name? (5)

- uitspraken over validiteit: niet absoluut
- perfecte validiteit: onmogelijk
- trade offs tussen ≠ validiteitstypes
  - interne versus externe validiteit
- ≠ valkuilen/bedreigingen per validiteitsprincipe gekend + manieren om deze te beperken (zie annex voor overzicht)

➢ per evaluatie:
  - oplijsten van belangrijkste valkuilen
  - keuzes maken inzake in te perken valkuilen
  - ✓ specifiek evaluatie-objectief (interne & externe validiteit)
  - ✓ beperkingen tijd, middelen, data

    →cf. '*shoestring evaluation*' (see Bamberger et al., 2004)

2. impactevaluatie: what's in a name? (6)

- ≠ bedreigingen voor interne validiteit
  - omgekeerde causaliteit ('*reversed causality*')
  - andere interventies ('interfering events')
  - seculiere tendenzen ('maturation of communities')
  - natuurlijke effecten van veroudering ('maturation of persons')
  - selectie (zelfselectie, 'creaming', uitval)
  - 'test'-effect
  - 'instrument'-effect

➢ hoe alternatieve verklaringen reduceren? (netto-impact distilleren uit bruto-impact)

➢ ≠ mogelijkheden met verschillende vereisten en kost

3. Experimentele benadering

- toevallige toewijzing aan 'interventie'groep en 'controle'groep (*randomisation*)

- $E_1 = C_1$ dan netto-impact: $E_2 - C_2$

➢ in principe: enkel behoefte aan ex-post data, maar meestal ook pre-interventiedata

- first best voor interne validiteit: ↓ zelfselectie
- maar vaak moeilijk of niet-wenselijk bij ontwikkelingsprojecten
  - probleem met full-coverage interventies
  - niet wenselijk omwille van morele problemen
  - duur + afwezigheid bij start van de interventies
  - experiment ≠ echte interventie
  - kan niet alle problemen van interne validiteit oplossen

4. Quasi-experimentele benaderingen (1)

- second-best, poging om interne validiteit te verhogen via
  1. constructie controlegroep
  2. statistische technieken (tijdens analyse van data)
  3. combinatie van 1&2

4.1. constructie controlegroep

- **Matching** ('cross-section')
  - constructie 'controle' groep op basis van karakteristieken interventiegroep
  - ! keuze variabelen voor matching
  - individuele matching: niet noodzakelijk (zie vb. slide 14)

EXHIBIT 9-C Evaluation of a Family Development Program Using Aggregate-Matched Controls

A program was started in Baltimore to serve poor families living in public housing by providing integrated services with the hope of helping families escape from long-term poverty. Services included access to special educational programs for children and adults, job training programs, teenage programs, special health care access, and child care facilities. To the extent possible, these services were delivered within the LaFayette Courts public housing project. Case managers assigned to the housing project helped families choose services appropriate to them. The special feature of this program was its emphasis on serving families rather than individuals. In all, 125 families were enrolled.

To constitute a comparison group, 125 families were chosen from a comparable public housing project, Murphy Homes. The impact of the family development program was then assessed by contrasting the enrolled families with the Murphy Homes sample. After a year of enrollment, the participating families were shown to be higher in self-esteem and sense of control over their fates, but positive impacts on employment and earnings had not yet occurred.

SOURCE: Adapted from Anne B. Shlay and C. Scott Holupka, *Steps Toward Independence: The Early Effects of the Lafayette Courts Family Development Center* (Baltimore: Johns Hopkins University, Institute for Policy Studies, 1991).

4. quasi-experimentele benaderingen (2)

- **regression-discontuinity**
  - gebruik van een duidelijke selectie-variabele voor indeling contrôle en interventiegroep
  - ~ experiment

- **generische controles**
  - controles= voorafbepaalde normen ivm. verwachte resultaten
  - bv. antropometrische indicatoren, geboortecijfers, sterftecijfers, sex ratios, ...

4. quasi-experimentele benaderingen (3)

- **voor/na** (‘*reflexive*’ controls)
  - éénvoudige V/N
    - ✓ zware veronderstelling & veel ‘valkuilen’ voor interne validiteit
    - ✓ best alleen voor ‘impact’ assessment over KT (zie bv. slide 17)
  - panelstudies
    - ✓ verschillende observaties voor en na
    - ✓ interessant in gevallen waar verschillende eenheden op verschillende manier worden blootgesteld aan interventie (zie bv. Slide 18)

EXHIBIT 10-C A Convincing Pre-Post Outcome Design for a Program to Reduce Residential Lead Levels in Low-Income Housing

The toxic effects of lead are especially harmful to children and can impede their behavioral development, reduce their intelligence, cause hearing loss, and interfere with important biological functions. Poor children are at disproportionate risk for lead poisoning because the homes available to low-income tenants are generally older homes, which are more likely to be painted with lead paint and to be located near other sources of lead contamination. Interior lead paint deteriorates to produce microscopic quantities of lead that children may ingest through hand-to-mouth activity. Moreover, blown or tracked-in dust may be contaminated by deteriorating exterior lead paint or roadside soil containing a cumulation of lead from the leaded gasoline used prior to 1980.

To reduce lead dust levels in low-income urban housing, the Community Lead Education and Reduction Corps (CLEARCorps) was initiated in Baltimore as a joint public-private effort. CLEARCorps members clean, repair, and make homes lead safe, educate residents on lead-poisoning prevention techniques, and encourage the residents to maintain low levels of lead dust through specialized cleaning efforts. To determine the extent to which CLEARCorps was successful in reducing the lead dust levels in treated urban housing units, CLEARCorps members collected lead dust wipe samples immediately before, immediately after, and six months following their lead hazard control efforts. In each of 43 treated houses, four samples were collected from each of four locations—floors, window sills, window wells, and carpets—and sent to laboratories for analysis.

Statistically significant differences were found between pre and post lead dust levels for floors, window sills, and window wells. At the six-month follow-up, further significant declines were found for floors and window wells, with a marginally significant decrease for window sills.

Since no control group was used, it is possible that factors other than the CLEARCorps program contributed to the decline in lead dust levels found in the evaluation. Other than relevant, but modest, seasonal effects relating to the follow-up period and the small possibility that another intervention program treated these same households, for which no evidence was available, there are few plausible alternative explanations for the decline. The evaluators concluded, therefore, that the CLEARCorps program was effective in reducing residential lead levels.

SOURCE: Adapted from Jonathan P. Duckart, "An Evaluation of the Baltimore Community Lead Education and Reduction Corps (CLEARCorps) Program," *Evaluation Review*, 1998, 22(3):373-402.

The advantage of panel studies is that the measures of the intervention and outcomes (e.g., TV viewing and aggressiveness, respectively) are related to each other through time lags and not as cross-sectional correlations. Thus, aggressiveness at Time 2 is examined as a function of viewing patterns measured at Time 1. Panel studies are especially appropriate for impact assessments of full-coverage programs whose dosage varies over individuals and over time. In the case of TV viewing, all the children participated in the sense that virtually

EXHIBIT 10-D Measuring the Effects of TV Violence on Children's Aggressive Behavior

In an attempt to provide rigorous answers to public concern over whether the viewing of TV programs depicting violence and aggression affect children's aggressive behavior, the National Broadcasting Company (NBC) sponsored an elaborate panel study of young children in which aggressiveness and TV viewing were measured repeatedly over several years.

In the main substudy, samples of elementary school classes, Grades 2 through 6, drawn from Fort Worth and Minneapolis schools, formed the base for a six-wave panel study, in which 400 male children in 59 classes were interviewed six times in the period 1970 to 1973. (Additional substudies were conducted with female elementary school children and with samples of high school students in the same cities.) At each interview wave, the children in the classes were asked to rate each other on aggressiveness using questionnaires that included such items as "Who is likely to punch and kick another child?" The questionnaires also picked up information about the socioeconomic background of the children. In addition, at every interview, the children were each asked to check those programs they had watched recently on lists of programs shown locally. The programs previously had been rated by media experts according to the amount of violence depicted in them. To check the accuracy of recall, several nonexistent programs were placed on the checklists. Additional interviews were conducted with the children's teachers and parents.

The analyses undertaken related the viewing of violence on TV at one interview time with rated aggressive behavior at subsequent interview times, controlling statistically for the initial level of the children's aggressiveness. The results estimated the additional amount of aggressiveness that resulted from high levels of exposure to violence on TV programs. While the direction of effects indicated a small increment in aggressiveness associated with high levels of viewing of TV violence, that increment was not statistically significant.

SOURCE: Adapted from J. R. Milavsky, H. H. Stipp, R. C. Kessler, and W. S. Rubens, *Television and Aggression: A Panel Study* (New York: Academic Press, 1982).

all viewed some TV. Nevertheless, some children viewed more programs containing violence than others and some watched more such programs at some times than at other times. Self-selection was to some degree controlled by statistically controlling the initial level of aggressiveness of the children under study.

It should be noted that the researchers in this study considered using randomized experiments to estimate the effects of viewing violent programs on subsequent aggressiveness but rejected that design as introducing an artificiality that would undermine the generalizability of their findings. It would be difficult, if not impossible, to recruit schoolchildren for experimentation, randomly allocate them to experimental and control groups, and then somehow prevent the controls from viewing any programs that contain aggressive or violent behavior. An experiment along those lines might be conducted for a very short period of time, on the order of a few days, but would be extremely

4. quasi-experimentele benaderingen (4)

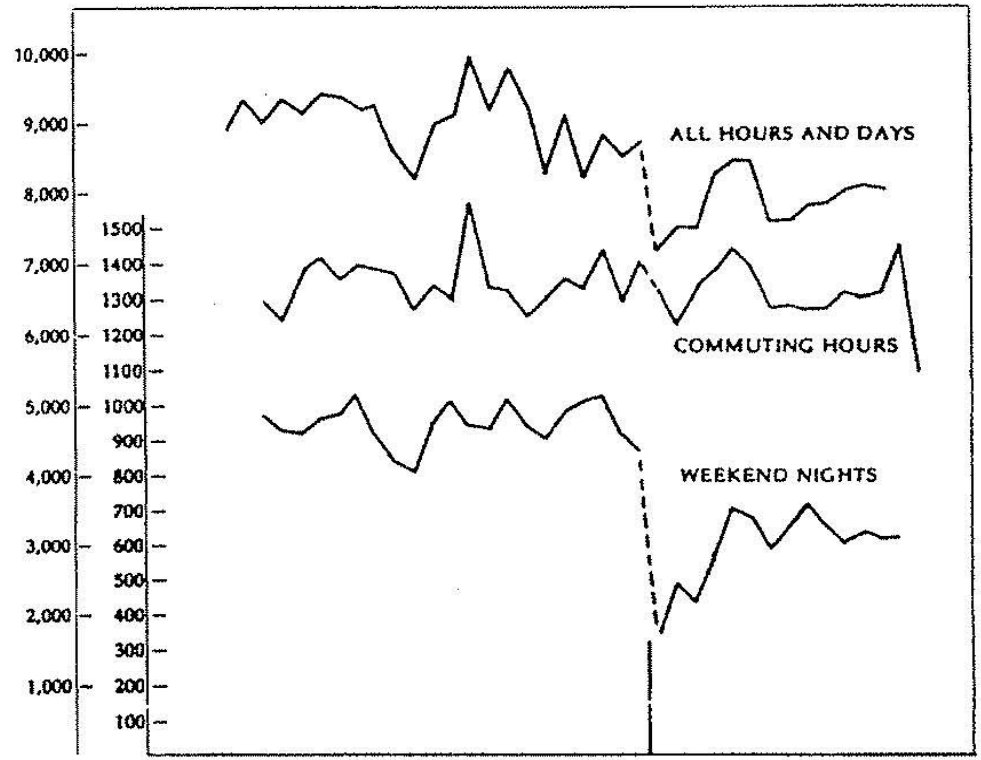
- tijdseries
  - ✓ veel observaties voor (min. 30) en na
  - ✓ geaggregeerde eenheden
  - ✓ trendanalyse op basis van observaties vóór interventie + projectie in de toekomst (= controle-groep)
  - ✓ gebruik van grafische methodes (zie voorbeeld slide 20)
  - ✓ first best om impact van full-coverage programma's te evalueren (eventueel in combinatie met cross-sectional)

EXHIBIT 9-1

An Analysis of the Impact of Compulsory Breathalyzer Tests on Traffic Accidents

In 1967, the British government enacted a new policy that allowed police to give Breathalyzer tests at the scenes of accidents. The test measured the presence of alcohol in the blood of suspects. At the same time, heavier penalties were instituted for drunken driving convictions. Considerable publicity was given to the provisions of the new law, which went into effect in October 1967.

The chart below plots vehicular accident rates by various periods of the week before and after the new legislation went into effect. Visual inspection of the chart clearly indicates that a decline in accidents occurred after the legislation, which affected most times of the week but had especially dramatic effects for weekend periods. Statistical tests verified that these declines are greater than could be expected from the chance component of these data.



SOURCE: Summary of H. L. Ross, D. T. Campbell, and G. V. Glass, "Determining the Social Effects of a Legal Reform: The British Breathalyzer Crackdown of 1967." *American Behavioral Scientist*, 1970, 13 (March/April): 494-509.

4. quasi-experimentele benaderingen (5)

4.2. statistische controles (‘*statistically equated controls*’)

- controle tijdens analyse
- zelfde effect als matching (zie vb. slide 22)
- meting ‘effect’ nauwkeurig
- beperkt tot ‘gekende’ en ‘gemeten’ variabelen
- verschillende technieken
  - multiple regressie (zie bv. slide 23)
    - ✓ regressie met afhankelijke variabele (‘resultaat’), onafhankelijke variabele (‘interventie’ dummy), controlevariabelen
    - ✓ effect wordt gemeten via regressie-coëfficient van onafhankelijke variable
  - two-stage regressie (zie bv. slide 24-25)
    - ✓ modelleren selectie (afhankelijke variabele: ‘selectie’) (slide 24)
    - ✓ modelleren resultaat (afhankelijke variabele: ‘resultaat’) (slide 25)

4.3. combinaties van 4.1 en 4.2.

EXHIBIT 9-E
Simple Statistical Controls in an Evaluation of the Impact of a Hypothetical Employment Training Project

I. Outcome comparison between men 35-40 who completed the training program and a sample of men 35-40 who did not attend the program

	Participants	Nonparticipants
Average wage rate	$7.75	$8.20
n =	1,000	1,000

II. Comparison after adjusting for educational attainment

	Participants		Nonparticipants	
	Less Than High School	High School	Less Than High School	HighSchool
Average wage rate	$7.60	$8.10	$7.75	$8.50
n =	700	300	400	600

III. Comparison adjusting for educational attainment and employment at the start of the training program (or equivalent data for nonparticipants)

	Participants		Nonparticipants			
	Less Than High School Unemployed	High School Unemployed	Less Than High School Unemployed	Less Than High School Employed	High School Unemployed	High School Employed
Average wage rate	$7.60	$8.10	$7.50	$7.83	$8.00	$8.60
n =	700	300	100	300	100	500

program were sampled from the same metropolitan area and also interviewed at the time the program started and one year after it ended. The men in both samples were asked about their current earnings, and hourly wage rates were computed.

In Panel I of Exhibit 9-E, the average posttraining wage rates of the two groups are compared without application of any statistical controls. Those who had participated in the project were earning an average of $7.75 per hour; those who had not participated, $8.20. Clearly, participants were earning less than nonparticipants; had this been the outcome of a randomized experiment, the difference would have been an unbiased estimate of the program effect. To the extent that participants and nonparticipants

EXHIBIT 9-F
Estimating the Effect of AA Attendance Using Regression Modeling

Does attendance at Alcoholics Anonymous (AA) meetings affect the drinking of individuals who have problems with alcohol? It is part of the AA philosophy that the problem drinker must make a voluntary commitment to participation, so self-selection becomes part of the intervention. Thus, any attempt to assess impact by comparing problem drinkers who attend AA with those who do not must deal with selection bias related to the natural differences between these groups. To attempt to equate the groups through statistical controls, researchers in the Palo Alto Veterans Affairs Health Care System used several approaches, one of which was a simple multiple regression model.

First, consideration was given to what variables might be related to AA participation. Based on prior research, three variables were identified—perceived seriousness of drinking, tendency to cope with problems by seeking information and advice, and sex. Two other control variables were selected because of their known relationship to drinking outcomes—baseline drinking scores and marital status. The outcome variable of interest was the amount of drinking measured on a Drinking Pattern scale.

These variables were measured on a sample of 218 individuals with drinking problems and used in a regression model with drinking outcome as the dependent variable and the other variables as independent (predictor) variables. The intervention variable, AA attendance (0 = no, 1 = yes), was also included as a predictor to assess its relation to the outcome when the other predictor variables were statistically controlled.

As shown in the summary below, two of the variables showed significant relationships to outcome, including the intervention variable, AA attendance. The significant negative coefficient for AA attendance indicates that those attending AA drank less at outcome than those not attending, controlling for the other variables in the model. To the extent that those other variables in this statistical model completely controlled for selection bias, the unstandardized regression coefficient shown for AA attendance estimates the program effect on the Drinking Pattern outcome variable.

Regression Results Predicting Drinking Outcome

Predictor Variable	Coefficient	Standard Error
Sex (0 = M, 1 = F)	−1.16	1.09
Information seeking	−.04	.12
Perceived seriousness of drinking	−.44	.57
Baseline drinking	.20*	.09
Married (0 = no, 1 = yes)	−1.69	1.25
AA attendance (0 = no, 1 = yes)	−2.82*	1.15
R^2 = .079		

*Statistically significant at $p \leq .05$.

SOURCE: Adapted with permission from Keith Humphreys, Ciaron S. Phibbs, and Rudolf H. Moos, "Addressing Self-Selection Effects in Evaluations of Mutual Help Groups and Professional Mental Health Services: An Introduction to Two-Stage Sample Selection Models." *Evaluation and Program Planning*, 1996, 19(4):301-308.

EXHIBIT 9-G
Estimating the Effect of AA Attendance Using Two-Stage Selection Modeling

By estimating selection effects separately from influences on the outcome variable, two-stage selection modeling has the potential to produce a better estimate of the effects of AA attendance than the one-stage multiple regression analysis presented in Exhibit 9-F. Three of the variables available to the researchers were expected to predict AA participation—perceived seriousness of drinking (those who believe their drinking is a problem are presumed more likely to participate), tendency to cope with problems by seeking information and advice, and sex (women are presumed more likely to seek help than men). These variables were used in the first-stage analysis to predict AA attendance rather than being included in a one-stage model predicting drinking outcome. For this application, the researchers used the Heckman procedure and fit a probit regression model to predict AA participation. As shown in the summary below, two of the variables showed significant independent relationships to attendance.

Stage 1: Probit Regression Predicting AA Attendance

Predictor Variable	Coefficient	Standard Error
Sex (0 = M, 1 = F)	.29	.19
Information seeking	.06*	.02
Perceived seriousness of drinking	.38*	.09
R^2 = .129		

*$p \leq .05$.

This selection model was then used to produce a new variable, Lambda, which estimates the probability that each individual will be in the intervention versus the control group. Lambda is then entered as a control variable in a second-stage regression analysis that attempts to predict the outcome variable, amount of drinking measured on the Drinking Pattern scale. Two outcome-related control variables were also included at this stage—baseline drinking scores and marital status. Finally, inclusion of the intervention variable, AA attendance (0 = no, 1 = yes), allowed assessment of its relation to the outcome when the other predictor variables, including the selection variable, were statistically controlled.

(Continued)

EXHIBIT 9-G
(Continued)

Stage 2: Least Squares Regression Predicting Drinking Outcome

Predictor Variable	Coefficient	Standard Error
Baseline drinking	.20*	.08
Married (0 = no, 1 = yes)	−1.68	1.23
Lambda	2.10	1.98
AA attendance	−6.31*	3.04
R^2 = .084		

*$p \leq .05$.

The significant coefficient for AA attendance shows that those participating drank less at outcome than those not attending, controlling for baseline drinking and self-selection. Indeed, on the 30-point Drinking Pattern scale, the estimated net effect of AA attendance was a reduction of more than 6 points. Note also that using the two-stage model indicates that the effect of AA attendance is nearly twice as large as the estimate derived in the earlier example using a one-stage regression model.

SOURCE: Adapted with permission from Keith Humphreys, Ciaran S. Phibbs, and Rudolf H. Moos, "Addressing Self-Selection Effects in Evaluations of Mutual Help Groups and Professional Mental Health Services: An Introduction to Two-Stage Sample Selection Models." *Evaluation and Program Planning*, 1996, 19(4):301-308.

now, that instead of trying to figure out what variables were related to selection, the evaluator was given the selection variable up front and could apply it case-by-case to allocate individuals into the intervention or control group according to their scores on that variable. In this circumstance, selection modeling should be a sure thing because there would be no uncertainty about how selection was done and the evaluator would have in hand the measured values that determined it.

A special type of constructed control group design, referred to as a **regression-discontinuity design,** is based on this concept. When this design is applicable, it generally provides less biased estimates of program effects than any of the other quasi-experimental impact assessment designs. Regression-discontinuity designs are appropriate for circumstances when the evaluator cannot randomly assign targets to intervention and control groups but could collaborate with program personnel to divide them systematically on the basis of need, merit, or some other qualifying condition and assign the neediest, most meritorious, and so forth to the intervention condition and those less needy or meritorious to the control condition.

5. niet-experimentele benaderingen

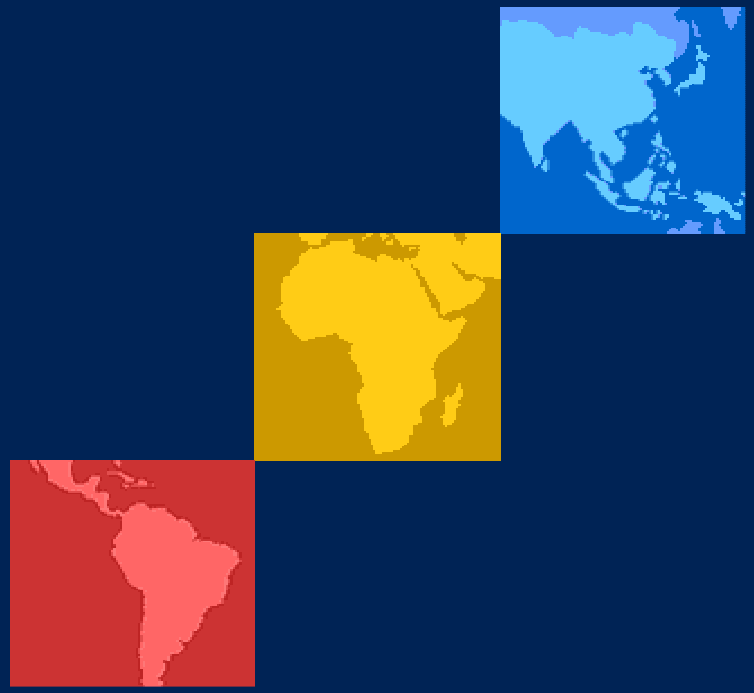
- Controle-groep: 'denkbeeldige' constructie door 'evaluator', op basis van:
  - auto-evaluaties door participanten
    - ✓ 'voldoening' participanten
    - ✓ 'empowering' effect
    - ✓ MAAR probleem van 'social desirable answering'
  - observatie tijdens veldbezoeken
  - interviews met lokale key-persons
  - archieven van projecten (data betreffende participanten, project, financiële data)
  - evaluatie door staf project
    - ✓ goede kennis van project implementatie
    - ✓ vaak overschatting van projectimpact
  - kennis, ervaring van evaluator
    - ✓ belang van kennis van het 'fenomeen' dat wordt geëvalueerd (> methodologie)

6. Bezint vooraleer je begint…

- doeltreffendheid interventie ok?
- innovatief element ?
- onvoldoende impactevaluaties? (check secondaire data, meta-evaluaties?)
- interventie kan herhaald worden ?
- voldoende investering van middelen?

7. Referenties

- Bamberger M., J. Rugh, M. Church and L. Fort (2004) "Shoestring evaluation: Designing Impact Evaluations under Budget, Time and Data Constraints", *American Journal of Evaluation*, vol. 25 (1): 5-37
- Center for Gobal Development (2006) *When will we ever learn? Improving Lives through Impact Evaluation*, Washington, Center for Global Development.
- Shadish, W., T.D. Cook and D. Campbell (2002) Chapters 2 & 3 (internal & external validity), pp. 37-94 in *Experimental and quasi-experimental designs for generalized causal inference*, Boston, Houghton Mifflin
- Rossi P.H., Freeman H.E. and Lipsey (1999) "Strategies for impact assessment", p. 235-278 in *Evaluation: A Systematic Approach,* London, Sage.



Dank

nathalie.holvoet@ua.ac.be

Annex: Threats to different validity types and how to lower them (see Shadish, Cook and Campbell, 2002)

STATISTICAL CONCLUSION VALIDITY

Threats (selection)	Ways of lowering (selection)
• Low statistical power ▪ ↑ false negative ▪ effect size estimates less precise	↑ sample size ↑ reliability of treatment implementation ↑ homogeneity of units ↑ reliability of measures
• unreliability of measures	• Improving quality of measures • Increasing the number of measurements
• restriction of range	• Use distinctly different treatment doses • Avoid floor and ceiling effects

Threats (selection)	**Ways of lowering (selection)**
• unreliability of treatment implementation ▪ but sometimes deliberate	↑ homogeneity of treatment implementation ↑ sample size
• extraneous variance in the experimental setting	• Control for 'distracting' factors • Measure sources of extraneous variance and use statistical controls in analysis
• heterogeneity of units (respondents)	↑ homogeneity on characteristics correlated with major outcomes ▪ ↓ external validity ▪ ↑ restriction of range
• inaccurate effect size estimation	Use appropriate tests

INTERNAL VALIDITY

Threats (selection)	**Ways of lowering (selection)**
• ambiguous temporal precedence	• experimental design
• selection ▪ self-selection ▪ creaming	• randomisation (experimental design) • two-stage statistical regression (Heckmann)
•history (interfering events)	• exposed & control group exposed to same history (= location) ↑ diffusion ↓ external validity

Threats (selection)	Ways of lowering (selection)
• maturation ▪ persons ▪ communities (secular trends)	• exposed & control group of same age, same location ↑ diffusion ↓ external validity
• regression artificats ▪ in particular when 'extreme scorers' were selected ▪ sometimes deliberate	• random assignment within the group of extreme scorers
• attrition (de-selection)	• early monitoring • ↓ response burden

Threats (selection)	Ways of lowering (selection)
• testing	• assessment of testing effect through specific designs • ↑ interval between tests
• instrumentation ▪ important in longitudinal design	• avoid switching instruments during studies

CONSTRUCT VALIDITY

Threats (selection)	Ways of lowering (selection)
• inadequate explication of constructs	• improve explication ▪ avoid constructs that are too general ▪ avoid using one construct to reflect more than one construct ▪ avoid wrong constructs
• mono-operation bias	• multiple operationalisation
• mono-method bias	• use of different methods
• confounding constructs with levels of constructs	• use several levels of treatment • specification of the level of treatment

Threats (selection)	Ways of lowering (selection)
• reactive self–report changes	• use external (not self-report) measures • use techniques that encourage accurate responding
• reactivity to the experimental situation	• make dependent variables less obvious (no pre-test) • reduce interactions of experimenter and participant • deception by using false hypotheses (ethical!) • use quasi-control participants • ensure confidentiality and anonymity

Threats (selection)	**Ways of lowering (selection)**
• novelty or disruption effects	• include 'innovation' and 'disruption' in the construct
• compensatory equalisation	• monitor treatment • interview staff, administrators
• compensatory rivalry	• unstructured interviews, direct observation to detect • avoid awareness about treatment (if ethical)
• resentful demoralisation	• unstructured interviews, direct observation to detect • avoid awareness about treatment (if ethical)

Threats (selection)	Ways of lowering (selection)
• treatment diffusion	• avoid physical proximity • avoid communication between exposed & control group • monitor and measure treatment implementation in both groups

EXTERNAL VALIDITY

Threats (selection)	Ways of lowering (selection)
• interaction of the causal relationship with units	• ↑ variability in units in present study + explicit testing for interaction (ex-ante) ▪ ↓ statistical conclusion validity • additional studies in other units + meta-evaluation (ex-post)
• interaction of the causal relationship over treatment variations	• ↑ variability in treatment ▪ ↓ statistical conclusion validity • additional studies for other treatment variations + meta-evaluation
• interaction of the causal relationship with outcomes	• ↑ variability in outcomes • additional studies for other outcomes + meta-evaluation

Threats (selection)	**Ways of lowering (selection)**
• interaction of causal relationship with settings	• ↑ variability in settings (multi-site studies) • additional tests in the other settings + meta-evaluation
• context-dependent mediation	• ↑ variability in contexts • additional tests in the other contexts