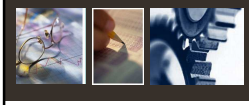
HOOFDSTUK 8

EXPERIMENTELE ONTWERPEN



Prof. dr. Henk Roose
UNIVERSITEIT GENT

Methodologie van de Sociale Wetenschappen, AJ 2019-2020
1ste bachelor Politieke & Sociale Wetenschappen

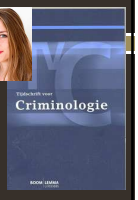
INLEIDING

- inzicht in methoden sociale wetenschappen?
- setting is debat in Tweede Kamer in Nederland 25/03/11
- hoofdrolspeler: Lilian Helder (PVV)
  - taakstraf
  - gevangenisstraf
  } recidive?

- > hoe vergelijk je op wetenschappelijk verantwoorde manier 'appels met peren'?

INLEIDING

- originele artikel uit 2009



Recidive na werkstraffen en na gevangenisstraffen

Een gematchte vergelijking*

Hilde Wermink, Arjan Blokland, Paul Nieuwbeerta & Nikolaj Tollenaar

Tabel 3: Gemiddelde jaarlijkse recidive na eerste werk-/gevangenisstraf, verschillende followupperioden

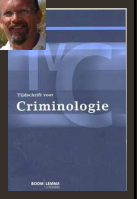
	Gemiddelde experimentele groep (N=2.123)	Gemiddelde controle-groep (N=2.123)	Absoluut verschil	t-stat	Sign.	Relatief verschil
1 jaar						
totaal	0,273	0,683	-0,410	-3,229	***	-0,60
vermogen	0,132	0,398	-0,266	-2,404	**	-0,67
geweld	0,044	0,109	-0,065	-2,255	**	-0,60
overig	0,097	0,175	-0,079	-3,698	***	-0,45

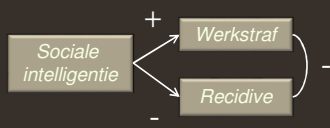
INLEIDING

Recidive, werkstraf en gevangenisstraf: een kritische bespreking

Frans Groenendijk & André van Delft

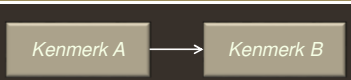
- kritiek 2013
  - causaliteit? > vinden samenhang
  - gemeenschappelijke oorzaak samenhang?
  - *vb.* 'sociale intelligentie'
- Lilian Helder toch gelijk?





SITUERING

- effectrelatie
  - samenhang
  - tijdsvolgorde



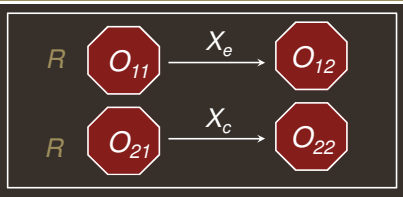
- niet zomaar causale conclusie trekken
  - neutraliseren alle andere mogelijke oorzakelijke factoren
  - vereist specifiek design: *experimenteel ontwerp*
- < natuurwetenschappen

NATUURWETENSCHAPPEN

- lange traditie
- Galileo Galilei (1564-1642), Francis Bacon (1561-1626)
  - "Nullius in verba" > empirie
  - actie en experiment verschaffen inzicht in Goddelijke creatie
  - Galilei en Toren van Pisa

SCHEMA KLASSIEK EXPERIMENT

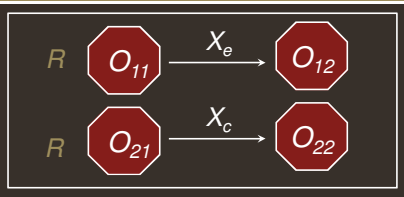
7



- voormeting: O_{11} en O_{21}
- nameting: O_{12} en O_{22}
- randomisering: R

SCHEMA KLASSIEK EXPERIMENT (2)

8



- verschillen na experimentele manipulatie: $O_{12} - O_{22}$
- minus verschillen vóór manipulatie: $O_{11} - O_{21}$
- *vb.* $(O_{12} - O_{22}) = 10,5$; $(O_{11} - O_{21}) = 2,7$
- *netto effect* = 10,5 – 2,7 = 7,8 (en niet 10,5)

SESAMSTRAAT

10

- onderdeel *Program Head Start*
  - Robert Kennedy, president Lyndon Johnson
  - < 'War on poverty'
  - gestart in 1965
- Sesamstraat tv-show voor kinderen
  - cognitieve en sociale vaardigheden aanscherpen
  - doelgroep: 3-5-jarigen

"One has only to listen to a child singing a television jingle...to realize that 'the tube' is teaching him something.... The medium is there; it is only the message which needs changing."

SESAMSTRAAT

11

- reacties in de pers...

THE SUNDAY TIMES

"Sesame Street, ...is in the vanguard of a television revolution. In a television season laden with mediocrity the series has been the one new creative jump, the show for the history books."
Norman Mark, Sunday Times Advertiser, Trenton, March 22, 1970.

The New York Times

"If the current television season has produced one undisputed hit, it is Sesame Street...."
Jack Gould, The New York Times, December 10, 1969.

SESAMSTRAAT (2)

12

- wat zegt wetenschappelijke literatuur?

Table 1: Experimental-minus-control differences in adjusted posttest mean scores showing favorable impact of Plaza Sésamo

	3-year-olds	4-year-olds	5-year-olds
General Knowledge	2.7	7.3***	4.8***
Numbers	4.4**	7.8***	6.2***
Letters and Words	2.9**	4.5***	5.1***
Relations	.5	1.6**	.7
Parts of the Whole	.5	.6	1.6*
Ability to Sort	.5	3.0***	3.2***
Classification Skills	2.0**	3.2***	2.3*
Embedded Figures	3.7	2.0*	.5
Oral Comprehension	3.1*	5.3**	2.4*

Bron: Diaz-Guerrero *et al.* (1976), "Plaza Sésamo in Mexico: an Evaluation," *Journal of Communication*, 26:146-154.

13

SAMENGEVAT: KLASSIEK EXPERIMENT

- toevalstoewijzing van eenheden aan condities
- het al dan niet optreden van de vermoedelijke oorzaak wordt door de onderzoeker gemanipuleerd
- voormeting om eventuele reeds bestaande verschillen in rekening te brengen
- nameting om bruto-effect van manipulatie na te gaan
- netto-effect: verschil in nameting minus verschil in voormeting

14

ENKELE ASPECTEN VAN NADERBIJ

- randomiseren en matchen
- voormeting
- geldigheid
  - externe
  - interne

15

RANDOMISEREN EN MATCHEN

- randomiseren beste strategie
- MAAR: kan niet altijd
  - "gefixeerde kenmerken": geslacht, leeftijd, *etc.*
  - kleine aantallen > toeval veronderstelt grote aantallen
  - ethische overwegingen
- > alternatief: matching
  - = voor elk element in de experimentele conditie zorgen we voor een vergelijkbaar element in de vergelijkingsconditie
  - op waarden van een beperkt aantal kenmerken (waarvan we verwachten dat ze relevant zijn)
  - = meteen zwakte

16

VOORMETING

- soms kan een voormeting niet
  - effect treedt pas op bij experimentele stimulus
  - *vb.* surveyonderzoek > non-respons
  - 'prepaid, unconditional incentive'

Tabel 18. Het effect van drankbonnetje op deelname aan korte vragenlijst ter plekke.

	Totaal aantal contactnames	*Deelname*		
	n_i	X_i	$\hat{p}_i$	$SE(\hat{p}_i)$
Drankbonnetje	1842	1478	0,802**	0,022
Geen drankbonnetje	1812	1272	0,702	0,036

Het verschil is significant ($z = 2,40$, $p = 0,008$).
Bron: Publieksonderzoek PBO99 en Vlaams-Brabant.

17

VOORMETING (2)

- consistent invullen in nameting of 'praktijkervaring'
  - > bedreigen interne geldigheid
- door voormeting gevoeliger voor stimulus
- (technisch: interactie-effect van voormeting op relatie stimulus en afhankelijke variabele)
  - eenheden worden zich bewust van studie
  - vormen zich een idee van wat onderzoekers bestuderen
  - conformeren aan verwachtingspatroon van onderzoekers
  - *> reactiviteit = effect van meten op wat gemeten wordt*
  - *cf.* Hawthorne effect – bedreigt externe geldigheid

18

ILLUSTRATIE SENSOA

- illustratie: studie over veilig vrijen

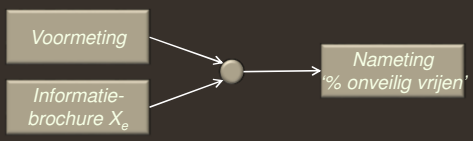
SENSOA
PRAAT OVER SEKS

- effectiviteit van informatiecampagne?
  - voormeting: vragen over seksueel gedrag
    - hoe vaak? veilig of niet? hoe?
    - wat weet je ervan? *etc.*
  - experimentele stimulus: brochures en affiches

ILLUSTRATIE SENSOA

21

- door voormeting…
  - … ga je meer aandacht schenken aan informatie over veilig vrijen: meer bewust van de gevaren > gedrag beïnvloeden
- zonder voormeting
  - minder aandacht > gedrag wordt niet beïnvloed



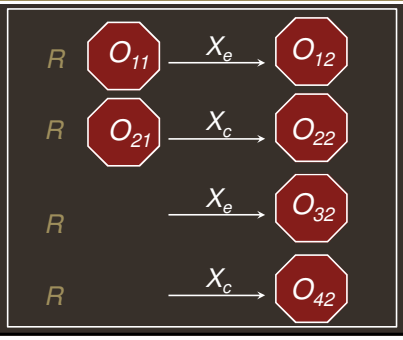
OPLOSSING?

22

- mogelijk interactie-effect van voormeting op relatie experimentele stimulus en outcome
- hoe ga je dat na?
  - 2 groepen mét voormeting
    - experimentele en vergelijkingsgroep
    - voormeting kan invloed hebben
  - 2 groepen zónder voormeting, enkel nameting
    - experimentele en vergelijkingsgroep
    - voormeting kan geen effect hebben—we doen er geen…
- *= Solomon four group design*

SOLOMON FOUR GROUP DESIGN

23

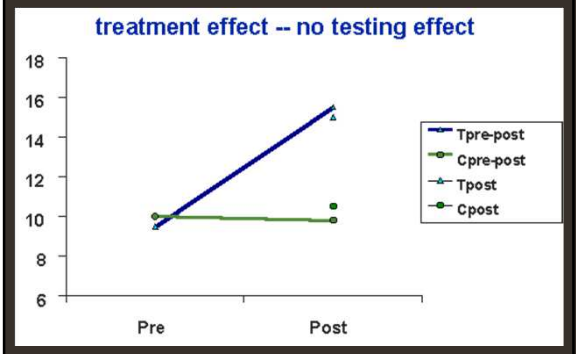


SOLOMON FOUR GROUP DESIGN (2)

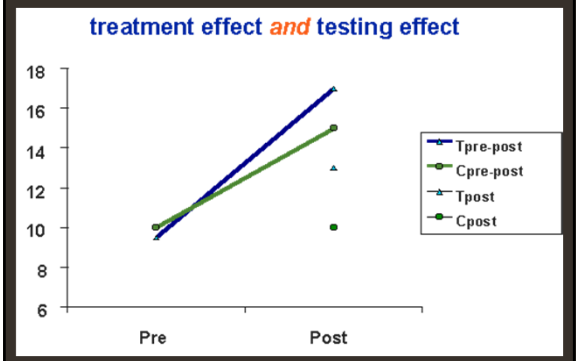
24

- als $O_{12} = O_{32}$ en $O_{42} = O_{22}$ (of zeer gelijkend): geen effect voormeting
- anders: wel effect van voormeting
- schematisch…

25



26

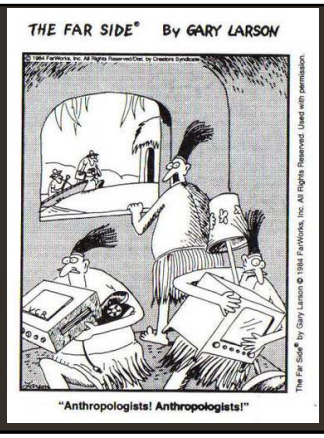


27

GELDIGHEID?

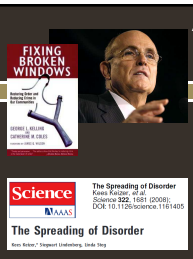
- externe geldigheid is zwakte
- = veralgemeenbaarheid naar personen/situaties buiten experimentele context
  - participanten zijn vaak speciale groep—*vb.* studenten—veralgemeenbaarheid naar algemene populatie?
    = populatiegeldigheid
  - artificiële setting—*vb.* labo—verhindert veralgemeenbaarheid naar 'normale' setting
    = naturalistische geldigheid
  - reactiviteit?
    - Hawthorne effect & placebo-effect

28



29

VELDEXPERIMENTEN

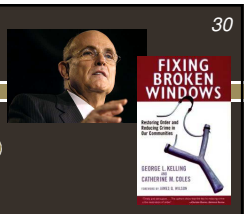


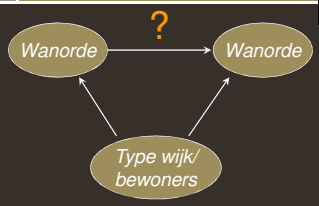
- 'broken windows' theorie
  - causaliteit?
  - > opzetten veldexperimenten

"Moreover, to our knowledge, research on the BWT has so far been correlational, so conclusions about causality are shaky. The BWT suggests that a setting with disorder triggers disorderly and petty criminal behavior, but it might be the other way around or both may be caused by a third variable." (Keizer *et al.*, 2008: 1681)

30

VELDEXPERIMENT





- reeks van 6 veldexperimenten: *injunctive norms*
  - > gaan mensen sneller normen overtreden als ze zien dat een norm geschonden is?

31

STUDIE 1: FLYERS

- > flyers aan stuur
- definiëren situaties
  - experimentele conditie: mét graffiti
  - vergelijkingsconditie: zonder graffiti
- randomiseren: op zelfde moment in de week, zelfde weer > 77 individuen in elke conditie
- > 69% versus 33% ($p < 0{,}001$)

32

STUDIE 2: SHORTCUT

- nemen kortere weg door werf
  - experimentele conditie: 4 fietsen gestald
  - vergelijkingsconditie: geen fietsen gestald
- randomiseren: op zelfde moment > 44 en 49 personen
- > 82% versus 27% ($p < 0{,}001$)

33

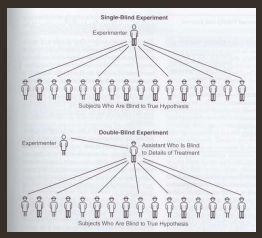
GELDIGHEID? (2)

- interne geldigheid is sterkte
  - grote zekerheid dat de factor waaraan we het effect toeschrijven, inderdaad de experimentele factor is
- toch bedreigingen
  - buitenexperimentele gebeurtenissen (*vnl.* 'local history')
  - maturatie/spontane veranderingen
  - testeffect
  - instrumentatie
  - statistische regressie
  - selectie
  - uitval

34

GELDIGHEID (3)

- toch bedreigingen (vervolg)
  - verwachtingen onderzoeker > 'double-blind'



35

QUASI-EXPERIMENTELE DESIGNS

- verschil met klassieke experiment?
  - geen randomisatie
  - of geen voormeting
  - of geen vergelijkingsgroep
- DUS:
  - minder controle over onderdelen design
  - minder controle voor storende factoren
  - interne geldigheid beperkter
  - causale uitspraken problematischer

36

TIJDREEKSONTWERP

O_{11} O_{12} O_{13} X_e O_{14} O_{15} O_{16}

- meerdere metingen vóór en ná experimentele stimulus
- twee types
  - zelfde groep: panelstudie
  - andere groepen: herhaald cross-sectioneel onderzoek
- beleidsevaluatieonderzoek a.d.h.v. indicatorenreeksen
  - monitoring: *vb.* criminaliteit in New York

37

Note. The vertical line denotes when the Marijuana Initiative Campaign began. The solid line represents HSS regression. The dotted line represents LSS regression.

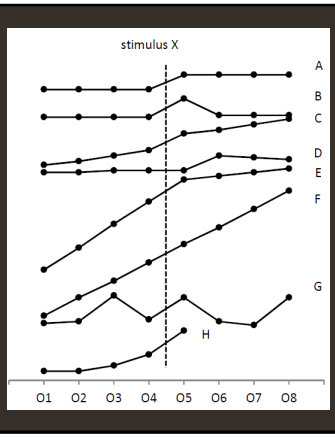
FIGURE 1—Thirty-day marijuana use regression plots for high sensation-seeking (HSS) and low sensation-seeking (LSS) adolescents in the combined Knox County, Tenn, and Fayette County, Ky, sample: April 1999–March 2003.

38

TIJDREEKSONTWERP (2)

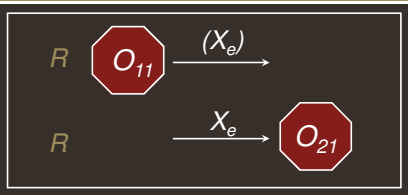
- geen controlegroep
- wel reeks voormetingen
  - grotere betrouwbaarheid
  - zicht op natuurlijke fluctuaties en evoluties
- laat toe uitgestelde effecten vast te stellen

39



40

AFZONDERLIJKE STEEKPROEF ONTWERP



- het is niet mogelijk een groep niet te onderwerpen aan de experimentele conditie
  - *vb.* grote mediacampagnes

41

AFZONDERLIJKE STEEKPROEF ONTWERP (2)

- twee voldoende grote toevalsteekproeven trekken uit populatie
  - eerste steekproef: voormeting, O_{11}
  - tweede steekproef: nameting, O_{21}
- effect = nameting O_{21} – voormeting O_{11}
  - beide groepen vergelijkbaar, want op toevalsbasis opgesteld
  - random assignment
- PROBLEEM: buitenexperimentele gebeurtenissen kunnen O_{21} beïnvloeden: 'history'

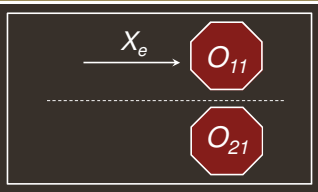
42

PRE-EXPERIMENTELE DESIGNS

- zijn eigenlijk geen experimentele designs
- laten geen causale verklaringen toe/interne geldigheid is zwak
  - of geen controlegroep
  - of geen toevallige samenstelling van de groepen
- enkel te gebruiken als het echt niet anders kan

43

BESTAANDE GROEPEN MET ALLEEN NAMETING

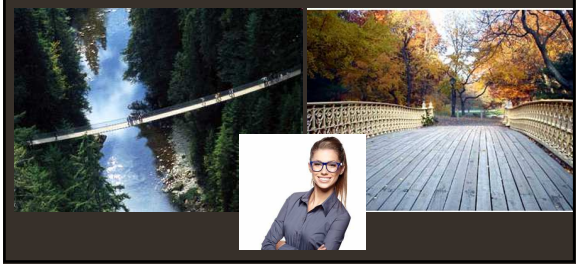


- twee groepen, geen voormeting, geen R
- PROBLEEM: selectie

44

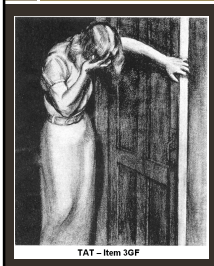
CAPILANO CANYON SUSPENSION BRIDGE

- seksuele aantrekking is groter bij grote emotie?



45

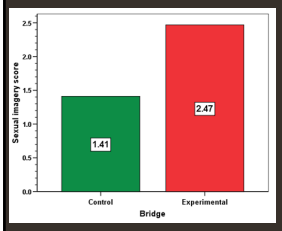
CAPILANO CANYON SUSPENSION BRIDGE

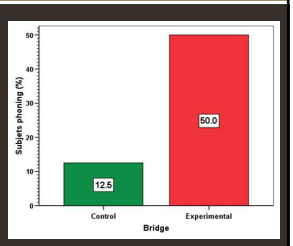

TAT – Item 3GF

- voorbijgangers gevraagd om verhaal te schrijven over afbeelding > beoordelen op seksuele inhoud
- + interviewster geeft telefoonnummer

46

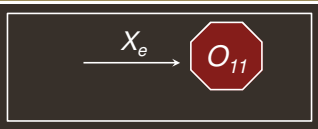
CAPILANO CANYON SUSPENSION BRIDGE





47

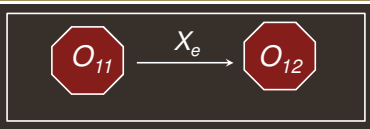
GROEP MET ALLEEN NAMETING



- slechts één groep, geen voormeting
- onafhankelijke variabele varieert niet
  - iedereen krijgt stimulus toegediend
  - effect zonder experimentele stimulus niet gekend
- geen uitspraak mogelijk over 'causaal effect'

48

GROEP MET VOOR- EN NAMETING



- nu wel nulmeting
  - voormeting: 'vergelijkingsgroep'
  - nameting: 'experimentele groep'
- PROBLEEM: buitenexperimentele gebeurtenissen, reactiviteit voormeting

49

ILLUSTRATIE

- antirookcampagne op school
- voordien: 25% rokers
- na één jaar campagne: 15% rokers
  - effect: 10 percentpunten
  - volledig te wijten aan campagne?
    - wat met leeftijdseffect? > idee van 'ze groeien eruit'
    - prijsstijging sigaretten?
    - andere campagnes in media?
    - door voormeting meer bewust van risico's? (= reactiviteit)

50

	Buitenexperimentele gebeurtenissen	Maturatie	Testeffect	Instrumentatie	Statistische regressie	Selectie	Uitval	Interactie-selectie en maturatie, testeffect, etc.
Alleen nameting	-	-	...	...	...	-	-	...
Voor- en nameting	-	-	-	-	?	+	+	-
Bestaande groepen met alleen nameting	+	?	+	+	+	-	-	-
Tijdreeksontwerp	-	+	+	?	+	+	+	+
Niet-equivalent ontwerp met vergelijkingsgroep	+	+	+	+	?	+	+	-
Afzonderlijke-steekproef ontwerp met voor- en nameting	-	-	+	?	+	+	-	-
Voormeting-nameting ontwerp met vergelijkingsgroep	+	+	+	+	+	+	+	+
Ontwerp met vergelijkingsgroep en alleen een nameting	+	+	+	+	+	+	+	+
Solomon vier-groepen ontwerp	+	+	+	+	+	+	+	+

Methodologie van de Sociale Wetenschappen

Een inleiding

Henk Roose & Bart Meuleman

Academiejaar 2013-2014

Cursustekst bij colleges 'Methodologie van de Sociale Wetenschappen'

Eerste jaar Bachelor Communicatiewetenschappen,

Politieke Wetenschappen,

Sociologie

Overzicht hoofdstukken

Voorwoord

Er bestaan heel wat goede methodologieboeken in Vlaanderen—het tweedelige handboek van Ron Lesthaeghe uit de jaren ’80, dat van Jaak Billiet uit de jaren ’90 en later dat van Jaak Billiet en Hans Waege (2006) zijn er alvast enkele. Waarom dan een nieuw inleidend handboek methodologie publiceren? Welnu, dit boek is gegroeid uit onze ervaringen bij het geven van hoorcolleges ‘Methodologie van de sociale wetenschappen’ aan Universiteit Gent en Katholieke Universiteit Leuven. Telkens je de vrij abstracte principes van onderzoeksmethodologie aan een grote groep studenten moet uitleggen, heb je nood aan sprekende illustraties, aan relevante voorbeelden. Meestal grijp je dan terug naar je eigen interessegebied of onderzoekdomein voor dit illustratiemateriaal en de stap naar een eigen handboek waarin precies die voorbeelden uitgebreid aan bod komen, is dan ook snel gezet. Bovendien willen we met dit boek aanknopen bij een reeks recente methodologische evoluties, zoals onder meer het gebruik van web-surveys, het toepassen van QCA, het benutten van Big Data of het aanwenden van mixed-methods.

We hebben dit handboek methodologie opgebouwd uit vier delen. In een eerste deel behandelen we de bouwstenen van sociaalwetenschappelijk onderzoek, met name theorie en waarneming. Hoe werken beide op elkaar in en hoe kunnen ze elkaar aanvullen? We besteden de nodige aandacht aan enkele epistemologische veronderstellingen alsook aan ethische codes en deontologische regels die (sociaal)wetenschappelijk onderzoek kenmerken. Deel twee is gewijd aan hetgeen een onderzoeker moet weten wanneer hij/zij een onderzoek wil opzetten: van het onderzoeksplan over de aard en wijze van ‘meten’ tot de principes van steekproeftrekking. Deel drie en vier gaan respectievelijk in op de brede waaiers aan kwantitatieve en kwalitatieve methoden/technieken die als onderzoeker tot je beschikking staan. Zo behandelen we experimenten, surveyonderzoek, veldonderzoek en participerende observatie, diepte-interviews en historisch-vergelijkend onderzoek broederlijk naast elkaar—en het is heus niet altijd zo geweest.

De methoden-strijd die heeft gewoed vanaf de jaren ’70 tussen de kwantitatieve en kwalitatieve benaderingen in de sociale wetenschappen, laten we achter ons—gegooi met

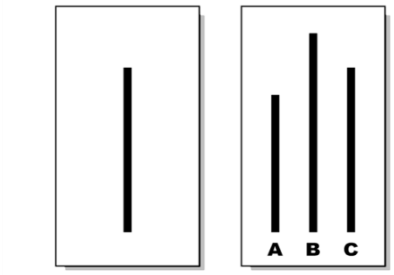
modder zoals toen, zal je hier niet vinden. Dit betekent echter niet dat we de inzichten en stellingnames van toen onbelangrijk achten. We zijn er namelijk van overtuigd dat inzicht en kennis in de mogelijkheden van een zo breed mogelijk spectrum aan onderzoekstechnieken en hun achterliggende epistemologische assumpties de kwaliteit van de onderzoekspraktijk ten goede komt—of die nu kwalitatief of kwantitatief gericht is.

Graag willen we een aantal mensen bedanken die hebben bijgedragen tot wat nu voorligt. Allereerst bedankt aan de studenten die onze colleges volgen en die ons met kritische bemerkingen en suggesties aanzetten tot het geven van een kristalheldere en duidelijke uitleg. Ook zijn we iedereen erkentelijk die eerdere versies van de tekst kritisch hebben doorgenomen: Stijn Daenekindt, Susan Lagaert, Bruno Vandenbussche en Rachel Waerniers. Ten slotte danken we Peter Laroy van Academia Press voor zijn input en ideeën om dit werk snel een mooie vorm mee te geven.

Henk Roose & Bart Meuleman
Gent, 29 februari 2014

Hoofdstuk 8. Experimentele ontwerpen.

Beeld jezelf de volgende situatie in: je wordt door een universiteit uitgenodigd om samen met zeven andere proefpersonen de lengte van lijnen te beoordelen. Eerst krijgt iedereen een kaart te zien met één lijn erop, gevolgd door een andere kaart met drie lijnen erop: lijnen a, b en c. Elke proefpersoon moet aangeven welke lijn even lang is als de lijn op het eerste kaartje door te antwoorden met 'a', 'b' of 'c'. Jij moet telkens als laatste of voorlaatste een antwoord geven. De proef houdt een totaal van 18 van die vergelijkingen in. Bij de eerste twee vergelijkingen van de lengte geven je collega-proefpersonen het correcte antwoord. Vanaf de derde vergelijking echter beginnen ze overduidelijk verkeerde antwoorden te geven. In totaal doen ze dit 12 van de 18 keer. Hoe reageer je hierop? Hoe ga je om met de groepsdruk die ontstaat om ook het foute antwoord te geven? Of ga je op je eigen oordeel af tegen de groepsdruk in?

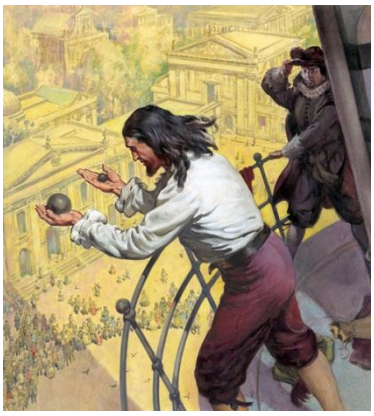


Deze situatie neemt je mee naar de experimenten van Solomon Asch (1955). Hij herhaalt dit experiment bij 123 proefpersonen. Bij 36,8% van de vergelijkingen begeven de proefpersonen onder de groepsdruk—de andere 'proefpersonen' zijn uiteraard handlangers van Asch—en geven een verkeerd antwoord in navolging van de opinie van meerderheid. Ter controle: in 'normale' omstandigheden zouden mensen in minder dan 1% van de vergelijkingen een fout antwoord geven. Ongeveer een kwart van de proefpersonen bezwijkt geen enkele keer onder de groepsdruk, 75% minstens één keer, 5% altijd. Asch toont hiermee aan hoe mensen geneigd zijn te conformeren onder groepsdruk. Zijn beroemde experimenten geven aan hoe sociale druk mensen ertoe aanzet om dingen te zeggen die nochtans pertinent onjuist zijn. Asch vormt met zijn ideeën de voedingsbodem voor de misschien nog beroemdere experimenten van zijn leerling Stanley Milgram (1974), waarin hij gehoorzaamheid aan autoriteitsfiguren beschrijft, zelfs al zijn de bevelen in strijd met iemands persoonlijke overtuiging—de herinnering aan situaties in de Duitse vernietigingskampen uit WO II is nooit veraf...

1. Inleiding

Experimentele ontwerpen kennen hun oorsprong in de natuurwetenschappen, zoals fysica, biologie of chemie, waar ze gelden als absolute standaard voor rigoureus onderzoek. Ook in verwante disciplines als geneeskunde en ingenieurswetenschappen zijn experimenten ingeburgerd als de uitgelezen manier om tot onwrikbare wetenschappelijke kennis te komen. Basisidee van een experiment is dat je als onderzoeker twee situaties creëert. In de eerste situatie manipuleer je één element en je gaat na of die ingreep een effect heeft op

een kenmerk van interesse. In de andere situatie manipuleer je niets—dit is de controlesituatie. Vervolgens kijk je of en in hoeverre het kenmerk van interesse in de eerste situatie verschilt van datzelfde kenmerk in de tweede situatie.



Zo wil de legende dat Galileo Galilei (1564-1642) een claim van Aristoteles experimenteel getest heeft op de toren van Pisa. Aristoteles postuleert in zijn theorie over zwaartekracht dat objecten vallen met een snelheid recht evenredig aan hun massa. Volgens Aristoteles vallen zwaardere objecten dus sneller dan lichte. Galilei wil dit idee testen, gaat op de toren van Pisa staan en laat twee bollen met een verschillende massa naar beneden vallen. En wat blijkt? Beide bollen vallen even snel. Later zijn die principes vastgelegd in de grondslagen van de dynamica door Isaac Newton (1643-1727). Dit experiment wordt trouwens nog een keertje overgedaan tijdens één van de maanlandingen in 1971, wanneer astronaut David Scott op de maan een hamer en een veer laat vallen en de valversnelling van beide voorwerpen dezelfde blijkt—opnieuw in tegenspraak met de ideeën van Aristoteles.

De eerste experimenten uit de fysica onderschrijven het idee dat alleen actie en experiment inzicht kunnen verschaffen in de werkelijkheid. Mensen als Galileo Galilei en Francis Bacon (1561-1626) zijn allemaal doordrongen van het principe dat je enkel door het toepassen van gedegen empirisch, natuurwetenschappelijk onderzoek via experimenten kennis over de Goddelijke creatie kan opbouwen. Inzichten op gezag van overgeleverde, geschreven bronnen worden niet zomaar meer aangenomen. Getuige daarvan is bijvoorbeeld het adagium van *The Royal Society of London for Improving Natural Knowledge* gesticht in 1660. Hun motto luidt 'Nullius in verba', wat vrij vertaald zoveel betekent als: neem niets op het gezag van een neergeschreven autoriteit aan. En hoewel de procedures om experimenteel onderzoek te verrichten door de eeuwen heen zijn aangepast, verbeterd en verfijnd, blijft het onderliggende basisontwerp ook nu nog als model dienen binnen de huidige natuurwetenschappen. Het experiment geldt er als ideaal, als dé manier om tot ware kennis over en inzicht in natuurverschijnselen te komen—perfect in lijn met het naturalisme (zie Hoofdstuk 3).

Ook in de sociale wetenschappen worden experimentele ontwerpen gebruikt om tot inzicht in oorzakelijke verbanden in de sociale werkelijkheid te komen. Sommige domeinen

lenen zich makkelijker tot het overnemen van methoden uit de natuurwetenschappen. Vooral de psychologie kent een lange traditie in het gebruik van experimenten—zij waren de eerste om de geliefkoosde methode van de natuurwetenschappen ook toe te passen binnen de sociale wetenschapen.[1] Vanaf het begin van de 20ste eeuw omarmt de psychologie het experiment om regelmatigheden in menselijk handelen te identificeren, te isoleren en causaal te duiden. Gehoopt wordt om algemene inzichten op te bouwen met betrekking tot universele patronen en wetmatigheden van het menselijk handelen. Vooral sinds de jaren '60 is het gebruik van experimentele ontwerpen in de rest van de sociale wetenschappen toegenomen. Onderzoekers passen dan experimentele ontwerpen toe in studies naar onder meer productiviteit bij arbeiders in de fabriek, naar onderwijsmethoden en hun effectiviteit en naar effecten van boodschappen uit massamedia op hun publiek.

> In *News that matters. Television and American opinion* schrijven Shanto Iyengar en Donald Kinder (1987) over de invloed van de media op politieke percepties in de Verenigde Staten. Ze besteden hiervoor zowel aandacht aan 'agenda setting' als aan 'priming' van nieuwsberichten. *Agenda setting* verwijst naar het inzicht dat sociale problemen die een prominente plaats krijgen in een nieuwsbulletin, precies ook die problemen zullen zijn die kijkers als het meest prangend, het meest belangrijk zullen beschouwen. *Priming* op zijn beurt duidt op veranderingen in de standaarden waarmee het publiek iets beoordeelt als een gevolg van de wijze waarop nieuwsberichten verhalend worden ingebed. In een nieuwsitem krijgen sommige aspecten van een issue immers meer aandacht of nadruk en dit beïnvloedt de wijze waarop het publiek iets/iemand beoordeelt of evalueert. Iyengar en Kinder passen op ingenieuze wijze een experimenteel ontwerp toe. Proefpersonen worden uitgenodigd naar de campus van Yale University om daar gedurende één week dagelijks naar het nieuws te kijken, zogezegd omdat thuis de kans om afgeleid te worden te groot is en om te verzekeren dat iedereen dezelfde nieuwsberichten in identieke omstandigheden zou kunnen bekijken. In werkelijkheid worden de proefpersonen op toevallige wijze over een aantal condities verdeeld waarin ze door de onderzoekers aangepaste nieuwsuitzendingen te zien krijgen. De nieuwsuitzendingen worden precies zo gemonteerd zodat ze variëren in de mate waarin bepaalde problemen meer aandacht kregen—zo krijgen de ernst van milieuvervuiling, de ongeschiktheid van de Verenigde Staten om zichzelf afdoende militair te verdedigen, economische problemen en werkloosheid verschillende aandacht. Vóór en ná die week wordt hen een enquête voorgelegd over wat ze als de belangrijkste problemen beschouwen in de Amerikaanse samenleving en op basis van welke criteria ze hun president beoordelen. Via een reeks experimenten tonen de

> auteurs aan dat agenda setting en priming een causale invloed uitoefenen op attitudes, politieke overtuigingen en zelfs op stemgedrag. Hiermee ontkrachten ze de idee dat de media slechts reeds bestaande attitudes en predisposities weten te versterken, maar dat nieuwsberichten heel krachtig zijn in het beïnvloeden van opinies en handelen. Nieuwsberichten—zo er niet wijs mee omgesprongen wordt—kunnen dus overtuigende wapens zijn in de politieke arena.

De publicatie van het boek van Campbell en Stanley (1963) luidt in die zin een periode in waar het experimentele ontwerp als wetenschappelijke standaardpraktijk naar voren wordt geschoven—ook buiten de psychologie.[2] Bovendien worden enkele designs geïntroduceerd die minder stringente variaties voorstellen op het klassieke, ideaaltypische experiment. Het devies bij veel Amerikaanse academici in de jaren '60 luidt immers dat sociale problemen op een wetenschappelijke manier aangepakt moeten worden en dat de effectiviteit van sociale interventies door middel van rigoureuze experimentele, maar tegelijk praktisch haalbare studies getest moet worden. Interventies die niet effectief zijn, moeten worden afgevoerd. Sociale hervormingen die werken, moeten worden verdergezet. Dit is het ideaal van Donald Campbell's zogenaamde 'experimenting society', waarin de samenleving zich aandient als laboratorium en sociale interventies aan de hand van experimenteel onderzoek worden getest op hun merites (Campbell & Cook, 1976; Cook & Campbell, 1979). Wetenschap en democratie zouden zo hand in hand naar een beloftevolle toekomst marcheren met de sociale wetenschapper als onpartijdige arbiter… Een bekend voorbeeld van experimentele studies in die optimistische geest is het evaluatieonderzoek van Sesamstraat.

> Sesamstraat is onderdeel van interventies in het kader van de *War on Poverty* in de Verenigde Staten in de jaren '60. Het eerste seizoen start in 1969 en het is de bedoeling om de cognitieve en sociale vaardigheden van kinderen aan te scherpen en hen voor te bereiden op de lagere school—vooral van kinderen met een niet-blanke, kansarme achtergrond. Het devies is dat 'als je de aandacht van kinderen kan vasthouden, dan kan je hen opleiden ook'. De serie kent een overdonderend succes en Sesamstraat krijgt overal in de wereld een plekje op de televisie. Na het eerste seizoen wordt de serie aan een experimenteel onderzoek onderworpen om te beoordelen of kijken naar Sesamstraat werkelijk de gewenste effecten had (zie Ball & Bogatz, 1970; Fisch, 2005). Dit rapport toont de positieve invloed van dit televisieprogramma aan. In een reeks van experimenten worden kinderen toegewezen aan condities die verschillen in de mate van blootstelling aan Sesamstraat. Analyses wijzen uit dat hoe

> frequenter kinderen kijken, hoe groter de positieve effecten: ze scoren hoger op vaardigheden zoals het herkennen van cijfers, letters en woorden, het sorteren en classificeren van vormen, etc.

Experimenten zijn uitermate geschikt om het oorzakelijke verband tussen twee of meer kenmerken objectief en onvertekend bloot te leggen—in lijn met de visie op causaliteit zoals in Hoofdstuk 2 uiteengezet: er is een verband tussen twee variabelen, de ene variabele gaat aan de andere vooraf en het vastgestelde verband is geen gevolg van of wordt niet meebepaald door andere variabele. Eén van die kenmerken, de experimentele stimulus met name, moet manipuleerbaar zijn. Dat wil zeggen dat je als onderzoeker die stimulus moet kunnen 'toedienen' aan proefpersonen. Voorbeelden van zo'n manipuleerbare kenmerken waarvan de effectiviteit kan worden onderzocht zijn televisieprogramma's, informatie- of reclamecampagnes via de media, onderwijsmethodes, geneesmiddelen, managementstijlen, etc. Wil je als sociale wetenschapper de wereld om je heen begrijpen in termen van oorzakelijke relaties, dan moet je die wereld aanpassen en erop ingrijpen zodat ze experimenteel 'onderzoekbaar' wordt. Om de volle kracht van een experimenteel ontwerp te benutten is het dus noodzakelijk om een paar relevant geachte kenmerken te isoleren en te manipuleren. Om te weten wat relevant is of niet, gebruik je theoretische inzichten binnen het onderzoekdomein in kwestie.

Dat isoleren en manipuleren is echter niet altijd makkelijk in complexe maatschappelijke settings. De complexiteit van processen die het menselijke handelen beïnvloeden laat nauwelijks toe om door middel van veldexperimenten *in situ* alle relevante kenmerken te isoleren, laat staan te controleren zoals in een laboratorium, om zo tot causale verklaringen te komen. Veldexperimenten vinden immers plaats in een 'natuurlijke', moeilijk beheersbare realistische setting tegenover laboratoriumexperimenten, die plaatsvinden in de strikt controleerbare, maar hoogst artificiële omgeving van het laboratorium. Zo is het bijvoorbeeld moeilijk om inzicht te krijgen door middel van experimenten over de samenhang tussen pakweg gezinssamenstelling—eventueel een echtscheiding—en stressbeleving bij jonge kinderen, of tussen iemands opleiding en zijn/haar cultuurbeleving. Noch gezinssamenstelling, noch opleiding laten zich immers makkelijk manipuleren om op toevallige wijze toegewezen te worden aan een experimentele en een vergelijkingsgroep. Hier botsen we op de technische—en in sommige gevallen uiteraard ook ethische (zie

paragraaf 6 hieronder voor verdere bespreking van ethische kwesties)—beperkingen van het experimentele ontwerp. Het is bijvoorbeeld uitgesloten om adolescenten te dwingen om sigaretten te roken, opdat je zou kunnen nagaan of roken op jonge leeftijd de kans op longkanker op latere leeftijd verhoogt. Net zo min is het denkbaar dat je patiënten een werkzaam geachte remedie tegen hart- en vaatziekten ontzegt, alleen maar om netjes een vergelijkingsgroep samen te stellen die de experimentele stimulus niet toegediend krijgt.

2. Onderdelen van het klassieke experimentele ontwerp

Het basisschema van een experiment vormt de formele neerslag van wat mensen doen als ze situaties vergelijken: ze veranderen één ding in een situatie en ze gaan na of het resultaat mét die verandering verschilt van wat er zou gebeuren zonder die verandering. Alle overige kenmerken in de verschillende situaties zijn gelijk of vergelijkbaar. Zoals de illustratie in de proef van Galilei aangeeft, waar de twee bollen naar beneden worden gegooid. De twee bollen verschillen alleen in massa. Voor de rest zijn ze vergelijkbaar—zoals hun volume, de hoogte van waar ze vallen, de omgevingstemperatuur en -luchtdruk, etc. In een experimenteel ontwerp staat dus vergelijken centraal, maar dan op een systematische en zorgvuldig doorgedachte manier.

Hoe zit dit schema van een klassiek experiment in elkaar? Vertrekpunt is dat je het causale effect van één kenmerk op een andere variabele wil isoleren. Er is met andere woorden een hypothese over het oorzakelijke effect van een kenmerk op een afhankelijke variabele. Om die oorzakelijke relatie te testen, creëer je twee situaties: één waarin dat kenmerk optreedt en één waarin dat kenmerk niet optreedt. Dan vergelijk je de uitkomst op de afhankelijke variabele tussen de situatie mét en zonder het kenmerk. Voor al het overige moeten beide situaties vergelijkbaar zijn. Dat variërende kenmerk heet je de stimulus of ook wel experimentele manipulatie en kan van velerlei aard zijn. De situatie waar de stimulus plaatsvindt, is de experimentele groep; de situatie zonder stimulus is de vergelijkingsgroep. Bijvoorbeeld in Sesamstraat is de stimulus de frequentie van kijken naar het Sesamstraat op televisie. Bij medisch onderzoek kan het gaan om een geneesmiddel waarvan je het positieve effect wil nagaan op een ziektebeeld of aandoening. De afhankelijke variabele is dan respectievelijk hoe goed kinderen scoren op een test in het herkennen van letters, woorden, cijfers, etc. en genezingsproces bij een bepaalde medische aandoening.

2.1. Basisschema klassiek experiment

In Figuur 8.1 worden de diverse onderdelen van een klassiek experimenteel ontwerp gevisualiseerd (dit is het 'pretest-posttest control group design' in het Engels). Het gaat om de experimentele versus de vergelijkingsgroep met het principe van randomiseren, de stimulus of onafhankelijke variabele en de voor- en nameting van de afhankelijke variabele. Tegelijk krijg je de uniforme notatie en grafische presentatie te zien die voor de verschillende onderdelen worden gebruikt.

Figuur 8.1. Schematische voorstelling onderdelen klassiek experimenteel ontwerp (voormeting-nameting ontwerp met vergelijkingsgroep).

R	O_1	X	O_2
R	O_3		O_4

Elke rij staat voor een situatie—ook wel conditie genoemd. De X verwijst naar de stimulus of experimentele manipulatie waaraan de experimentele groep wordt blootgesteld en waarvan je het oorzakelijke effect wil onderzoeken. Bij een zuiver experiment heb je als onderzoeker zelf in de hand wie aan de stimulus wordt blootgesteld en wie niet. O_i staat voor observatie, een score op een variabele. De O's en X in een rij hebben betrekking op groepen personen in éénzelfde situatie: O_1, X en O_2 slaan op de experimentele groep, O_3 en O_4 op de vergelijkingsgroep. De links-rechts dimensie verwijst naar de voortschrijdende tijd, de gebeurtenissen in de kolommen gebeuren op hetzelfde moment, respectievelijk op t_0 en t_1. Het symbool R staat voor randomisering, ofwel dat proces dat instaat om gelijkwaardige groepen—dat wil zeggen statistisch equivalente groepen—te creëren vóór het experiment van start gaat.

2.2. Randomiseren en matchen

Zoals hierboven gezegd, functioneren experimentele ontwerpen op een systematische en weldoordachte vergelijking. Uiteraard wil je als sociale wetenschapper niet vervallen in het spreekwoordelijke vergelijken van appelen met peren. Met andere woorden, je wil groepen vergelijken die met elkaar vergelijkbaar zijn, behalve op het hebben ontvangen van een experimentele stimulus. Randomiseren is nu precies die procedure die vergelijkbare of

gelijkwaardige groepen kan creëren. Proefpersonen worden immers op louter toevallige basis aan de experimentele of vergelijkingsconditie toegewezen: iedereen heeft een gekende, gelijke kans om ofwel in de experimentele dan wel in de vergelijkingsgroep te belanden. Op die manier zijn beide groepen op toevalsfouten na aan elkaar gelijkwaardig of statistisch equivalent—dit voor alle kenmerken en niet alléén voor de afhankelijke variabele. Zo is $O_1 = O_3$ op toevalsfouten na. Door randomisering schakel je dus zo veel mogelijk systematische verschillen tussen de groepen uit. Wanneer een vergelijkingsgroep op basis van toevallige toewijzing is samengesteld, wordt ook wel eens de term controlegroep gebruikt in plaats van het meer algemene begrip vergelijkingsgroep. In de praktijk is randomisatie eenvoudig. Zodra je je groep van proefpersonen voor het experiment hebt samengesteld, kan je door middel van het opgooien van een muntstuk of door geblinddoekt naamkaartjes uit een hoed te selecteren op toevalsbasis de experimentele en controlegroep samenstellen.

Hoewel toevallige toewijzing of randomisering een essentieel onderdeel is van een succesvol klassiek experimenteel ontwerp, is het niet steeds mogelijk. Bepaalde kenmerken zoals geslacht of leeftijd—kenmerken die als het ware vastkleven aan een persoon—kan je eenvoudigweg niet toevallig toewijzen. Het is bijvoorbeeld immers onmogelijk om proefpersonen toevallig te verdelen over de twee geslachtscategorieën om het effect van geslacht op een afhankelijke variabele na te gaan. Ook wanneer je slechts met een beperkt aantal proefpersonen werkt, functioneert randomiseren niet optimaal. De werking van het toeval is namelijk gediend met grote aantallen. Wanneer je slechts een paar tientallen proefpersonen moet verdelen over experimentele condities, kan je er niet vanuit gaan dat de groepen niet systematisch van elkaar zullen verschillen. Een alternatieve werkwijze dringt zich op om alsnog vergelijkbaarheid tussen groepen te maximaliseren.

> Een aantrekkelijke vrouw wordt als nog aantrekkelijker beschouwd wanneer mannen een sterke emotie ervaren, zoals angst, dan wanneer ze niet beïnvloed worden door emotie. Twee Canadese onderzoekers willen deze uitspraak empirisch testen (Dutton & Aron, 1974). Hiervoor laten ze een bloedmooie interviewster mannelijke voetgangers tegenhouden die ofwel een gevaarlijke hangbrug willen oversteken—de Capilano Canyon Suspension Bridge—ofwel een gewone brug willen gebruiken. Zij legt hun uit dat ze een project moet uitvoeren voor school over de relatie tussen landschap en creatieve expressie en of ze hiervoor een enquête willen invullen. Als ze klaar zijn

met het invullen van de enquête, scheurt de interviewster een hoekje van het formulier en noteert haar telefoonnummer 'voor het geval de respondenten geïnteresseerd zouden zijn om meer te weten te komen over haar onderzoek'. Wat blijkt? 9 van de 18 geïnterviewden aan de hangbrug tegenover 2 van de 16 bij de gewone brug bellen haar terug ($\lambda^2 = 5{,}7$, $p < 0{,}02$). Levert dit bevestiging voor een verband tussen hevige emotie en seksuele aantrekking? Welnu, op het eerste gezicht wel, maar veldexperimenten hebben vaak te kampen met differentiële groepssamenstelling en dit is ook hier mogelijk het geval. Het zou wel eens kunnen dat de mannen die de avontuurlijke brug willen oversteken sowieso avontuurlijker aangelegd zijn—steeds op zoek naar kicks—en dat het precies dit kenmerk is—veeleer dan de angst voor brug—die hen doet terugbellen naar de knappe interviewster.

Zo'n alternatieve toewijzing—zij het niet zo vaak gebruikt—is het zogenaamde matchen. Hierbij probeer je vergelijkbare groepen samen te stellen op basis van zo veel mogelijk relevant geachte kenmerken of op de combinatie ervan. Dus, wanneer je een proefpersoon met een bepaald kenmerk in de experimentele groep indeelt, wijs je een andere persoon met hetzelfde kenmerk toe aan de vergelijkingsgroep. Er vigeren drie soorten matching, namelijk precisie- of paarsgewijs matchen, frequentie- of groepsmatchen en zwak matchen. Bij precisie- of paarsgewijs matchen ga je voor elke combinatie van kenmerken bij een lid van de experimentele groep zorgen voor eenzelfde combinatie van kenmerken in de vergelijkingsgroep. Je gaat als het ware op zoek naar een 'tweeling'.

Het oorspronkelijke tweelingenonderzoek stamt uit de psychologie en genetica, waarbij eeneiige tweelingen worden gebruikt om effecten van verschillen in opvoeding op gedrag te achterhalen. Eeneiige tweelingen beschikken immers over precies hetzelfde genetische materiaal en verschillen alleen in de opvoeding of socialisatie die ze hebben gehad—bijvoorbeeld tweelingen die opgroeien in verschillende gezinnen. Experimenteel tweelingenonderzoek heeft dan ook een belangrijke bijdrage geleverd aan het zogenaamde *nature-nurture* debat: zijn verschillen tussen mensen voornamelijk toe te schrijven aan hun opvoeding of aan hun genetische kenmerken (Collins et al.,

2000)? Dit soort onderzoek is later uitgebreid naar zogenaamde sibling-analyses, waarbij geen tweelingen maar kinderen van hetzelfde gezin worden geanalyseerd. Deze kinderen hebben een vergelijkbare familiale achtergrond, maar kunnen bijvoorbeeld verschillen naar opleiding (Van Eijck, 1997).

Als je wil precisiematchen op een combinatie van de kenmerken geslacht, leeftijd en opleidingsniveau bijvoorbeeld, dan neem je voor iedere man jonger dan 45 jaar met een universitaire opleiding in de experimentele groep een man jonger dan 45 jaar met een universitair diploma op. Of voor een zestigjarige vrouw zonder diploma in de experimentele groep zoek je een even oude dame van zestig zonder diploma in de vergelijkingsgroep. Frequentie- of groepsmatchen houdt enkel rekening met elk kenmerk afzonderlijk—zonder de combinatie van kenmerken in stand te houden (Chapin, 1947). Hier zorg je bijvoorbeeld dat voor elke man in de ene conditie, ook een man zit in de andere conditie. Of dat voor iedere 65-plusser in de ene groep, een 65-plusser in de andere groep belandt. Het matchen op een combinatie van kenmerken is bij frequentiematchen niet aan de orde. Zwak matchen ten slotte, houdt alleen rekening met de vergelijkbaarheid van gemiddelde en/of spreiding van een reeks kenmerken tussen de verschillende condities.

Er is echter een fundamenteel probleem met matchen, met name: op welke kenmerken—of combinatie van kenmerken—moet je experimentele en vergelijkingsgroep matchen? Individuele proefpersonen verschillen op wel duizend kenmerken van elkaar en je kan niet op voorhand weten op welke van die kenmerken statistische equivalentie gerealiseerd moet worden. Uiteraard is het niet zinvol om condities te matchen op kenmerken die niet verondersteld worden samen te hangen met de afhankelijke variabele. Die kennis over welke kenmerken wel of niet relevant zijn, haal je uit het beschikbare theoretische materiaal. Dit materiaal is echter niet steeds voor handen. Bovendien wordt matchen op alle relevant geachte kenmerken al snel een moeilijke, zo niet onmogelijke opdracht. Zodra het aantal kenmerken waarop je wil matchen groot wordt—zeker als het om precisiematchen gaat—, dan blijkt het vinden van een gelijkaardige doublure een haast onmogelijke taak.

2.3. Effect van de stimulus

Op welke manier meet je het effect van stimulus X op de afhankelijke variabele? Hoe ga je bijvoorbeeld na of het kijken naar Sesamstraat een effect heeft op een reeks cognitieve vaardigheden bij kinderen? Centraal in het berekenen van het effect van de experimentele stimulus is de opsplitsing in een experimentele en een vergelijkingsgroep.[3] Het effect van de stimulus is immers het verschil tussen de voor- en nameting in de experimentele groep minus het verschil in voor- en nameting in de vergelijkingsgroep. Meer formeel en gebruik makend van de notatie uit Figuur 8.1:

Netto-effect stimulus = $(O_2 - O_1) - (O_4 - O_3)$

waarbij $(O_2 - O_1)$ = bruto-effect stimulus, oftewel: experimentele stimulus + alle storende factoren

en $(O_4 - O_3)$ = alle storende factoren

In de experimentele conditie indiceert het verschil tussen de na- en voormeting $(O_2 - O_1)$ het bruto-effect van de experimentele stimulus. $O_2 - O_1$ omvat echter ook mogelijke veranderingen in de score op de afhankelijke variabele die zijn opgetreden buiten de stimulus om, die dus een precieze, correcte inschatting van het effect van de stimulus verstoren. Die storende factoren—zogenaamde buitenexperimentele gebeurtenissen—kunnen van velerlei aard zijn, zoals onder meer spontane veranderingen/ontwikkelingen in proefpersonen of gebeurtenissen die niets van doen hebben met de stimulus maar die toch de afhankelijke variabele beïnvloeden. Het nut van de vergelijkingsconditie zit hem nu precies in het feit dat je het effect van die storende, buitenexperimentele gebeurtenissen kan becijferen. Het verschil tussen na- en voormeting in de vergelijkingsgroep $(O_4 - O_3)$ geeft namelijk de grootte van hun effect aan. Zo de experimentele groep en vergelijkingsgroep statistisch equivalent zijn—vandaar dat randomisering essentieel is—dan kan je het netto-effect van de stimulus berekenen als het verschil tussen de verschillen in na- en voormeting uit beide condities, of $(O_2 - O_1) - (O_4 - O_3)$. Met andere woorden, er is sprake van een effect van de stimulus als dit netto-effect significant verschillend is van nul.

2.4. Voormeting

De functie van de voormeting bestaat er onder andere in na te gaan of de statistische equivalentie door middel van randomisering wel degelijk geslaagd is. Dankzij randomiseren

moeten de experimentele condities immers op toevalsfouten na gelijkwaardig zijn—op alle, ook op niet-gemeten kenmerken die mogelijk samenhangen met de afhankelijke variabele. Met een voormeting kunnen we die equivalentie alvast checken bij de afhankelijke variabele via een eenvoudige *t*-test voor verschillen tussen O_1 en O_3.

Het moet gezegd dat de voormeting zoals ze in dit ideaaltypische, klassiek experimentele ontwerp wordt beschreven, twee moeilijkheden inhoudt. Ten eerste, is een voormeting niet steeds mogelijk, omdat er geen respons—of de score op een afhankelijke variabele—is zonder stimulus of omdat de voormeting op zijn minst ongepast is. Als je bijvoorbeeld een onderwijsmethode wil toetsen op efficiëntie bij het aanleren van geheel nieuwe vaardigheden bij jonge schoolgaande kinderen, dan kunnen de nieuwe vaardigheden vóór het toedienen van de experimentele stimulus niet gemeten worden omdat ze er eenvoudigweg niet zijn. Of denk maar aan studies over het effect van soorten pleidooien van advocaten in de rechtszaal op uitspraken over schuld/onschuld van een volksjury. Andere illustraties van afwezigheid van respons zonder stimulus zijn legio in de methodologische literatuur omtrent het reduceren van non-respons in surveyonderzoek. In die gevallen kan je gebruik maken van een ontwerp met vergelijkingsgroep en alleen een nameting zoals weergegeven in Figuur 8.2 (‘posttest-only control group design’ in het Engels).

Figuur 8.2. Schematische voorstelling ontwerp met vergelijkingsgroep met alleen nameting.

R	X	O_1
R		O_2

Non-respons in surveyonderzoek baart sociale wetenschappers sinds de jaren ’90 zorgen. Steeds meer mensen weigeren immers om deel te nemen aan het invullen van een enquête of het face-to-face laten afnemen van een interview door een enquêteur. Het is nochtans van groot belang voor de representativiteit van het onderzoek dat de non-respons tot een minimum beperkt blijft. De kans op vertekende resultaten stijgt aanzienlijk bij een grote, selectieve non-respons. Vandaar dat je een uitgebreide literatuur aantreft over manieren om die non-respons te reduceren. Het gebruik van een zogenaamde ‘unconditional, monetary incentive’ is één van de weinige methodes die consistent hogere responspercentages blijkt op te leveren. Die methode baseert zich op de sociale ruiltheorie: als je iets aan iemand geeft, is die persoon geneigd iets

terug te doen als vorm van wederkerigheid (Gouldner, 1960). Hieronder vind je een toepassing van een ontwerp met vergelijkingsgroep en alleen een nameting bij een experiment over het effect van drankbonnetje op de deelname aan een publieksenquête in theaterzalen in Vlaanderen (Roose, 2006). Bij het binnenkomen krijgt de helft van het publiek een enquête toegestopt met de vraag die na de voorstelling in te vullen en terug te geven. De andere helft krijgt bij hun enquête een drankbonnetje te gebruiken in de foyer ter waarde van twee euro. Het deelnamepercentage in de experimentele groep is 80,2% tegenover 70,2% in de vergelijkingsgroep, een verschil van 10 percentagepunten in de lijn van de theorie en statistisch significant ($z = 2{,}40$, $p < 0{,}01$). Het berekenen van het deelnamepercentage is hier dus pas mogelijk na het toedienen van de experimentele stimulus. Er is wel voor randomisering gezorgd: het toeval—het opgooien van een muntje—bepaalde of een contactname gepaard ging met een drankbonnetje of niet.

Ten tweede, een voormeting kan ook een potentieel gevaar of risico inhouden. Het afnemen van een voormeting kan namelijk een invloed uitoefenen op de score in de nameting en zo de interne geldigheid van het experiment ondermijnen. Proefpersonen kunnen bijvoorbeeld de neiging hebben om consistent te zijn in de nameting met wat ze zich herinneren van hun antwoorden in de voormeting. Of de praktische ervaring opgedaan tijdens het invullen van een test in de voormeting kan ervoor zorgen dat ze sowieso beter gaan scoren op een soortgelijke test in de nameting. De voormeting zorgt op die manier eigenlijk voor een soort praktijkervaring die er zonder voormeting niet zou zijn geweest. Beide vallen onder de noemer van de zogenaamde testeffecten (zie paragraaf 3.3 hieronder).

Voorts kan een voormeting ook de externe geldigheid of veralgemeenbaarheid van een experiment ondergraven. De vragen in een voormeting kunnen immers een invloed uitoefenen op de wijze waarop de proefpersonen in de wereld staan. Ze kunnen meer aandacht gaan schenken aan dingen uit hun omgeving die te maken hebben met de inhoud van de vragen uit de voormeting—en dus ook meer gevoelig worden voor de experimentele stimulus zelf. In dat geval is er sprake van een interactie-effect van de voormeting en de experimentele stimulus op de afhankelijke variabele. Een voorbeeld zal één en ander verduidelijken. Stel dat je het effect wil meten op de kennis en seksuele praktijk bij Vlaamse jongeren van een mediacampagne om de verspreiding van seksueel overdraagbare aandoeningen tegen te gaan. In de voormeting leg je vragen voor over wat ze bijvoorbeeld

weten over AIDS, hoe AIDS wordt overgedragen, etc. alsook over hoe vaak ze vrijen, of ze veilig vrijen, welke voorbehoedsmiddelen ze gebruiken, etc. Precies door die voormeting gaan jongeren misschien meer aandacht schenken aan informatie over veilig vrijen verspreid via de media en worden ze zich bewust van de gevaren, wat op zijn beurt hun seksuele praktijk stuurt. Mocht er geen voormeting geweest zijn, zouden ze minder aandacht aan boodschappen omtrent veilig vrijen besteed hebben, en zou de mediacampagne minder effectief geweest zijn. Dus, de aan- of afwezigheid van een voormeting beïnvloedt de relatie tussen campagne en kennis/seksuele praktijk—een typisch voorbeeld van een interactie-effect. Dit interactie-effect treedt uiteraard niet altijd op en hangt af van de inhoud van het onderzoek en aard van de voormeting. Vooral bij opinie- en marketingonderzoek wordt het Solomon vier-groepen-ontwerp frequent gebruikt om vertekening door zulke interactie-effecten tegen te gaan.

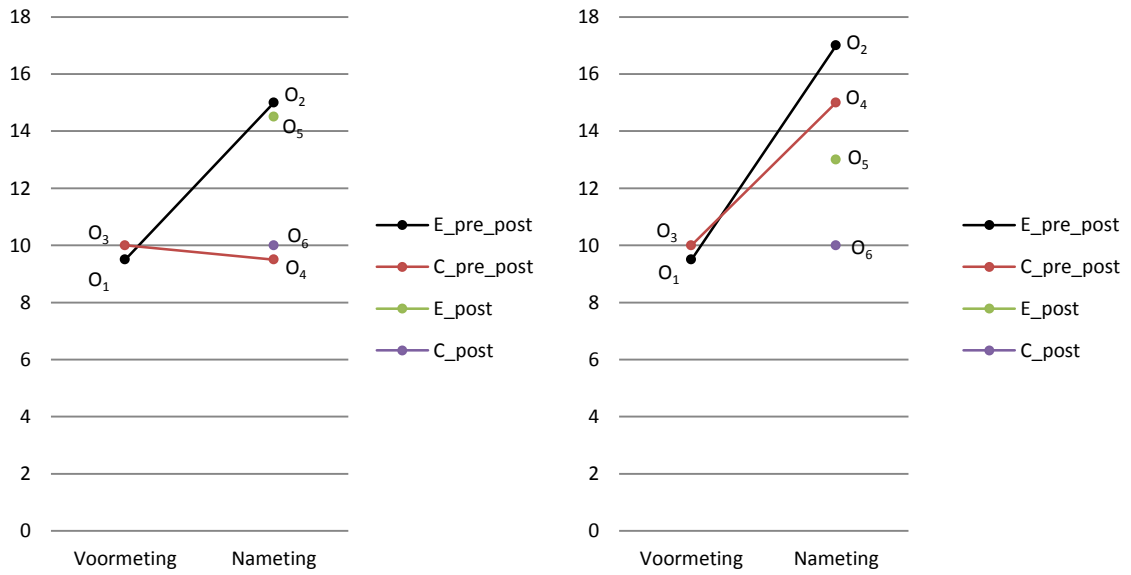
SENSOA
PRAAT OVER SEKS

Het Solomon vier-groepen-ontwerp uit Figuur 8.3 laat toe expliciet rekening te houden met de mogelijk funeste invloed van een voormeting en maakt het mogelijk om zowel het testeffect als het interactie-effect van de stimulus met voormeting te becijferen. Het vier-groepen-ontwerp voegt namelijk twee groepen toe aan het klassieke ontwerp: een experimentele conditie zonder voormeting en een controleconditie zonder voormeting—telkens met randomisering. Als O_2 en O_5 alsook O_4 en O_6 zeer gelijkend zijn en slechts op toevalsfouten van elkaar verschillen, kunnen we besluiten dat er geen effect van de voormeting is. Als die scores wel significant van elkaar verschillen, dan moet je ervan uitgaan dat de voormeting een effect heeft gehad op de afhankelijke variabele. Eén en ander wordt grafisch weergegeven in Grafiek 8.1.

Figuur 8.3. Schematische voorstelling Solomon vier-groepen-ontwerp.

R	O_1	X	O_2
R	O_3		O_4
R		X	O_5
R			O_6

Grafiek 8.1. Weergave effect stimulus, zonder testeffect (links) en effect stimulus met testeffect (rechts).



Als stimulus X een positief effect heeft, dan moeten we de volgende resultaten terugvinden: $O_2 > O_1$ én $O_2 > O_4$ én $O_5 > O_6$ én $O_5 > O_3$. De grafiek links geeft deze situatie weer. De scores op de afhankelijke variabele bij de experimentele situatie mét en zonder voormeting zijn gelijkaardig ($O_2 \approx O_5$). Dit is evenzo bij de vergelijkingsgroep mét en zonder voormeting ($O_4 \approx O_6$). Mét een testeffect of interactie-effect tussen voormeting en stimulus op de grafiek rechts liggen de waarden op de afhankelijke variabele tussen de experimentele groepen mét en zonder voormeting ($O_2 \neq O_5$) en de controlegroepen mét en zonder voormeting ($O_4 \neq O_6$) een stuk uiteen. Ofwel, meer formeel, dan is: $O_5 – O_6 \neq O_2 – O_4$. De grootte van het effect van de stimulus is met andere woorden afhankelijk geworden van de aan- of afwezigheid van een voormeting.

3. Bedreigingen van de interne geldigheid

De logica van een experimenteel ontwerp is dat alleen de onafhankelijke variabele—lees: de stimulus—de afhankelijke variabele beïnvloedt. Zodra een ander kenmerk de score op de afhankelijke variabele stuurt, dan is de interne geldigheid bedreigd. Dat wil zeggen dat je niet met aan zekerheid grenzende waarschijnlijkheid het effect op de afhankelijke variabele causaal kan toeschrijven aan de experimentele stimulus. Basisidee is namelijk dat experimentele en vergelijkingsgroep alleen verschillen in het al dan niet toegediend krijgen van een stimulus en dat ze op alle andere kenmerken statistisch equivalent zijn. Wanneer je de mogelijke effecten van andere variabelen op de afhankelijke variabele niet kan uitschakelen via het experimentele ontwerp, is het mogelijk dat je die storende effecten verwart met de invloed van de experimentele stimulus en is de interne geldigheid bedreigd. In het Engels spreek je van zogenaamde ‘confounding variables’. Dat zijn storende factoren in experimenteel onderzoek die geen onderdeel vormen van de hypothese die getest wordt—het effect van de stimulus—maar die wel samenhangen met de kenmerken van interesse en zo de oorzakelijke redenering kunnen bedreigen of ondergraven. Er zijn met andere woorden alternatieve verklaringen denkbaar én mogelijk voor een eventuele verandering in de afhankelijke variabele. Storende factoren kunnen van velerlei aard zijn en hierna bespreken we er enkele.[4]

3.1. Buitenexperimentele gebeurtenissen

Dit zijn gebeurtenissen of trends die plaatsvinden tussen de voor- en nameting, los van de experimentele stimulus, en die een invloed uitoefenen op de afhankelijke variabele. In het Engels wordt hiervoor ook de term ‘history’ gebruikt. Hoe groter de periode tussen voor- en nameting, hoe groter vanzelfsprekend de kans dat er zich buitenexperimentele gebeurtenissen voordoen. Stel dat je bijvoorbeeld een onderzoek verricht naar het effect van de informatiecampagnes van het *Nucleair Forum*, dat als belangengroep de mogelijkheden van kernenergie en haar toepassingen bekend wil maken bij de brede bevolking. Eén week na de voormeting is er een meltdown in één van de reactoren van de kerncentrale van Fukushima (Japan) ten gevolge van een aardbeving en de daaropvolgende tsunami. Hierdoor wordt het nieuws en aanverwante duidingsprogramma’s wekenlang gemonopoliseerd door items over kernenergie, gevolgen van bloostelling aan straling, radioactiviteit, etc. Je kan aannemen dat zo’n buitenexperimentele gebeurtenis, zoals de meltdown van een kernreactor met massale mediabelangstelling, zowel de scores bij de experimentele als de vergelijkingsgroep stuurt.

Een andere vorm van buitenexperimentele gebeurtenis is de zogenaamde ‘intra-session history’ of ‘local history’. Dit zijn unieke gebeurtenissen in die zin dat ze alleen in de

experimentele óf in de vergelijkingsgroep optreden en zich dus opwerpen als mogelijke alternatieve verklaringen voor het effect van de experimentele stimulus. Het kan gaan om inleidende opmerkingen van de onderzoeker, haar/zijn grapjes, een brand aan de overkant van de straat, etc.

3.2. Maturatie of spontane veranderingen

Maturatie verwijst naar het proces dat proefpersonen gedurende de looptijd van een experiment veranderen op psychologisch, biologisch of emotioneel vlak. De veranderingen zijn niet een gevolg van een specifieke gebeurtenis—zoals hierboven—maar zijn gewoon te wijten aan het verstrijken van de tijd *per se*. Die veranderingen gebeuren dus 'spontaan'. Proefpersonen worden bijvoorbeeld ouder, vermoeid, verveeld, slaperig, minder geconcentreerd tijdens hun deelname aan een experiment, wat de score op de onafhankelijke variabele kan beïnvloeden los van de experimentele stimulus.

3.3. Testeffect

Bij een testeffect heeft het ondergaan van een voormeting een invloed op de scores in de nameting. Proefpersonen kunnen de neiging hebben om consistent te zijn doorheen het experiment en soortgelijke antwoorden willen geven op voor- en nameting. Dan stuurt de herinnering aan de antwoorden in de voormeting de score op de afhankelijke variabele in de nameting—en is niet alleen de experimentele stimulus ervoor verantwoordelijk. Een andere illustratie van een testeffect vind je in de literatuur over IQ-tests. Zo blijken studenten die voor een tweede keer dezelfde of zelfs een lichtjes andere IQ-test invullen, het over het algemeen beter te doen dan op hun eerste test (Kaufman, 2009). Hiervoor reserveert men ook wel eens de term 'practice effect': studenten zijn niet intelligenter geworden door het invullen van die eerste test, ze hebben zich de testprocedures eigen gemaakt—ze zijn met andere woorden gewoon beter geoefend in het invullen van IQ-testen.

3.4. Instrumentatie

Met instrumentatie wordt verwezen naar veranderingen in betrouwbaarheid van het gebruikte meetinstrument. Veelal ligt falend materiaal hiervan aan de oorsprong, zoals bijvoorbeeld een weegschaal die naarmate meer proefpersonen worden gewogen, minder gevoelig wordt. Hierdoor weegt iedereen in de nameting ogenschijnlijk minder dan in de

voormeting—dit komt niet door een werkelijke gewichtsafname, maar door het slecht werken van de weegschaal. Ook het wijzigen van het personeel dat de experimenten uitvoert en/of begeleidt, kan een bijkomend effect genereren.

3.5. Statistische regressie

Statistische regressie—ook wel regressie naar het gemiddelde genoemd—kan voorkomen wanneer groepen worden geselecteerd op basis van extreem hoge of extreem lage scores op de afhankelijke variabele. Stel dat leerlingen uit het eerste leerjaar worden gekozen om deel te nemen aan extra lessen, omdat ze heel slecht scoorden op een toets rekenvaardigheid. Met die extra lessen, die zijn in dit geval de experimentele stimulus, wil de school hun lage scores opkrikken. Bij een tweede, identieke of soortgelijke toets in rekenvaardigheid zouden de scores van diezelfde leerlingen gemiddeld bijna zeker hoger zijn dan op de eerste toets. Die stijging zou plaatsvinden niet zozeer vanwege een werkelijk effect van de extra lessen of door het hierboven beschreven 'practice effect', maar wel door toevalsfouten in de voor- en nameting. Toevalsfouten bij extreme waarden worden immers verondersteld hoger te zijn dan bij gemiddelde scores. Louter door toeval dus zullen de grote toevalsfouten in de eerste meting niet meer voorkomen in de tweede meting en de extreem lage scores van de leerlingen uit de eerste rekentest zullen sowieso richting gemiddelde evolueren in de tweede toets. Die verschuiving of regressie naar het gemiddelde heeft met andere woorden louter te maken met een correctie van de per toeval te laag gemeten scores op de eerste rekenvaardigheidstest. Of in meer technische termen gesteld: statistische regressie is een onvermijdelijk nevenproduct van de niet perfecte test-hertest correlatie voor groepen die geselecteerd zijn op hun extreme waarden op de afhankelijke variabele.

3.6. Selectie

Selectie verwijst naar een verschillende samenstelling van de groepen die je wil vergelijken. Normaal gezien zorgt randomisering bij de samenstelling van de experimentele en vergelijkingsgroep ervoor dat ze equivalent zijn op alle kenmerken op t_0—gemeten en ongemeten. Dit kan je nagaan door te controleren of de waarden op de variabele(n) van interesse tussen de verschillende condities in de voormeting op toevalsfouten na gelijk aan elkaar zijn. Zo is het makkelijk te begrijpen dat equivalentie beter verzekerd is bij grote

aantallen proefpersonen die op toevallige wijze toegewezen worden, dan bij kleine aantallen. Dus randomisering moge dan wel niet steeds perfect werken, het is wel de enige—en essentiële—remedie tegen een verschillende samenstelling van de experimentele en de vergelijkingsgroep.

3.7. Uitval

Bij uitval of zogenaamde experimentele mortaliteit gaat het om proefpersonen die tijdens de duur van het experiment (tussen t_0 en t_1) uitvallen of afhaken. Als veel proefpersonen afhaken halfweg het experiment, dan kunnen we niet met zekerheid weten of de resultaten hetzelfde zouden zijn mochten ze niet zijn uitgevallen. Een experimenteel onderzoek naar het effect van een bepaald opleidingsprogramma of onderwijsmethode bijvoorbeeld neemt al snel enkele weken, soms zelfs maanden, in beslag. Wanneer na verloop van tijd enkele leerlingen afhaken, dan kan dit aanleiding geven tot subtiele vertekeningen. Zo de uitval louter toevallig gespreid is over de verschillende condities, dan is er niets aan de hand. Dan blijft de samenstelling van de experimentele en vergelijkingsgroep equivalent en zijn de groepen vergelijkbaar. Wanneer echter in de experimentele groep bijvoorbeeld vooral minder goed presterende leerlingen uitvallen na verloop van tijd—en met andere woorden de uitval afhankelijk is van de condities—dan krijg je differentiële uitval en dus een verschillende samenstelling van de groepen die je wil vergelijken met mogelijk ernstige vertekening tot gevolg. Het is dus boodschap om van de proefpersonen die uitvallen, kenmerken bij te houden, zodat je eventueel nadien kan nagaan hoe selectief de uitval is.

Verder kan uitval ook te wijten zijn aan ethische bekommernissen. Stel dat je in een onderzoek naar de werking van een geneesmiddel vaststelt dat sommige proefpersonen in de experimentele groep onverwachte, kwalijke neveneffecten en bijwerkingen van het medicament ondervinden. Dan wordt uiteraard de behandeling voor die personen stopgezet en door hun uitval is de oorspronkelijke samenstelling van de experimentele groep niet meer dezelfde.

3.8. Verwachtingen onderzoeker

Ook het gedrag van de onderzoeker kan de interne geldigheid van een experiment beïnvloeden. Hij/zij heeft doorgaans bepaalde verwachtingen omtrent de afloop van het experiment en kan die verwachtingen op de proefpersonen projecteren. Meestal gebeurt

dit niet op een expliciete manier, maar op een indirecte, onbewuste en ongewilde manier door bijvoorbeeld bepaalde vormen van non-verbale communicatie, oogcontact, toon van de stem, etc. Door de reacties van de proefleider op gedrag van de proefpersonen, kan hij/zij ongewild net dat gedrag gaan aanmoedigen of bevestigen dat in de lijn ligt van de resultaten die hij/zij verwacht of bevestigd wenst te zien.

Vandaar de noodzaak om te werken met een zogenaamd double-blind experiment. In dit geval weten de onderzoekers die in contact komen met de proefpersonen, niets af van de details van de hypothese die wordt onderzocht of van de stimulus die al dan niet wordt toegediend. 'Blind' verwijst naar het feit dat de proefpersonen niet weten of ze tot de experimentele dan wel de vergelijkingsgroep behoren, 'double blind' duidt op het idee dat noch de proefleider, noch de proefpersonen op de hoogte zijn van de details van het experiment en de onderzochte hypothese. Op die manier kan je het effect van verwachtingen van de onderzoeker uitschakelen. In experimenteel onderzoek naar effecten van medische behandelingen is double-blind de standaard, waarin de vergelijkingsgroep veelal een placebo krijgt toegediend. Een placebo is een stimulus die geen werkzame bestanddelen heeft maar die enkel vormelijk gelijkt op de echte stimulus—in het geval van geneesmiddelen is dit een medicijn dat geen actieve, genezende stof bevat, maar wel in de vorm van een pil wordt aangeboden. Noch de proefpersonen, noch de proefleider weten zo of hij/zij een werkelijke stimulus toedient of toegediend krijgt (een echt medicijn) of slechts een placebo (een 'leeg' medicijn zoals bijvoorbeeld een suikerpil).

4. Bedreigingen van de externe geldigheid

Tot zover hebben we bedreigingen voor de interne geldigheid besproken. Dit zijn bedreigingen die rechtstreeks de score op de afhankelijke variabele systematisch sturen—een score die dus verkeerdelijk geïnterpreteerd kan worden als het resultaat van de experimentele stimulus. Externe geldigheid heeft te maken met het feit of en in hoeverre je bevindingen uit een experiment kan veralgemenen buiten de strikte context van de experimentele situatie. Externe geldigheid staat dus gelijk met veralgemeenbaarheid. Wanneer een experiment externe geldigheid ontbeert, dan zijn de bevindingen strikt genomen niet veralgemeenbaar buiten de experimentele setting. Meer technisch gaat het dus om een interactie-effect van de experimentele stimulus met een beperkte set van

condities op de afhankelijke variabele, waardoor het effect van de stimulus mogelijk slechts beperkt is tot die specifieke set van condities. Als experimentele resultaten niet toepasbaar zijn op andere, meer alledaagse natuurlijke settings of andere populaties, dan beperkt dit de waarde van inzichten uit zulk onderzoek, alsook de aantrekkingskracht van dit 'ideale' design voor sociaalwetenschappelijk onderzoek.

Strikt genomen—en dit wordt vaak vergeten—is veralgemenen als vorm van inductie nooit volkomen te rechtvaardigen op basis van vastgelegde methodologische of statistische criteria. Dit is in tegenstelling tot bedreigingen van de interne validiteit die binnen de limieten en logica van de inductieve statistiek kunnen worden beschreven en geduid. Externe geldigheid houdt altijd een oefening in extrapoleren in: je wil steeds de bevindingen uit een specifieke setting en met een bepaalde steekproef aan proefpersonen gaan toepassen buiten die setting en op alle mensen. Dit is niet op logica gebaseerd, maar op cumulatie van ervaringen uit vroeger onderzoek. Je moet bepaalde veronderstellingen en assumpties maken om de resultaten te veralgemenen vanuit de specifieke condities van het experiment, te weten onderzoek op proefpersonen met een specifieke opleiding, leeftijd en geslacht in een bepaalde geografische regio op een bepaald moment met die specifieke oriëntatie van het magnetische veld, luchtdruk, etc. Uit ervaring weten we dat die oriëntatie van het magnetische veld geen effect heeft op de manier waarop een onderwijsmethode de leerprestaties van zesjarigen beïnvloedt. De bedreigingen van externe geldigheid zijn dus vermoedens over wat redelijkerwijs zou kunnen interageren met de experimentele stimulus—en dus de veralgemeenbaarheid beperken—en vermoedens over wat we kunnen negeren.[5]

Externe geldigheid kan betrekking hebben op twee dingen: de mate waarin resultaten uit het experiment veralgemeenbaar zijn naar een ruimere populatie enerzijds en naar een natuurlijke, levensechte setting buiten de strikt gecontroleerde en artificiële context van het laboratorium anderzijds. Dit geeft aanleiding tot twee soorten externe geldigheid, te weten populatie- en naturalistische geldigheid.

4.1. Populatiegeldigheid

Populatiegeldigheid heeft betrekking op de mate waarin je de resultaten uit een experiment dat is gevoerd bij een specifieke steekproef van proefpersonen, kan veralgemenen naar een ruimere populatie. Veralgemenen naar populaties buiten de bevolking waaruit een

steekproef getrokken is, vergt argumentatie en eventueel empirisch bewijsmateriaal. In de V.S. wordt heel wat experimenteel psychologisch onderzoek gevoerd op eerstejaars studenten psychologie. Zolang de deelname als een leerervaring voor de student kan doorgaan, wordt het als ethisch en gepast beschouwd ze als proefpersonen in een experiment te laten fungeren. Een probleem is uiteraard of je resultaten bekomen bij een populatie psychologiestudenten kan veralgemenen naar andere studentenpopulaties (bijvoorbeeld studenten wiskunde, politieke en sociale wetenschappen of letteren) of—sterker nog—naar de ruimere bevolking? Het is steeds aan te raden experimentele resultaten te repliceren bij andere populaties om de externe geldigheid van de bevindingen te vergroten.

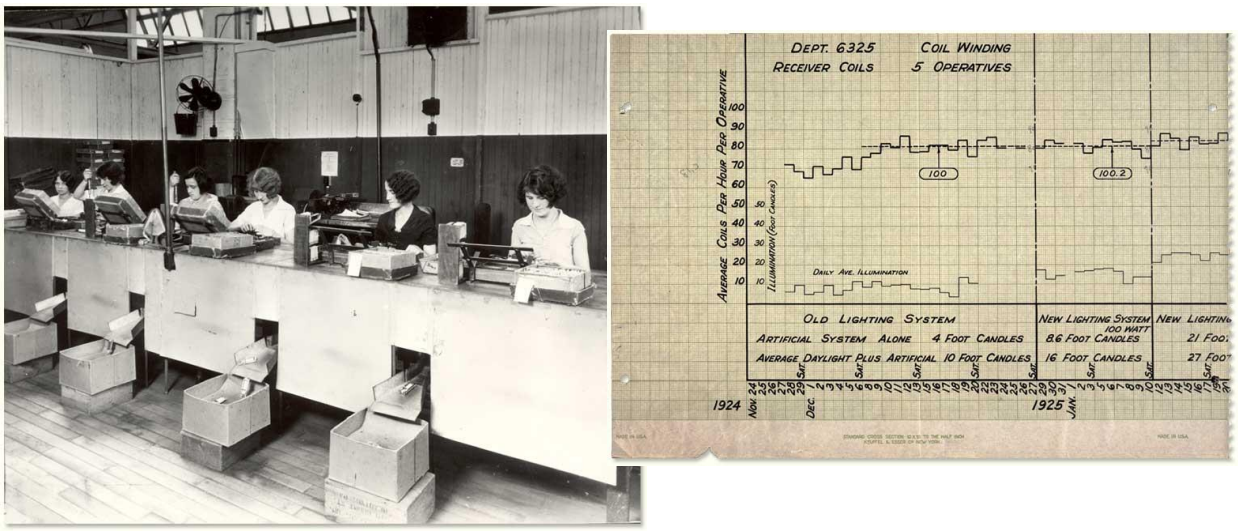
> Volgens Henrich en collega's hebben heel veel Westerse studies met problemen van populatiegeldigheid te kampen—zeker als je wil veralgemenen naar 'de mens' (Henrich, Heine & Norenzayan, 2010). Doorgaans zijn resultaten van onderzoek naar variatie in gedrag bij de mens gebaseerd op Westerse populaties. En dit is volgens hen een ongewone subpopulatie van de menselijke bevolking. Zij reserveren er het acroniem 'WEIRDS' voor: met name mensen van "Western, Educated, Industrialized, Rich, and Democratic Societies" (Henrich, Heine & Norenzayan, 2010: 61).

4.2. Naturalistische geldigheid

Bij naturalistische geldigheid (of 'ecological validity' in het Engels) staat de vraag centraal of je bevindingen uit de artificiële, hoogst kunstmatige experimentele situatie kan doortrekken naar de 'echte' sociale wereld. Kan je resultaten bekomen uit testen in een gecontroleerde, kunstmatige laboratoriumsetting zomaar extrapoleren naar 'real-life' situaties? Twee begrippen zijn relevant om die vraag te beantwoorden, met name alledaags realisme en reactiviteit. Alledaags realisme verwijst naar de mate waarin experimentele condities lijken op situaties die je ook in het dagelijkse leven terugvindt. Stel dat je geïnteresseerd bent in het effect van een methode om de werking van het geheugen aan te scherpen en als test vraag je proefpersonen vierletterwoorden te memoriseren. Het alledaags realisme van het experiment zou groter zijn, mocht je niet met lijsten van vierletterwoorden werken om de onderwijsmethode te testen, maar wel bijvoorbeeld met artikels uit een krant die proefpersonen binnen een bepaalde tijd moeten lezen en waarover dan feitenvragen worden gesteld. Reactiviteit is de mate waarin mensen zich anders gaan gedragen in de

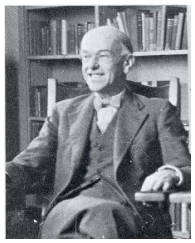
experimentele setting dan in de ‘echte’ wereld, precies omdat ze zich bewust zijn dat ze onderdeel uitmaken van een studie. Met andere woorden, er gaat een effect uit van het meten op datgene wat er gemeten wordt. Een klassieke illustratie van reactiviteit is de studie van Roethlisberger & Dickson (1939) in de Hawthorne vestiging van de Western Electric Company in Chicago in de jaren ‘20-’30 (zie ook Franke & Kaul, 1976; Gillespie, 1991).

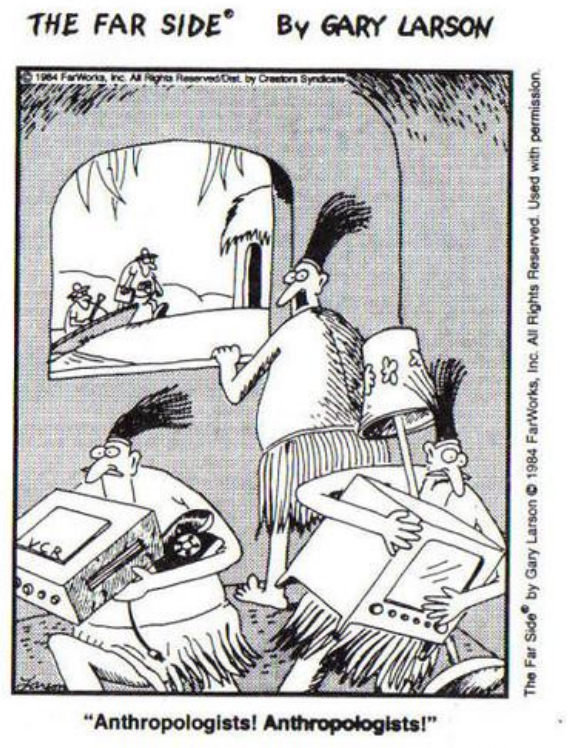
Tijdens de jaren ’20 en begin jaren ’30 vinden een reeks veldexperimenten plaats in de Hawthorne vestiging van de Western Electric Company in Chicago. In deze fabriek wordt allerlei elektrische apparatuur massaal geproduceerd, zoals verlichting, alarminstallaties, onderdelen van telegrafen zoals relais, etc. De context van de experimenten is een discussie over de productiviteit van arbeiders en de wijze waarop je die productiviteit kan verhogen. Toen vigeerde het Taylorisme dat een verregaande rationalisering van het massaproductieproces voorstond: via een doorgevoerde arbeidsdeling wordt de productie uiteengerafeld in kleine handelingen afgestemd op de kennis en vaardigheden van iedere werknemer. Bijkomend streng toezicht op de arbeidsdiscipline moet dan zorgen voor een efficiënte productie. Verschillende condities in de fabriek worden gemanipuleerd. In de zogenaamde ‘illumination studies’ wordt de belichting in de fabriek gemanipuleerd en nagegaan in welke mate de intensiteit van de belichting een invloed heeft op de productiviteit. Een sterkere belichting geeft aanleiding tot een hogere productie. Een zwakke belichting geeft echter ook aanleiding tot een hogere productie dan bij de voormeting.



Bij de ‘relais assembly studies’ worden enkele werknemers geïsoleerd in een kleine groep en worden de frequentie en duur van de pauzes gemanipuleerd alsook de

verloning. En wat blijkt? De productiviteit gaat bijna steevast de hoogte in—zelfs bij een terugkeer naar de oorspronkelijke situatie van vóór de experimenten. Eén interpretatie—die van Elton Mayo (1880-1949), alvast voor de relais assembly experimenten—is dat een goede verstandhouding tussen de werknemers in de groep zorgen voor een goed moreel en dus voor een hogere productiviteit. Deze interpretatie geeft aanleiding tot het ontstaan van de zogenaamde human relations beweging in management, waarin nu ook aandacht wordt besteed aan de sociale omgeving en arbeidstevredenheid van de werknemer ter verklaring van productiviteit (Mayo, 1933). Een andere interpretatie, eentje die van belang is in de discussie omtrent de veralgemeenbaarheid van experimentele resultaten naar andere settings, is dat de productiviteit stijgt vanwege de aandacht die de arbeiders krijgen van de onderzoekers. Vandaar dat het Hawthorne-effect begrepen moet worden als een effect op een afhankelijke variabele als gevolg van de aandacht van onderzoekers. De bevindingen kunnen dan uiteraard niet zomaar veralgemeend worden naar situaties waarin de aandacht van onderzoekers afwezig is.





In laboratoriumexperimenten is het alledaags realisme en bijhorende naturalistische geldigheid doorgaans laag en het experimenteel realisme hoog. In veldexperimenten, die plaatsvinden 'in het veld' dat wil zeggen in natuurlijke settings zoals de fabriek, de metro, de klas, etc., zijn veelal beide hoog—maar veldexperimenten missen dan weer de controle op

mogelijke storende factoren die de interne geldigheid kunnen ondergraven. Het gaat dus om een afweging van interne en externe geldigheid bij de keuze voor een laboratorium- dan wel een veldexperiment. Goede sociale wetenschap is uiteraard gediend met beide. Een andere bron van bedreiging voor de externe geldigheid is een interactie-effect tussen stimulus en voormeting, zoals beschreven in paragraaf 2.4 als we over de gevaren van een voormeting hebben besproken.



In Groningen is in 2008 een reeks veldexperimenten opgezet om na te gaan of en in hoeverre de aanwezigheid van norm-overtredend handelen (bijvoorbeeld de aanwezigheid van graffiti of zwerfvuil) andere normovertredingen, zoals stelen, stimuleert en in de hand werkt (Keizer, Lindenberg & Steg, 2008). Eén van de bekommernissen van de onderzoekers is om de situaties in de experimentele en vergelijkingsgroep levensecht te laten lijken—zoals je kan zien op de foto's. Zo hebben ze een vergelijkbaar steegje in Groningen mét en zonder graffiti nagegaan of toevallige passanten meer of minder geneigd waren het witte publiciteitsbriefje dat aan hun fiets was vastgemaakt, op de grond te gooien of mee te nemen. Zo blijkt dat 69% (van 77 proefpersonen) het briefje op de grond gooit in de 'graffiti-conditie' tegenover 33% (ook van 77 proefpersonen) in het 'propere' steegje. Of wanneer een envelop uit de brievenbus steekt waarin duidelijk zichtbaar een briefje van vijf euro zit, dan blijkt dat in omstandigheden waar norm-overtredend gedrag is opgetreden—bijvoorbeeld als de brievenbus zich in een buurt bevindt met graffiti of met veel zwerfvuil—de kans groter is dat mensen dit bankbriefje zullen stelen (respectievelijk 27% en 25% tegenover 13% zonder graffiti of zwerfvuil). Deze bevindingen leggen een reeks causale mechanismen bloot onderliggend aan de 'Broken Windows theory'.

5. Alternatieven op de klassieke experimentele ontwerpen

Het is niet steeds mogelijk om aan de strenge eisen van het klassieke experiment te voldoen—en zeker niet in de sociale wetenschappen waar je als onderzoeker niet altijd volledige controle hebt over wie op welk moment de experimentele stimulus krijgt toegediend. Bovendien krijg je in veldexperimenten doorgaans niet de kans of mogelijkheid om proefpersonen op volkomen toevallige wijze aan een experimentele en vergelijkingsgroep toe te wijzen. Randomisering is dan onmogelijk. Vaak is er sprake van zelfselectie: de groepen zijn 'in vivo' al aanwezig en hebben zichzelf uit eigen keuze al dan niet blootgesteld aan de stimulus.

Dit deel gaat dieper in op de mogelijkheden om vanuit praktische of ethische redenen van het ideale design af te wijken en gebruik te maken van alternatieve ontwerpen, de zogenaamde quasi- en pre-experimentele designs. Quasi-experimentele ontwerpen hebben een aantal kenmerken gemeen met een zuiver experimenteel design, maar missen of toevallige toewijzing aan condities, of een vergelijkingsgroep, of een voor- en nameting. Bij pre-experimentele ontwerpen zijn de verschillen met het zuiver experiment nog meer uitgesproken: ze ontberen meerdere essentiële kenmerken van het zuivere experiment. Wel hebben ze alle gemeen dat ze makkelijker haalbaar zijn in de praktijk, maar strikte garanties op interne geldigheid ontberen. Uitspraken over causaliteit en causale relaties moeten daarom met nog meer omzichtigheid gebeuren dan bij klassieke experimentele ontwerpen al het geval is. Of zoals Samuel Stouffer, de prominente Amerikaanse socioloog en methodoloog, opmerkt: “Sometimes, believe it or not, we have only one cell [i.e. één meetmoment bij één groep]. When this happens, we do not know much of anything. But we can still fill pages of social science journals with “brilliant analysis” if we use plausible conjecture in supplying missing cells from our imagination. […] The tragicomic part is that most of the public, including, I fear, many social scientists, are so accultured that they ask for no better data.” (Stouffer, 1950: 357-358)

Sociologische verbeelding mag dan al belangrijk zijn, een goed geoefend inzicht in onderzoeksmethoden en hun mogelijkheden en beperkingen is dat nog meer. En juist omdat bij quasi- en pre-experimentele ontwerpen zo’n volledige experimentele logica ontbreekt, is het belangrijk dat onderzoekers in staat zijn om te beoordelen op welke aspecten die ontwerpen (on)voldoende garanties bieden om oorzakelijke redeneringen toe te laten.

5.1. Quasi-experimentele ontwerpen

5.1.1. Tijdreeksontwerp

Centraal in een tijdreeksontwerp is het herhaaldelijk en op regelmatige basis meten van scores op een afhankelijke variabele (O_1, O_2, O_3, O_4, etc.) in één enkele conditie (zie Figuur 8.4). In die reeks metingen wordt een experimentele stimulus X geïntroduceerd, een vergelijkingsgroep is afwezig. De metingen die vóór de introductie van X gebeuren, kan je als voormetingen beschouwen. Zij die na de stimulus plaatsvinden, zijn nametingen. De groepen bij wie de herhaalde metingen gebeuren, hoeven ook niet steeds uit precies

dezelfde personen te bestaan. Soms nemen dezelfde proefpersonen aan alle metingen deel, soms gaat het bij iedere meting om een andere toevalsteekproef uit eenzelfde populatie. Tijdreeksontwerpen komen geregeld voor, denk maar aan overheden of organisaties (scholen, bedrijven) die administratieve statistieken bijhouden over bijvoorbeeld werkloosheid, verkeersongevallen, criminaliteit, productiviteit, etc. en die het effect van een beleidsmaatregelen willen proberen in te schatten op trends en ontwikkelingen in die cijferreeksen.

Figuur 8.4. Schematische voorstelling tijdreeksontwerp.

O_1 O_2 O_3 O_4 X O_5 O_6 O_7 O_8

Eigenlijk vormt het tijdreeksontwerp—mét maar vaak zónder vergelijkingsgroep—het geliefkoosde experimentele ontwerp van heel wat onderzoek in de natuurwetenschappen. Fysici meten bijvoorbeeld herhaaldelijk de valversnelling van een object met een bepaalde massa, waarna ze bij een soortgelijk object—op een verschil in massa na—opnieuw herhaaldelijk de valversnelling meten. Waarom kan dit design dan toch niet op dezelfde status rekenen binnen de sociale wetenschappen als in de natuurwetenschappen? De grootste bedreiging voor de interne geldigheid van het tijdreeksontwerp zijn buitenexperimentele gebeurtenissen. Als je niet kan controleren voor gebeurtenissen die ongeveer tegelijk met de experimentele stimulus plaatsvinden, dan kan je nooit met zekerheid die rivaliserende verklaringen ontkrachten—een grote zwakte dus. De sociale werkelijkheid laat immers zelden toe alle gebeurtenissen die enigszins relevant kunnen zijn voor de score op de afhankelijke variabele, in kaart te brengen, laat staan te controleren of te becijferen. Natuurwetenschappers daarentegen werken binnen strikt gecontroleerde laboratoriumcondities die ervoor zorgen dat je de stimulus wel perfect experimenteel kan isoleren en dat je in steeds dezelfde, constante omstandigheden werkt. Onder de noemer van buitenexperimentele gebeurtenissen zou je ook het effect van weersomstandigheden en seizoenen kunnen catalogeren. Tijdreeksen lopen immers doorgaans over een lange tijdsperiode, waardoor seizoenfluctuaties in weersomstandigheden, temperatuur, lichtintensiteit, etc.—zo ze samenhangen met de afhankelijke variabele—verward kunnen worden met de introductie van de experimentele stimulus. Denk bijvoorbeeld maar aan

verkeersongevallen en weersomstandigheden, productiviteit van arbeiders en eet-slaappatronen die samenhangen met seizoenen.

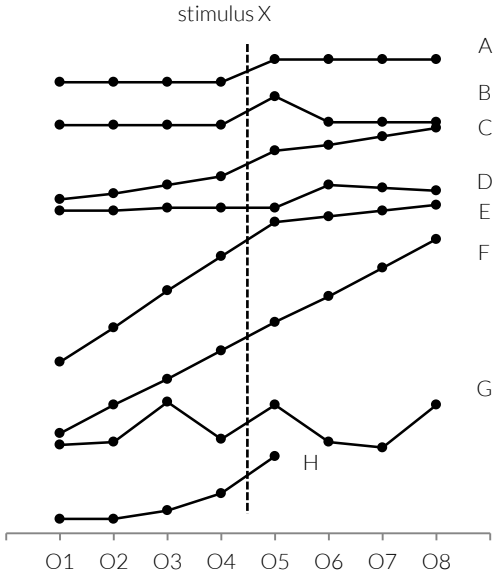
Net door de herhaalde metingen over een lange periode laat dit ontwerp toe om enkele bedreigingen van de interne geldigheid voor een stuk af te weren en mogelijke rivaliserende verklaringen te weerleggen—of toch alvast hun geldigheid enigszins te ontkrachten. Maturatie, testeffect en instrumentatie—en eigenlijk ook buitenexperimentele gebeurtenissen—kunnen hierdoor wat ingeschat worden. Waarom zouden ze immers alleen tussen O_4 en O_5 voor verandering in de afhankelijke variabele zorgen en niet tussen de andere metingen? Met andere woorden, het zou wel heel erg toevallig zijn mochten de storende factoren zich precies op het moment van de stimulus manifesteren zonder ook op andere momenten in de tijdreeks voor effect te zorgen. Vandaar dat bij de analyse van tijdreeksen er steeds voldoende metingen moeten gebeuren en dat eventuele trendbreuken en hun mogelijke oorzaken grondig moeten worden onderzocht. IJzersterke garanties met betrekking tot causale redeneringen zoals bij een klassiek experiment, kan een tijdreeksontwerp echter niet geven. Boodschap is dus om materiaal uit tijdreeksen niet te vlug in causale termen te interpreteren en voorzichtig om te springen met al te verregaande conclusies, als de resultaten niet in een diversiteit aan settings, omstandigheden, etc. gerepliceerd zijn geweest.

Bij tijdreeksen stelt zich bovendien het probleem wanneer het legitiem is om te spreken van een effect van de experimentele gebeurtenis. Het is immers niet steeds duidelijk als een bepaald patroon in tijdreeksdata geïnterpreteerd kan worden in termen van een trendbreuk als gevolg van stimulus X of niet. In Figuur 8.5 vind je een aantal mogelijk patronen van tijdreeksen na het introduceren van een experimentele gebeurtenis X (naar Campbell & Stanley, 1963: 38).

In de volgende gevallen zou je hoogstwaarschijnlijk besluiten dat X en effect heeft gehad: A, B en mogelijk ook C, D en E. Bij A is er een onmiddellijk effect dat bovendien duurzaam blijkt te zijn. Bij B laat het effect zich ook onmiddellijk voelen, maar gaat het veeleer om een korte termijn effect. C is vergelijkbaar met A, zij het dat het effect zich bovenop een opwaartse trend manifesteert. In tijdreeks D is er al meer onduidelijkheid en kan je spreken van een zogenaamd ‘lagged effect’ of een uitgesteld effect van stimulus X dat zich pas bij O_6 laat optekenen. Bij E is er ofwel sprake van een effect van X die een bestaande trend afremt, ofwel heb je te maken met een plafondeffect waarbij een initieel lineaire ontwikkeling

afbuigt als een bovengrens wordt benaderd. Of je kan hierbij bijvoorbeeld ook denken aan een logistische evolutie. Voor reeksen F, G en H zijn er niet veel argumenten om een causale redenering toe te laten. Iedere interpretatie in termen van een effect van X is uiterst speculatief en op basis van het tijdreeksontwerp alleen eigenlijk niet of nauwelijks gerechtvaardigd.

Figuur 8.5. Mogelijke tijdreekspatronen met introductie van stimulus X (naar Campbell & Stanley, 1963: 38).



5.1.2. Niet-equivalent ontwerp met vergelijkingsgroep

Een zeer populair quasi-experimenteel ontwerp dat veel gelijkenis vertoont met het klassiek experimenteel design, is het niet-equivalent ontwerp met vergelijkingsgroep ('nonequivalent control group design' in het Engels). Hierin werk je met een voor- en nameting bij een experimentele en vergelijkingsgroep, maar er is geen randomisering toegepast om statistische equivalentie en dus vergelijkbaarheid tussen beide condities te garanderen—vandaar niet-equivalent (zie Figuur 8.6). Veelal gaat het om bestaande groepen die vóór het experiment al een soort eenheid vormen, zoals klassen in scholen, afdelingen, teams of ploegen in organisaties, etc. Het toedienen van de experimentele stimulus aan één van de groepen moet wel op toevallige wijze gebeuren en wordt verondersteld onder de controle van de onderzoeker te zijn.

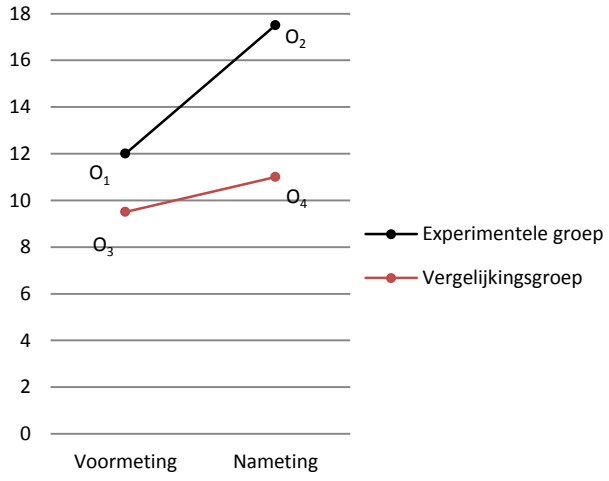
Figuur 8.6. Schematische voorstelling niet-equivalent ontwerp met vergelijkingsgroep.

O_1 X O_2

O_3 O_4

Precies het feit dat de groepen niet op toevallige wijze via randomisering zijn samengesteld, maar dat dit ontwerp gebruik maakt van vooraf reeds bestaande groepen—wat de praktische toepasbaarheid en haalbaarheid verhoogt—, vormt een inherente zwakte. Via het verschil tussen de experimentele en vergelijkingsgroep in de voormeting kan je alvast evalueren in hoeverre de condities vergelijkbaar zijn met betrekking tot de afhankelijke variabele ($O_3 – O_1$). Als andere kenmerken van de proefpersonen ter beschikking zijn die zouden kunnen samenhangen met de afhankelijke variabele, dan doe je er goed aan ook die te controleren op vergelijkbaarheid. Het kan gaan om variabelen als geslacht, leeftijd, IQ, werkervaring, etc. Wanneer aan die criteria wordt voldaan en beide condities soortgelijke personen weten aan te trekken (toch op die variabelen die we hebben kunnen nagaan), dan kan je aannemen dat dit ontwerp robuust is voor buitenexperimentele gebeurtenissen, maturatie, testeffect en instrumentatie. Als die bedreigingen even groot zijn voor de experimentele als voor de vergelijkingsconditie, dan worden ze immers uitgezuiverd als het verschil tussen na- en voormeting in de vergelijkingsgroep wordt afgetrokken van het verschil tussen na- en voormeting in de experimentele groep: ($O_2 – O_1$) – ($O_4 – O_3$). Alleen het effect van de experimentele stimulus blijft over.

Grafiek 8.2. Weergave interactie-effect selectie en maturatie.



Het grootste gevaar met het niet-equivalent ontwerp met vergelijkingsgroep komt echter van een mogelijk interactie-effect van selectie met maturatie, buitenexperimentele gebeurtenissen, testeffect, etc. enerzijds en statistische regressie anderzijds. Dit zijn de grootste bedreigingen van de interne geldigheid. Ten eerste, kunnen de groepen doordat ze verschillen in aanvangsscores op de afhankelijke variabele, op een verschillende wijze spontaan evolueren. Een voorbeeld van het interactie-effect van selectie en maturatie maakt één en ander duidelijk (zie Grafiek 8.2). Stel dat je de scores van twee klassen vergelijkt op een toets rekenvaardigheid. De ene klas scoort merkelijk hoger dan de andere in de voormeting: $O_1 > O_3$. De klas die goed scoort in de voormeting, volgt een bijkomende cursus rekenvaardigheid na de gewone lesuren. Het initiële verschil tussen beide klassen wordt nog groter in de nameting ($O_4 – O_2 > O_3 – O_1$), hoewel ook de vergelijkingsgroep in positieve zin geëvolueerd is ($O_4 > O_3$). Kan je hieruit besluiten dat de cursus rekenvaardigheid effectief is geweest? Misschien ontwikkelden de klassen zich spontaan op een verschillend tempo—getuige hiervan misschien de verschillende score in de voormeting. De conclusie dat de bijkomende cursus verantwoordelijk is voor het verschil verliest dus duidelijk aan geldigheid. Een alternatieve verklaring, zoals een verschillend tempo van spontaan veranderen, kan immers niet zomaar van de hand gedaan worden.

Ten tweede, statistische regressie komt geregeld voor in dit ontwerp als de groepen proefpersonen zijn geselecteerd op extreme scores op de afhankelijke variabele of op kenmerken die sterk met die afhankelijke variabele samenhangen. Het verschil in de ontwikkeling van scores tussen voor- en nameting in de experimentele en vergelijkingsgroep is dan niet alleen toe te schrijven aan de experimentele stimulus maar ook aan de gevolgen van statistische regressie. Stel dat je een experimenteel onderzoek wil voeren naar het effect van een bepaalde psychotherapeutische behandeling op iemands ontevredenheid met haar/zijn persoonlijkheid. Ontevredenheid met persoonlijkheid is de afhankelijke variabele. De experimentele groep wordt samengesteld op basis van personen die een expliciete aanvraag hebben ingediend voor die behandeling—en die dus hoogstwaarschijnlijk extreem laag scoren op de afhankelijke variabele. Voor de vergelijkingsgroep kies je ‘normale’ personen, die geen aanvraag tot behandeling hebben ingediend en die tevreden zijn met hun persoonlijkheid. De experimentele groep zal op de nameting sowieso in de richting van het gemiddelde evolueren en hoger scoren, mogelijk niet alleen vanwege de therapie maar omdat statistische regressie optreedt. Zo’n regressie

naar het gemiddelde zal voor de vergelijkingsgroep niet zo sterk zijn of zelfs afwezig. Op die manier zou je de valse indruk kunnen krijgen dat de therapie effectief is. Als de condities zijn samengesteld op basis van zo'n vorm van 'zelfselectie', zijn beide condities niet in dezelfde mate gevoelig voor statistische regressie en verliezen causale redeneringen aan geldigheid.

5.1.3. Afzonderlijke-steekproef ontwerp met voor- en nameting

Dit ontwerp ('separate-sample pretest-posttest design' in het Engels) is uitermate geschikt om effecten van grootschalige publiciteits- en mediacampagnes in te schatten. Bij zulke campagnes ben je immers niet in staat om een deel van de populatie als vergelijkingsgroep te laten fungeren, omdat je er eenvoudigweg niet voor kan zorgen dat iemand de experimentele stimulus niet ontvangt. De volledige bevolking is er namelijk de doelgroep van—vergelijk dit geval met het meten van effecten van mediaboodschappen in *News that matters* (1987), waar proefpersonen in de strikt gecontroleerde omgeving van een laboratorium op de campus van Yale University werden gelokt en daar speciaal aangepaste nieuwsuitzendingen te zien kregen.

Figuur 8.7. Afzonderlijke-steekproef ontwerp met voor- en nameting.

R	O_1	(X)	
R		X	O_2

Dit ontwerp vertrekt van twee op toevalsbasis samengestelde, equivalente subgroepen (zie Figuur 8.7). Eén van die groepen gebruik je als vergelijkingsgroep waar een voormeting plaatsheeft (O_1). De andere groep is de experimentele groep die na de stimulus X gemeten wordt (O_2). Het effect van de experimentele stimulus wordt berekend via $O_2 – O_1$. Hoewel dit design weinig garanties op interne geldigheid biedt, is het vaak de enige mogelijkheid om effecten na te gaan en hierin zit precies zijn waarde. Het afzonderlijke-steekproef ontwerp met voor- en nameting is uitermate gevoelig voor buitenexperimentele gebeurtenissen, maturatie en uitval. Alle kunnen ze zorgen dat de verschillen tussen voor- en nameting niet terug te voeren zijn tot de experimentele stimulus. De bedreiging van interne geldigheid wordt groter naarmate de periode tussen O_1 en O_2 toeneemt.

5.2. Pre-experimentele ontwerpen

5.2.1. Groep met alleen nameting

Onderzoek binnen de sociale wetenschappen maakt gebruik van dit ontwerp om effecten te analyseren, hoewel er geen enkele methodologische garantie is om causale redeneringen te rechtvaardigen. Meestal gaat het om tentatieve inschatting van mogelijke effecten van beleidsmaatregelen op de waargenomen resultaten. Dit ontwerp met alleen een nameting ('one-shot case study' in het Engels) is eigenlijk het minimale referentiepunt voor experimentele ontwerpen (Figuur 8.8). De basisonderdelen van een experimenteel ontwerp zijn immers afwezig: er is geen vergelijking mogelijk met een conditie waarin de stimulus niet heeft plaatsgehad en er worden geen verschillen tussen een huidige en een vroegere situatie gemeten in de afhankelijke variabele. Grond voor causale inferentie is er dus niet.

Figuur 8.8. Schematische voorstelling ontwerp met alleen nameting.

$$X \qquad O_1$$

5.2.2. Groep met voor- en nameting

De groep met voor- en nameting ('one-group pretest-posttest design' in het Engels) is al een stap vooruit op het vorige ontwerp, maar er zijn niettemin heel wat factoren die de interne geldigheid van dit design ondermijnen. Er wordt in dit ontwerp slechts gewerkt met één conditie, met name de experimentele groep, waarbij een voormeting plaatsvindt, de experimentele stimulus wordt toegediend, en een nameting gebeurt (zie Figuur 8.9). Vergelijking met een conditie die geen stimulus heeft gekregen is onmogelijk. Vandaar dat er heel wat variabelen zijn buiten de stimulus X die een alternatieve verklaring kunnen bieden voor het gevonden verschil tussen O_2 en O_1: buitenexperimentele gebeurtenissen, maturatie, testeffect, instrumentatie, statistische regressie.

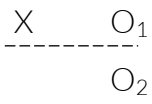
Figuur 8.9. Schematische voorstelling ontwerp met voor- en nameting.

$$O_1 \quad X \quad O_2$$

5.2.3. Bestaande groepen met alleen nameting

In dit ontwerp ('static-group comparison design' in het Engels) is er sprake van een experimentele en vergelijkingsgroep, maar een voormeting en randomisering ontbreken allebei (zie Figuur 8.10)—eigenlijk is dit ontwerp nauw verwant met het ontwerp met een vergelijkingsgroep en alleen nameting (vergelijk met Figuur 8.2). Eén conditie heeft de experimentele stimulus gekregen, de andere niet en het verschil tussen O_2 en O_1 wordt toegeschreven aan het effect van X. Er is echter geen enkele formele grond op basis waarvan de verschillen tussen O en O aan stimulus X kunnen toegewezen worden. Misschien verschillen de scores tussen de condities op de afhankelijke variabele al vóórdat stimulus X is toegediend en is er sprake van selectie of differentiële rekrutering. Of uitval kan ervoor zorgen dat personen met bepaalde kenmerken die samenhangen met de afhankelijke variabele makkelijker uitvallen in de experimentele groep dan in de vergelijkingsgroep—dit is selectieve drop-out.

Figuur 8.10. Schematische voorstelling ontwerp bestaande groepen met alleen nameting.



In Tabel 8.1 vind je een samenvatting van alle besproken experimentele ontwerpen (zuivere, quasi- en pre-experimentele designs), alsook de belangrijkste bedreigingen voor interne geldigheid. Een minteken betekent dat het ontwerp weinig garantie biedt, een plusteken indiceert dat het ontwerp controleert voor die bedreiging, een vraagteken staat voor mogelijke bron van bedreiging. Wanneer er geen symbool wordt toegekend, betekent het dat die bedreiging niet relevant is.

Tabel 8.1. Bronnen voor bedreiging van interne geldigheid bij verschillende experimentele ontwerpen.

	Buitenexperimentele gebeurtenissen	Maturatie	Testeffect	Instrumentatie	Statistische regressie	Selectie	Uitval	Interactie selectie en maturatie, testeffect, etc.
Alleen nameting	-	-	...	...	...	-	-	...
Voor- en nameting	-	-	-	-	?	+	+	-
Bestaande groepen met alleen nameting	+	?	+	+	+	-	-	-
Tijdreeksontwerp	-	+	+	?	+	+	+	+
Niet-equivalent ontwerp met vergelijkingsgroep	+	+	+	+	?	+	+	-
Afzonderlijke-steekproef ontwerp met voor- en nameting	-	-	+	?	+	+	-	-
Voormeting-nameting ontwerp met vergelijkingsgroep	+	+	+	+	+	+	+	+
Ontwerp met vergelijkingsgroep en alleen een nameting	+	+	+	+	+	+	+	+
Solomon vier-groepen ontwerp	+	+	+	+	+	+	+	+

Legende: '+' betekent dat factor onder controle is, '-' wijst op duidelijke zwakte in design, '?' duidt op een mogelijke bron voor problemen, '...' wil zeggen dat factor niet relevant is.

6. Ethische bekommernissen

In experimenten zijn ethische kwesties vaak aan de orde. Het zijn onderzoeksontwerpen die vaak veel vergen van de proefpersonen en hen confronteren met gevoelens of gedrag die in normale sociale situaties misschien zelden of nooit voorkomen. Vaak komt ook manipulatie van die gevoelens en handelingen voor—denk maar aan de experimenten van Asch en Milgram. In sommige gevallen is een lichte misleiding nodig over de ware toedracht van het onderzoek. Dit misleiden van proefpersonen mag alleen in overweging genomen worden wanneer op een andere manier het onderzoek niet zou kunnen uitgevoerd worden. Uiteraard worden achteraf de proefpersonen gede-brieft en wordt hun de ware toedracht van het experiment meegedeeld. Het spreekt ook vanzelf dat proefpersonen nooit in fysiek gevaarlijke situaties mogen gebracht worden.

Foto's van subject in experiment van Asch.

Vandaar dat voor het uitvoeren van experimenten met mensen bepaalde richtlijnen en regels gelden die gegrondvest zijn in een drietal ruime ethische principes. Een eerste ethisch principe dat relevant is in het kader van experimenteel onderzoek is respect voor personen. Individuen moeten beschouwd en behandeld worden als autonome actoren en mensen met een verminderde autonomie moeten worden beschermd—bijvoorbeeld medische experimenten in ontwikkelingslanden. Dit principe mondt uit of geeft aanleiding tot het principe van 'informed consent'. Deelname moet vrijwillig zijn en proefpersonen moeten voldoende geïnformeerd worden over de mogelijke gevolgen van hun deelname. Een tweede moreel handvest is het principe van 'beneficence': vermijd het toebrengen van schade en probeer tegelijk de voordelen te maximaliseren en de nadelen te minimaliseren. Een protocol van het experimentele ontwerp is hierin belangrijk. Zo'n protocol bevat onder andere informatie over de financiële sponsors van het onderzoek, de institutionele inbedding of affiliatie, het verloop van de studie, etc. en wordt voorgelegd aan een ethische commissie ter goed- of afkeuring. Het derde morele principe is het idee van rechtvaardigheid, met name een afweging tussen zij die de voordelen van het onderzoek ontvangen en diegenen die er de lasten van dragen—rijk versus arm, Noord versus Zuid. Een faire selectie van proefpersonen moet dit principe verzekeren.

[1] Voor een historisch overzicht van het experiment, zie bijvoorbeeld Morawski (1988) en Webster & Sell (2007).

[2] Die ideeën van Campbell en Stanley in de jaren ’60 zijn uiteraard niet volledig nieuw—al van in de jaren ’30 wordt gestreefd om wetenschappelijke kennis te gebruiken om beleid en bedrijfsleven te informeren, zie bijvoorbeeld Stephan (1935) of Roethlisberger & Dickson (1939).

[3] We gaan ervan uit dat de kenmerken op het metrische niveau gemeten zijn, zodat we met geobserveerde gemiddeldes kunnen werken in de voor- en nametingen.

[4] We volgen hier de opsomming van Campbell & Stanley (1963). Voor wie op zoek is naar meer materiaal over bedreigingen van interne geldigheid, verwijzen we door naar Shadish, Cook & Campbell (2002) of Cook & Campbell (1979).

[5] Voor verdere discussie over externe geldigheid, zie Mook (1983), Vissers, Heyne, Peters & Geurts (2001) en Zelditch (2007).