Methoden en Statistiek 1

true

3 juli 2020

Contents

Voorwoord

In dit boek hebben we gepoogd om de belangrijkste kwantitatieve methoden en statistische technieken die relevant zijn voor de Geesteswetenschappen uit te leggen. De tekst is waar mogelijk gevrijwaard van wiskundige afleidingen en formules, omdat we die voor studenten Geesteswetenschappen minder bruikbaar achten. Indien nodig geven we de belangrijkste formules voor statistische beschrijvingen en analyses in een aparte paragraaf.

Het tekstboek bevat ook aanwijzingen over hoe de besproken statistische analyses en visualisaties uitgevoerd kunnen worden in twee veelgebruikte programma's, nl. SPSS (versie 22 en later) en R (versie 3.0 en later). Ook deze aanwijzingen staan los van de hoofdtekst, in afzonderlijke paragrafen.

Graag willen we onze mede-docenten danken voor de vele discussies en voorbeelden die op enige wijze verwerkt zijn in dit tekstboek. Onze studenten danken we voor hun nieuwsgierigheid en nauwkeurigheid die geleid heeft tot deze versie van dit tekstboek.

Ook betonen wij grote dank aan Willemijn Heeren, Gerrit Bloothooft, Marijn Struiksma, Margot van den Berg, Els Rose, Tobias Quené en Kirsten Schutter voor hun adviezen, data, en/of commentaar bij het manuscript van dit digitale tekstboek.

Utrecht, december 2016 - juli 2020

Hugo Quené, https://www.hugoquene.nl

Huub van den Bergh

Notatie

In aansluiting op het internationale gebruik en op de conventies van Engelstalige tijdschriften gebruiken we de punt als decimaalteken; we schrijven dus $\frac{3}{2} = 1.5$. Hierbij is een waarschuwing op zijn plaats: het decimale symbool kan

verschillen tussen computers, en zelfs tussen programma's op dezelfde computer. Controleer dus welk decimaal symbool gebruikt wordt door (elk programma op) jouw computer.

Licentie

This document is licensed under the *GNU GPL 3* license (for details see https://www.gnu.org/licenses/gpl-3.0.en.html). It was created with the `bookdown` package (Xie, 2020) in Rstudio.

Deel I: Methodologie

Chapter 1

Inleiding

In dit tekstboek worden de grondbeginselen, methoden en technieken van empirisch wetenschappelijk onderzoek besproken, zowel in algemene zin als toegespitst op het brede domein van taal en communicatie. We zullen ons bezighouden met vragen als: Wat is een goede onderzoeksvraag? Welke methode is de beste om de onderzoeksvraag te beantwoorden? Hoe kunnen onderzoekers zinnige en valide conclusies trekken uit (statistische analyses van) hun gegevens? In dit tekstboek beperken we ons tot de belangrijkste grondbeginselen, en tot de belangrijkste methoden en technieken. In dit eerste hoofdstuk zullen we een overzicht geven van verschillende typen en vormen van wetenschappelijk onderzoek. In het vervolg van dit tekstboek geven we de meeste aandacht aan methoden van wetenschappelijk onderzoek waarbij empirische observaties uitgedrukt worden in de vorm van getallen (kwantitatief), die geanalyseerd worden met behulp van statistische technieken.

1.1 Wetenschappelijk onderzoek

Om te beginnen moeten we een vraag stellen die terugslaat op de allereerste zin hierboven: wat is eigenlijk wetenschappelijk onderzoek? Wat is het verschil tussen wetenschappelijk en niet-wetenschappelijk onderzoek (bijv. door onderzoeksjournalisten)? Onderzoek dat een wetenschapper uitvoert, hoeft nog geen wetenschappelijk onderzoek te zijn. Evenmin is journalistiek onderzoek per definitie onwetenschappelijk omdat het door een journalist wordt uitgevoerd. In dit tekstboek hanteren we de volgende definitie (Kerlinger and Lee, 2000, p.14):

> "Scientific research is systematic, controlled, empirical, amoral, public, and critical investigation of natural phenomena. It is guided by theory and hypotheses about the presumed relations among such phenomena."

Wetenschappelijk onderzoek is systematisch en gecontroleerd. Wetenschappelijk onderzoek is zodanig ontworpen dat we geloof kunnen hechten aan de conclusies, omdat die conclusies goed onderbouwd zijn. Het onderzoek kan door anderen herhaald worden, met (hopelijk) dezelfde resultaten. Deze eis van repliceerbaarheid maakt ook dat wetenschappelijk onderzoek zeer nauwgezet wordt ontworpen en uitgevoerd (zie Hoofdstukken 3 en **??**). De sterkste vorm van controle is die van een wetenschappelijk experiment; we besteden daarom in dit tekstboek veel aandacht aan experimenteel onderzoek (§1.5). Mogelijke alternatieve verklaringen voor het onderzochte verschijnsel worden één voor één onderzocht en zo mogelijk uitgesloten, zodat tenslotte slechts één verklaring overblijft (Kerlinger and Lee, 2000). Die verklaring vormt dan onze wetenschappelijk onderbouwde conclusie of theorie over het onderzochte verschijnsel.

Ook wordt in de definitie gesteld dat wetenschappelijk onderzoek *empirisch* van aard is. De conclusies die de onderzoeker trekt moeten uiteindelijk gebaseerd zijn op (systematische en gecontroleerde) waarnemingen of observaties van een verschijnsel in de werkelijkheid — bijvoorbeeld op de waargenomen inhoud van een tekst, of op het waargenomen gedrag van een proefpersoon. Als die waarneming ontbreekt, dan kunnen de eventuele conclusies niet logisch verbonden worden met de werkelijkheid, waardoor ze geen wetenschappelijke waarde hebben. Vertrouwelijke gegevens uit een onbekende bron, of inzichten verkregen in een droom of in een mystieke beleving, zijn niet empirisch onderbouwd, en kunnen dus niet de basis vormen van een wetenschappelijke theorie.

1.1.1 Theorie

Het doel van wetenschappelijk onderzoek is te komen tot een theorie over een deel van de werkelijkheid. Die theorie is te zien als een coherente en consistente verzameling van "justified true beliefs" (Morton, 2003). In deze overtuigingen, en in de theorie, wordt geabstraheerd van de complexe werkelijkheid van de natuurlijke verschijnselen, naar een abstract mentaal *construct*, dat uit zijn aard niet rechtstreeks waarneembaar is. Voorbeelden van dergelijke constructen zijn: leesvaardigheid, intelligentie, activatie-niveau, verstaanbaarheid, omvang van iemands actieve woordenschat, schoenmaat, woon-werk-afstand, introvertheid, etc.

Een onderzoeker definieert in een theorie niet alleen verschillende constructen, maar ook specificeert hij de *verbanden* of relaties tussen deze constructen. Pas wanneer zowel de constructen gedefinieerd zijn als de relaties tussen de constructen gespecificeerd zijn, kan een onderzoeker komen tot een systematische verklaring van het onderzochte verschijnsel. Deze verklaring of theorie kan weer de basis zijn van een *voorspelling* over het onderzochte verschijnsel: het aantal gesproken talen op de wereld zal verminderen in de 21e eeuw; teksten zonder voegwoorden zullen moeilijker te begrijpen zijn dan teksten met voegwoorden; kinderen die tweetalig opgroeien zullen niet slechter presteren op school dan eentalige kinderen.

Wetenschappelijk onderzoek is er in vele verschillende typen en vormen, die op verschillende manieren ingedeeld kunnen worden. In de volgende sectie 1.2 bespreken we een indeling op basis van paradigma, de manier waarop de onderzoeker tegen de werkelijkheid aankijkt. Onderzoek kan ook ingedeeld worden op een continuüm van 'zuiver theoretisch' naar 'toegepast'. Een derde manier om onderzoek in te delen is gericht op het type onderzoek, bijvoorbeeld instrumentatieonderzoek (§1.3), beschrijvend onderzoek (§1.4), en experimenteel onderzoek (§1.5).

1.2 Paradigmata

Eén criterium om typen onderzoek te onderscheiden is op basis van het gebruikte paradigma, de manier waarop de onderzoeker tegen de werkelijkheid aankijkt. In dit tekstboek besteden we nagenoeg alleen aandacht aan het empirisch-analytisch paradigma, omdat dit het meest uitgewerkte en meest invloedrijke paradigma is. Heden ten dage kan deze benadering opgevat worden als 'de' standaardopvatting, waar andere paradigma's zich min of meer tegen afzetten.

Binnen het *empirisch-analytische paradigma* onderscheiden we twee varianten: het positivisme en het kritisch-rationalisme. Beide stromingen hebben gemeen dat er aangenomen wordt dat er wetmatigheden zijn die 'ontdekt' kunnen worden: verschijnselen kunnen beschreven en verklaard worden in abstracte termen (constructen). Het verschil tussen beide stromingen binnen de empirisch-analytische traditie is gelegen in de pretentie van de uitspraken die gedaan worden. Volgens de positivisten is het mogelijk om uitspraken te doen vanuit feitelijke waarnemingen naar een theorie. Op basis van de observaties kunnen we generaliseren naar een algemeen geldende regel, door middel van inductie. (De vogels die ik zie, die hoor ik ook fluiten, dus alle vogels fluiten.)

De tweede stroming is het kritisch-rationalisme. De aanhangers van deze stroming keren zich tegen bovengenoemde inducties: al hoor ik talloze vogels ook fluiten, dan nog kan ik geen zekerheid verkrijgen over de veronderstelde algemene regel. Maar we kunnen het wel omkeren, en proberen aan te tonen dat de veronderstelde algemene regel of hypothese *niet* juist is. Hoe werkt dat? Op basis van de algemeen geldende regel kunnen we voorspellingen afleiden voor specifieke observaties, door middel van deductie. (Als alle vogels fluiten, dan moet het zo zijn dat alle vogels in mijn steekproef fluiten.) Als niet alle vogels in mijn steekproef fluiten, dan is de algemene regel blijkbaar onjuist. Dit wordt het falsificatie-principe genoemd; we bespreken dat uitgebreider in sectie 2.4.

Ook aan het kritisch-rationalisme kleven echter tenminste twee bezwaren. Met het falsificatieprincipe kunnen waarnemingen (empirische feiten, observaties, onderzoeksresultaten) gebruikt worden om theoretische uitspraken te doen (met betrekking tot hypothesen). Strikt genomen moet een veronderstelde algemene regel meteen verworpen worden na één geslaagde falsificatie (een van de vogels in mijn steekproef fluit niet): als theorie en observatie niet overeenstemmen,

dan faalt de theorie, volgens de kritisch-rationalisten. Maar om te komen tot een observatie moet een onderzoeker vele keuzes maken (bijv.: hoe maak ik een goede steekproef, wat is een vogel, hoe bepaal ik of een vogel fluit?), die de geldigheid van de observaties onzeker kunnen maken. Er kan dus ook iets mis zijn met de waarnemingen zelf (horen), of met de operationalisaties van de gebruikte constructen (vogels, fluiten).

Een tweede probleem is dat er in de praktijk eigenlijk zeer weinig theorieën zijn die werkelijk iets uitsluiten. Wanneer er discrepanties waargenomen worden tussen theorie en observaties, dan wordt de theorie bijgeschaafd, zodat de nieuwe observaties toch weer binnen de theorie passen. Theorieën worden dan ook zelden volledig verworpen.

Een tweede paradigma is de kritische benadering. Het *kritische paradigma* onderscheidt zich van andere paradigmata in de nadruk op maatschappelijke bepaaldheden; 'de' werkelijkheid bestaat niet, ons beeld ervan is een voorlopige, door maatschappelijke oorzaken bepaalde werkelijkheid. Inzicht in de maatschappelijke verhoudingen heeft zelf dus ook invloed op die werkelijkheid. Onze wetenschapsopvatting zoals verwoord in bovengenoemde definities van onderzoek en theorie wordt in het kritische paradigma dan ook afgewezen. Kritische onderzoekers menen dat onderzoeksprocessen niet los gezien kunnen worden van de maatschappelijke context waarin het onderzoek is verricht. Deze laatste visie wordt overigens overgenomen door steeds meer onderzoekers, ook door hen die andere paradigmata aanhangen.

1.3 Instrumentatie-onderzoek

Onderzoek is, zoals gezegd, een gesystematiseerde en gecontroleerde wijze om empirische gegevens te verzamelen en te interpreteren. Onderzoekers streven naar inzicht in natuurlijke verschijnselen, en in de wijze waarop (de constructen van) die verschijnselen met elkaar samenhangen. Een voorwaarde hiervoor is dat de onderzoeker deze verschijnselen daadwerkelijk kan meten, d.i. uitdrukken in een observatie (bij voorkeur in de vorm van een getal). Instrumentatieonderzoek is voornamelijk gericht op de constructie van instrumenten of methoden om verschijnselen, gedrag, vaardigheden, attitudes, etc. meetbaar te maken. De ontwikkeling van goede meetinstrumenten is bepaald geen sinecure: het is ambachtelijk handwerk, waarbij de constructeur vele valkuilen moet zien te vermijden. Het meetbaar maken van verschijnselen, van gedrag of van constructen noemen we de *operationalisatie*. Een concrete leestoets is bijvoorbeeld op te vatten als een operationalisatie van het abstracte construct 'leesvaardigheid'.

We kunnen een nuttig onderscheid maken tussen het abstracte theoretische construct en het gemeten construct, ofwel een onderscheid tussen: het begrip-zoals-bedoeld en het begrip-zoals-bepaald. Het is uiteraard de bedoeling dat het begrip-zoals-bepaald (de toets, de vragenlijst, de observatie) het begrip-zoals-bedoeld (het theoretische construct) zo goed mogelijk benadert. Indien het

theoretische construct goed wordt benaderd, dan spreken we van een adequate of valide meting.

Bij operationalisatie van een begrip-zoals-bedoeld moeten talloze keuzen gemaakt worden. Zo moet het CITO (Centraal instituut voor toetsontwikkeling) elk jaar tekstbegripstoetsen construeren om de leesvaardigheid van eindexamenkandidaten te meten. Daarvoor moet allereerst een tekst gekozen of geredigeerd worden. Deze tekst mag niet te moeilijk, maar ook niet te makkelijk zijn voor de doelgroep. Voorts mag het onderwerp van de tekst niet al te bekend zijn, omdat anders de bij sommige leerlingen aanwezige algemene kennis kan interfereren met de meningen en standpunten die in de tekst naar voren gebracht worden. Vervolgens moeten de vragen zó ontworpen worden dat de verschillende passages in de tekst aan bod komen. Ook moeten de vragen zó samengesteld zijn dat het theoretische construct 'leesvaardigheid' adequaat geoperationaliseerd wordt. Tot slot moet ook nog rekening gehouden worden met de examens uit voorafgaande jaren; het nieuwe examen mag immers niet al te veel afwijken van oude examens.

Een construct moet dus op de juiste wijze geoperationaliseerd zijn, om observaties te verkrijgen die niet alleen valide zijn (een goede benadering van het abstracte construct, zie Hoofdstuk 5) maar die ook betrouwbaar zijn (ongeveer gelijke observaties bij herhaalde meting, zie Hoofdstuk **??**). In ieder onderzoek zijn de validiteit en de betrouwbaarheid van een meting cruciaal; we besteden dan ook twee hoofdstukken aan deze begrippen. Maar in instrumentatieonderzoek zijn deze begrippen zelfs essentieel, omdat dit type onderzoek juist beoogt om valide en betrouwbare instrumenten te leveren, die een goede operationatisatie zijn van het abstracte construct-zoals-bedoeld.

1.4 Beschrijvend onderzoek

Met beschrijvend onderzoek bedoelen we onderzoek dat voornamelijk gericht is op de beschrijving van een bepaald natuurlijk verschijnsel in de werkelijkheid. De onderzoeker richt zich dus vooral op een *beschrijving* van het verschijnsel: het huidige vaardigheidsniveau, het verloop van een proces of een discussie, de wijze waarop de lessen Nederlands in het voortgezet onderwijs vorm worden gegeven, de politieke voorkeur van stemmers vlak voor verkiezingen, de samenhang tussen het aantal uren zelfstudie en het eindcijfer dat een student behaalt, etc. Kortom, ook de onderwerpen voor beschrijvend onderzoek kunnen zeer divers zijn.

Voorbeeld 1.1: (Dingemanse et al., 2013) hebben opnames van conversaties gekozen of gemaakt in 10 talen. Uit die opgenomen conversaties zijn woorden genomen waarmee een luisteraar om "open verduidelijking" vraagt: woordjes als *hè* (Nederlands), *huh* (Engels), *ã?*

> (Siwu). Van deze woorden is de klankvorm en het toonhoogteverloop vastgesteld, met akoestische metingen en met fonetische transcripties door experts. Een conclusie van dit beschrijvende onderzoek luidt dat deze tussenvoegsels in de verschillende talen veel meer op elkaar lijken (in klankvorm en toonhoogteverloop) dan op grond van toeval te verwachten is.

Dit voorbeeld illustreert dat beschrijvend onderzoek niet ophoudt als de gegevens (klankvormen, toonhoogteverloop) beschreven zijn. Vaak zijn *verbanden* tussen de verzamelde gegevens ook zeer interessant (zie §1.1). Zo wordt in opiniepeilingen naar het stemgedrag bij verkiezingen vaak een verband gelegd tussen het gepeilde stemgedrag enerzijds, en leeftijd, geslacht en opleidingsniveau van de respondent anderzijds. En evenzo wordt in onderwijskundig onderzoek een verband gelegd tussen aantal uren studietijd enerzijds, en studiesucces van de respondent anderzijds. Dit type van beschrijvend onderzoek, waarbij een correlatie wordt vastgesteld tussen mogelijke oorzaken en mogelijke gevolgen, wordt ook aangeduid als *correlationeel onderzoek*.

Het essentiële verschil tussen beschrijvend en experimenteel onderzoek is gelegen in de vraag naar oorzaak en gevolg. Op basis van beschrijvend onderzoek kan een causaal verband tussen oorzaak en gevolg *niet* goed vastgesteld worden. Uit beschrijvend onderzoek zou kunnen blijken dat er een samenhang is tussen een bepaald soort voeding en een langere levensduur. Is het voedingspatroon dan ook de oorzaak van de langere levensduur? Dat hoeft bepaald niet het geval te zijn: het is ook mogelijk dat dat soort voeding vooral genuttigd wordt door mensen die relatief hoog opgeleid en welvarend zijn, en door deze *andere* factoren ook relatief langer in leven zijn[1]. Om vast te kunnen stellen of er een causaal verband is, moeten we experimenteel onderzoek opzetten en uitvoeren.

1.5 Experimenteel onderzoek

Experimenteel onderzoek wordt gekenmerkt doordat de onderzoeker een bepaald aspect van de onderzoeksomstandigheden systematisch varieert (Shadish et al., 2002). Het effect van deze manipulatie staat dan centraal in het onderzoek. Een onderzoeker vermoedt bijvoorbeeld dat een bepaalde nieuwe lesmethode zal resulteren in betere prestatie van de leerlingen dan de huidige lesmethode. De onderzoeker wil deze hypothese toetsen door middel van een experimenteel onderzoek. Hij manipuleert het type onderwijs: sommige klassen of groepen krijgen les volgens de nieuwe experimentele lesmethode en andere klassen of

[1] Het is zelfs mogelijk dat het onderzochte voedingspatroon de oorzaak is van een relatief *kortere* levensduur, maar dat dit negatieve effect gemaskeerd wordt door de sterkere, positieve effecten van opleidingsniveau en welvaartsniveau op de levensduur.

groepen krijgen les op de traditionele wijze. Het effect van de nieuwe lesmethode wordt geëvalueerd door de prestaties van de twee soorten schoolklassen te vergelijken, na de 'behandeling' met de oude vs. nieuwe lesmethode.

Experimenteel onderzoek heeft als voordeel dat we de onderzoeksresultaten doorgaans mogen interpreteren als het gevolg van de experimentele manipulatie. Omdat de onderzoeker het onderzoek systematisch controleert en slechts één aspect (i.c. de lesmethode) varieert, kunnen eventuele verschillen tussen de prestaties van de twee categorieën alleen toegeschreven worden aan het veranderde kenmerk, i.c. aan de lesmethode. Dit veranderde kenmerk moet dan logischerwijs wel de oorzaak zijn van de geobserveerde verschillen. Experimenteel onderzoek is dus gericht op de evaluatie van causale verbanden.

Deze redenering vereist wel dat proefpersonen (of schoolklassen, in bovenstaand voorbeeld) volgens het toeval, aselect (Eng. 'at random'), worden toegewezen aan de experimentele condities (i.c. de oude of nieuwe lesmethode). Deze aselecte toewijzing (Eng. 'random assignment') is de beste methode om eventuele niet-relevante verschillen tussen de behandelcondities uit te sluiten. Een dergelijk experiment met aselecte toewijzing van proefpersonen aan condities wordt een *gerandomiseerd experiment* genoemd (Eng. 'randomized experiment', 'true experiment',). Om bij ons voorbeeld te blijven: als de onderzoeker de oude lesmethode zou inzetten bij jongens, en de nieuwe lesmethode bij meisjes, dan is een eventueel verschil in prestaties niet meer uitsluitend toe te schrijven aan het gemanipuleerde kenmerk (de lesmethode), maar ook aan een niet-gemanipuleerd maar wel relevant kenmerk, hier het geslacht van de leerlingen. Zo'n mogelijk verstorend kenmerk wordt een *storende variabele* (Eng. 'confound') genoemd. In Hoofdstuk **??** bespreken we hoe we deze storende variabelen kunnen neutraliseren, door random toewijzing van proefpersonen (of schoolklassen) aan de experimentele condities, in combinatie met andere maatregelen.

Er is ook experimenteel onderzoek waarbij een bepaald aspect (zoals lesmethode) wel systematisch varieert, maar waarbij de proefpersonen of schoolklassen *niet aselect* zijn toegewezen aan de experimentele condities; dit wordt *quasi-experimenteel* onderzoek genoemd (Shadish et al., 2002). In het bovenstaande voorbeeld is daarvan sprake als de gebruikte lesmethode onderzocht wordt, met gegevens van schoolklassen waarvan niet de onderzoeker maar de docent bepaald heeft of de oude of nieuwe lesmethode gebruikt wordt. Als de nieuwe lesmethode betere prestaties zou opleveren, dan weten we *niet* met zekerheid dat het verschil in prestaties toe te schrijven is aan de lesmethode. Ook het enthousiasme of de werkstijl van de docent kan een storende variabele zijn geweest in dit quasi-experiment. In dit tekstboek zullen we verschillende voorbeelden van quasi-experimenteel onderzoek tegenkomen.

Binnen het type van experimenteel onderzoek kunnen we een verdere verdeling aanbrengen, tussen laboratoriumonderzoek en veldonderzoek. In beide typen experimenteel onderzoek wordt een aspect van de werkelijkheid gemanipuleerd. Het verschil tussen beide typen onderzoek is gelegen in de mate waarin de

onderzoeker in staat is om allerlei storende aspecten van de werkelijkheid onder controle te houden. In laboratoriumonderzoek kan de onderzoeker zeer exact bepalen onder welke omgevingscondities de observaties worden gedaan, en kan de onderzoeker dus ook vele mogelijke storende variabelen onder controle houden (denk aan verlichting, temperatuur, omgevingslawaai, etc.). In veldonderzoek is dit niet het geval. De onderzoeker is 'in het vrije veld' niet in staat om alle (mogelijk relevante) aspecten van de werkelijkheid volledig onder controle te houden.

Voorbeeld 1.2: Margot van den Berg onderzocht samen met collega's van de Universiteit van Ghana en de Universiteit van Lomé hoe meertalige sprekers hun talen gebruiken als zij eigenschappen zoals kleur, grootte en waarde moeten benoemen door middel van een zogenaamde *Director-Matcher task* (Van den Berg et al., 2017). In deze taak gaf de ene onderzoeksdeelnemer (de directeur) aanwijzingen aan een ander (de uitvoerder) om een reeks voorwerpen in een bepaalde volgorde neer te zetten. Zo konden in een kort tijdsbestek veel voorkomens van eigenschapswoorden worden verzameld ('Zet de gele auto naast de rode auto maar boven de kleine slipper'). De gesprekken werden opgenomen, uitgeschreven en vervolgens onderzocht op taalkeuze, moment van taalwisseling en type grammaticale constructie. Bij dergelijk veldwerk kunnen echter allerlei niet-gecontroleerde aspecten in de omgeving van invloed zijn op de geluidsopnames, en daarmee op de gegevens: "kakelende kippen, een buurman die z'n motor aan het repareren is en 'm om de haverklap moet starten terwijl je een gesprek aan het opnemen bent, keiharde regen op het aluminium dak van het gebouw waar de interviews plaats vinden." (Margot van den Berg, pers.comm.)

Voorbeeld 1.3: Bij het luisteren naar gesproken zinnen kunnen we uit de oogbewegingen van een proefpersoon afleiden, hoe die gesproken zinnen worden verwerkt. In een zgn. 'visual world'-taak krijgen luisteraars een zin te horen (bijv. "Bert zegt dat het konijn is gegroeid"), terwijl ze kijken naar meerdere afbeeldingen op het scherm (meestal 4, bijv. een schelp, pauw, zaag, en wortel). Luisteraars blijken vooral te kijken naar de afbeelding die geassocieerd is aan het woord dat ze op dat moment mentaal verwerken: als ze het woord *konijn* verwerken, dan kijken ze naar de wortel (exacter gezegd: ze kijken vaker en langer naar de wortel dan naar de andere afbeeldingen). Met een zgn. 'eye tracker' kan worden vastgesteld naar welke positie van het scherm de proefpersoon kijkt (door observatie van de pupillen).

De onderzoeker kan zo dus observeren welk woord op welk moment mentaal verwerkt wordt (Koring et al., 2012). Dergelijk onderzoek kan het beste uitgevoerd worden in een laboratorium, met controle over achtergrondgeluiden, verlichting, en positie van de ogen t.o.v. computerscherm.

Laboratoriumonderzoek en veldonderzoek hebben beide voordelen en nadelen. Het grote voordeel van laboratoriumonderzoek is natuurlijk de mate waarin de onderzoeker allerlei externe zaken onder controle kan houden. In een laboratorium zal het experiment niet vaak verstoord worden door een startende motor of door een regenbui. Dit voordeel van laboratoriumonderzoek is echter ook een belangrijk nadeel, nl. dat het onderzoek plaatsvindt in een min of meer kunstmatige omgeving. Het is dan nog maar de vraag in hoeverre resultaten die onder kunstmatige omstandigheden verkregen zijn, ook zullen gelden in het leven van alledag buiten het laboratorium. Dit laatste is dan ook een punt in het voordeel van veldonderzoek: het onderzoek wordt verricht onder natuurlijke omstandigheden. Het nadeel van veldonderzoek is dan weer dat er in het veld kan van alles gebeuren wat de onderzoeksresultaten beïnvloedt, maar waar de onderzoeker geen controle over kan houden (zie het bovenstaande voorbeeld). De keuze die een onderzoeker maakt tussen beide typen experimenteel onderzoek wordt uiteraard sterk bepaald door de vraagstelling van het onderzoek. Sommige vraagstellingen laten zich beter in laboratoriumsituaties onderzoeken, terwijl andere beter in veldsituaties onderzocht kunnen worden (zoals bovenstaande voorbeelden illustreren).

1.6 Vooruitblik

Dit tekstboek bestaat uit drie delen. Deel I (hoofdstukken 1 tot en met 7) van dit tekstboek behandelt methoden van onderzoek, en geeft een toelichting bij allerlei termen en begrippen die van belang zijn bij het ontwerpen en opzetten van goed wetenschappelijk onderzoek.

In deel II (hoofdstukken 8 tot en met 12) van het tekstboek behandelen we de beschrijvende statistiek (Eng. ‘descriptive statistics’) en in deel III (hoofdstukken 13 tot en met 17) behandelen we de elementaire technieken uit de toetsende statistiek (Eng. ‘inferential statistics’). Met deze laatste twee delen streven we drie doelen na.

Allereerst willen we dat je in staat bent om artikelen en andere verslagen waarin statistische verwerkings- en toetsingstechnieken zijn gebruikt, kritisch te beoordelen. Ten tweede willen we dat je dat je de noodzakelijke kennis en inzicht hebt in de belangrijkste statistische procedures. Ten derde willen we met deze

statistische delen bereiken dat je in staat bent om zelfstandig statistische bewerkingen uit te voeren voor je eigen onderzoek, bijvoorbeeld voor je stage of eindwerkstuk.

Deze drie doelen zijn geordend in volgorde van belangrijkheid. Wij menen dat een adequate en kritische interpretatie van statistische resultaten en de conclusies die daaraan verbonden kunnen worden van groot belang is voor alle studenten. Om die reden besteden we in dit tekstboek dan ook relatief veel aandacht (in deel I) aan de 'filosofie' of methodologie achter de besproken statistische technieken en analyses. Ook geven we aan hoe je de besproken statistische analyses zelf kunt uitvoeren in SPSS (een populair pakket voor statistische analyses) en in R (een wat moeilijker, maar ook krachtiger en veelzijdiger pakket, met stijgende populariteit). Beide statistische pakketten zijn geïnstalleerd in de computerleerzalen van de Faculteit Geesteswetenschappen. SPSS is beschikbaar via SurfSpot.nl voor een sterk gereduceerde prijs. R is vrijelijk beschikbaar via www.R-project.org. Meer achtergrond over het gebruik van R is te vinden via https://hugoquene.github.io/emlar2020/.

Chapter 2

Hypothese-toetsend onderzoek

2.1 Inleiding

Veel empirisch onderzoek heeft tot doel om verbanden vast te stellen tussen (vermeende) oorzaken en hun (vermeende) gevolgen. De onderzoeker wil weten of de ene variabele van invloed is op de andere. Het onderzoek toetst de hypothese dat er een verband is tussen de vermeende oorzaak en het vermeende gevolg (zie Tabel 2.1). De beste methode om zo'n causaal verband vast te stellen, en dus om de hypothese te toetsen, is het experiment. Een goed opgezet en goed uitgevoerd experiment is de 'gouden standaard' in veel wetenschappelijke disciplines, omdat het goede waarborgen biedt voor de validiteit van de conclusies (zie Hoofdstuk 5. Anders gezegd: de uitkomsten van een goed experiment vormen de sterkst mogelijke evidentie voor een verband tussen de onderzochte variabelen. Zoals besproken in Hoofdstuk 1 zijn er ook vele andere vormen van onderzoek, en kunnen hypotheses ook op andere wijze en volgens andere paradigmata onderzocht worden, maar we beperken ons hier tot experimenteel onderzoek.

Table 2.1: Mogelijke oorzaken en mogelijke gevolgen.

onderwerp	vermeende oorzaak	vermeend gevolg
handel	buitentemperatuur	aantal verkochte ijsjes
zorg	type behandeling	mate van herstel
onderwijs	lesmethode	prestatie in toets
taal	beginleeftijd van onderwijs	mate van taalbeheersing
onderwijs	klassegrootte	schoolprestatie algemeen
zorg	temperatuur	hoogte van malaria-gebieden
taal	leeftijd	spreeksnelheid

onderwerp	vermeende oorzaak	vermeend gevolg
zorg	ligtijd voedsel op grond	mate van bacteriële besmetting

In experimenteel onderzoek wordt het effect onderzocht van een door de onderzoeker gemanipuleerde variabele op een andere variabele. In de inleiding is al een voorbeeld gegeven van een experimenteel onderzoek. Een nieuwe lesmethode werd beproefd door leerlingen te verdelen over twee groepen. De ene groep kreeg les volgens een nieuwe methode, terwijl de andere groep het gebruikelijke onderwijs genoot. De onderzoeker hoopte en verwachtte dat zijn nieuwe lesmethode een gunstig effect zou hebben, d.w.z. dat het zou leiden tot betere prestaties.

In hypothese-toetsend onderzoek wordt nagegaan of de onderzochte variabelen inderdaad met elkaar samenhangen op de verwachte wijze. In deze definitie staan twee termen centraal: 'variabelen' en 'op de verwachte wijze'. Voordat we nader ingaan op experimenteel onderzoek zullen we deze termen nader beschouwen.

2.2 Variabelen

Wat is een variabele? Grofweg is een variabele een eigenschap van objecten of personen die kan variëren, en die dus verschillende waarden kan aannemen. Laten we twee eigenschappen van personen bekijken: het aantal broers en zussen, en het geslacht van de moeder van die persoon. De eerste eigenschap kan variëren tussen personen, en is dus een variabele (tussen personen). De tweede eigenschap kan niet variëren: als er een moeder is, dan is die altijd en per definitie van het vrouwelijke geslacht. De tweede eigenschap is dus niet een variabele, maar een constante eigenschap.

In onze wereld bestaat bijna alles in een variabele hoeveelheid of hoedanigheid of mate. Ook een eigenschap die lastig te definiëren is, zoals de populariteit van een persoon in een groep, kan een variabele vormen. We kunnen immers personen in een groep rangschikken van meer tot minder populair. Voorbeelden van variabelen zijn er te over:

- van *personen*: hun lengte, hun gewicht, schoenmaat, spreeksnelheid, aantal broers en zussen, aantal kinderen, politieke voorkeur, inkomen, geslacht, populariteit in een groep, enz.
- van *teksten*: het totaal aantal woorden ('tokens'), aantal verschillende woorden ('types'), aantal spelfouten, aantal zinnen, aantal leestekens, enz.
- van *woorden*: de gebruiksfrequentie, aantal lettergrepen, aantal klanken, grammaticale woordsoort, enz.

- van *objecten* zoals auto's, telefoons, enz.: het gewicht, aantal componenten, energieverbruik, kostprijs, enz.
- van *organisaties*: het aantal werknemers, postcode, omzet, aantal burgers of klanten of patiënten of leerlingen, aantal operaties of diploma's of transacties, rechtsvorm, enz.

2.3 Onafhankelijke en afhankelijke variabelen

In hypothese-toetsend onderzoek kennen we twee soorten variabelen: de afhankelijke en de onafhankelijke variabele. De *onafhankelijke variabele* is dat wat het veronderstelde effect teweeg moet brengen. De onafhankelijke variabele is het aspect dat in een onderzoek door de onderzoeker gemanipuleerd wordt. In het voorbeeld waar een experiment uitgevoerd wordt om het effect van een nieuwe lesmethode te evalueren, vormt die lesmethode de onafhankelijke variabele. Wanneer de prestaties van de leerlingen die de nieuwe lesmethode gevolgd hebben vergeleken worden met de prestaties van leerlingen die alleen traditioneel schrijfonderwijs gevolgd hebben, dan neemt de onafhankelijke variabele twee waarden aan. Deze twee waarden (ook wel *niveau's* genoemd) van de onafhankelijke variabele kunnen we in dit voorbeeld benoemen als "experimenteel" en "controle", of als "nieuw" en "oud". We zouden de waarden van de onafhankelijke variabele ook kunnen uitdrukken als een getal, 1 resp. 0. Deze getallen hebben geen numerieke betekenis (we zouden de waarden ook 17 resp. 23 kunnen noemen), maar worden hier enkel gebruikt als willekeurige etiketten om verschillende groepen te onderscheiden. De gemanipuleerde variabele wordt 'onafhankelijk' genoemd omdat de gekozen (gemanipuleerde) waarden van deze variabele in een onderzoek niet afhankelijk zijn van iets anders: de onderzoeker is onafhankelijk in zijn of haar keuze van de gekozen waarden. Een onafhankelijke variabele wordt ook wel *factor* of soms *predictor* genoemd.

Het tweede type variabele is de *afhankelijke variabele*. De afhankelijke variabele is de variabele waarvoor we het veronderstelde effect verwachten. De onafhankelijke variabele veroorzaakt dus mogelijkerwijs een effect op de afhankelijke variabele, of: men veronderstelt dat de waarde van de afhankelijke variabele afhankelijk is van de waarde van de onafhankelijke variabele — vandaar hun benamingen. De afhankelijke variabele is dus datgene wat we meten of observeren. Een geobserveerde waarde van de afhankelijke variabele wordt ook wel *responsie* of *score* genoemd; ook de afhankelijke variabele zelf wordt vaak zo aangeduid. In het voorbeeld waar een experiment uitgevoerd wordt om het effect van een nieuwe lesmethode op de prestaties van leerlingen te evalueren, vormen die prestaties van de leerlingen de afhankelijke variabele. Andere voorbeelden zijn de spreeksnelheid, of de score op een vragenlijst, of het aantal malen dat een product verkocht wordt (zie Tabel 2.1). Kortom, in principe kan elke variabele als afhankelijke variabele gebruikt worden. Het is voornamelijk

de vraagstelling die bepaalt welke afhankelijke variabele gekozen wordt, en hoe deze gemeten wordt.

De onafhankelijke en afhankelijke variabelen dienen we overigens nadrukkelijk *niet* te interpreteren als 'oorzaak' resp. 'gevolg'. Het doel van het onderzoek is immers om overtuigend aan te tonen dàt er een (causaal) verband bestaat tussen de onafhankelijke en de afhankelijke variabele. In Hoofdstuk 5 zullen we echter zien hoe complex dat is.

De onderzoeker varieert de onafhankelijke variabele en observeert of dit resulteert in verschillen in de afhankelijke variabele. Als de waarden van de afhankelijke variabele verschillen voor en na de manipulatie van de onafhankelijke variabele, dan nemen we aan dat dit een gevolg is van de manipulatie van de onafhankelijke variabele. Er is sprake van een relatie tussen beide variabelen. Als de waarde van de afhankelijke variabele niet verschilt onder invloed van de waarden van de onafhankelijke variabele, dan is er geen verband tussen beide variabelen.

Voorbeeld 2.1: (Quené et al., 2012) onderzochten of een glimlach of frons invloed heeft op hoe luisteraars gesproken woorden verwerken. De woorden werden door de computer uitgesproken (gesynthetiseerd) in verschillende fonetische varianten, en wel op zo'n manier dat die woorden klonken alsof ze neutraal, of met een glimlach, of met een frons waren uitgesproken. Luisteraars moesten de woorden zo snel mogelijk classificeren als 'positief' danwel 'negatief' (qua betekenis). In dit onderzoek vormt de fonetische variant (neutraal, glimlach, frons) de onafhankelijke variabele, en de snelheid waarmee de luisteraars oordelen vormt de afhankelijke variabele.

2.4 Falsificatie en nul-hypothese

Het doel van wetenschappelijk onderzoek is om te komen tot een coherente verzameling van "justified true beliefs" (Morton, 2003). Een wetenschappelijke overtuiging moet dus deugdelijk onderbouwd en gerechtvaardigd zijn (en coherent met andere overtuigingen). Hoe komen we tot zo'n goede onderbouwing en rechtvaardiging? Daarvoor moeten we eerst terug naar het zgn. inductieprobleem van (Hume, 1739). Hume constateerde dat het logisch onmogelijk is om een bewering te generaliseren van een aantal specifieke gevallen (de waarnemingen in een onderzoek) naar een algemene regel (alle mogelijke waarnemingen in het universum).

Het probleem met deze generalisatie of inductie zullen we illustreren met de overtuiging 'alle zwanen zijn wit'. Als ik 10 zwanen heb gezien die allemaal wit zijn, dan zou ik dat kunnen beschouwen als een onderbouwing voor deze overtuiging. Deze generalisatie zou echter ook onterecht kunnen zijn: misschien bestaan er ook niet-witte zwanen, al heb ik die niet gezien. Meer algemeen: de inductie van specifieke waarnemingen naar een generalisatie houdt altijd een risico in, en kan niet gedaan worden "met behoud van waarheid". Er zit dus altijd een logische 'sprong' in, waardoor de generalisatie niet zonder risico is. Een regel die wel opgaat voor alle waargenomen specifieke gevallen ('alle zwanen zijn wit') hoeft daarmee nog niet een algemene regel te zijn. Hetzelfde inductie-probleem blijft bestaan als ik 100 of 1000 witte zwanen heb gezien. Maar wat als ik één zwarte zwaan heb gezien? Dan weet ik meteen, met zekerheid, dat de overtuiging dat alle zwanen wit zijn, niet waar is. Dit principe gebruiken we ook in wetenschappelijk onderzoek.

Laten we terugkeren naar ons eerdere voorbeeld waarin we hebben verondersteld dat een nieuwe lesmethode beter is dan een oude lesmethode; deze overtuiging noemen we H1. Laten we deze redenering nu eens omdraaien, en ons baseren op de complementaire overtuiging[1] dat de nieuwe methode *niet* beter is dan de oude; deze overtuiging noemen we de nul-hypothese of H0. Deze overtuiging H0 'alle methoden hebben gelijk effect' is analoog aan de overtuiging 'alle zwanen zijn wit' uit het voorbeeld in de vorige alinea. Hoe moeten we nu toetsen of de overtuiging of hypothese H0 waar is? Laten we daarvoor een representatieve steekproef van leerlingen trekken (zie Hoofdstuk **??**), en laten we de leerlingen volgens het toeval toewijzen aan de nieuwe of oude lesmethode (waarden van onafhankelijke variabele); we observeren vervolgens alle prestaties (afhankelijke variabele) van alle deelnemende leerlingen, volgens hetzelfde protocol voor alle gevallen. Vooralsnog veronderstellen we dat H0 waar is. We verwachten dus ook geen verschil tussen de prestaties van de verschillende groepen leerlingen. Als de leerlingen van de nieuwe methode desalniettemin veel beter blijken te presteren dan de leerlingen van de oude methode, dan vormt dat waargenomen verschil de figuurlijke zwarte zwaan: het gevonden verschil (dat in tegenspraak is met H0) maakt het onwaarschijnlijk dat H0 waar is (*mits* het onderzoek valide was; meer daarover in het volgende hoofdstuk). Omdat H0 en H1 elkaar uitsluiten, is het dan dus ook erg waarschijnlijk dat H1 wèl waar is. En omdat we onze onderbouwing baseerden op H0 en niet op H1, kunnen sceptici ons niet van partijdigheid beschuldigen: we probeerden immers juist aan te tonen dat er géén verschil was tussen de prestaties van de leerlingen uit de twee groepen.

Deze methode wordt *falsificatie* genoemd, omdat we kennis verwerven door hypothesen te verwerpen (falsifiëren) en niet door hypothesen te aanvaarden (verifiëren). Deze methodologie is ontwikkeld door de wetenschapsfilosoof Karl Popper (Popper, 1935, 1959, 1963). De falsificatie-methode heeft interessante overeenkomsten met de evolutietheorie. Door variatie tussen de individuen kun-

[1] Twee beweringen zijn complementair als ze elkaar wederzijds uitsluiten, zoals H1 en H0 in dit voorbeeld.

nen sommigen zich succesvol voortplanten, terwijl veel anderen voortijdig sterven en/of zich niet voortplanten. Op analoge wijze kunnen sommige tentatieve beweringen niet weerlegd worden, en kunnen deze dus 'overleven' en 'zich voortplanten', terwijl veel andere beweringen weerlegd worden en dus 'sterven'. In de woorden van (Popper, 1963, p.51):

> " … to explain (the world) … as far as possible, with the help of laws and explanatory theories …there is no more rational procedure than the method of trial and error — of conjecture and refutation: of boldly proposing theories; of trying our best to show that these are erroneous; and of accepting them tentatively if our critical efforts are unsuccessful."

Een goede wetenschappelijke bewering of theorie dient dus falsifieerbaar of weerlegbaar of toetsbaar te zijn (Popper, 1963), d.w.z. het moet mogelijk zijn om de onjuistheid van die bewering of theorie aan te tonen. De wetenschappelijke onderbouwing en daarmee de plausibiliteit van een toetsbare bewering neemt toe, naarmate die bewering vaker en onder meer wisselende omstandigheden bestand is gebleken tegen falsificatie. 'Het klimaat wordt warmer' is een goed voorbeeld van een bewering die steeds beter bestand blijkt te zijn tegen falsificatie, en die daarmee steeds sterker wordt.

Voorbeeld 2.2: 'Alle zwanen zijn wit' en 'de gemiddelde temperatuur van de aarde stijgt sinds 1900' zijn falsifieerbare, en daarom wetenschappelijk bruikbare beweringen. Maar hoe zit dat met de volgende beweringen?
a. Goud lost op in water.
b. Zout lost op in water.
c. Vrouwen praten meer dan mannen.
d. De muziek van Coldplay is beter dan die van U2.
e. De muziek van Coldplay verkoopt beter dan die van U2.
f. Als een patiënt een duiding van de psychoanalyticus afwijst, dan is dat het gevolg van weerstand omdat de duiding van de psychoanalyticus juist is.
g. De stijging van de gemiddelde temperatuur van de aarde is het gevolg van menselijke activiteiten.

2.5 De empirische cyclus

In het voorafgaande hebben we op een vrij globale manier kennis gemaakt met experimenteel onderzoek. In deze paragraaf beschrijven we het verloop

van experimenteel onderzoek meer systematisch. Er zijn in de loop der tijd verschillende schema's opgesteld waarin onderzoek in fasen beschreven wordt. De bekendste van deze schema's is waarschijnlijk wel de empirische cyclus van (De Groot, 1961).

In de empirische cyclus worden vijf onderzoeksfasen onderscheiden: de observatiefase, de inductiefase, de deductiefase, de toetsingsfase en de evaluatiefase. In de laatste fase worden tekortkomingen en alternatieve interpretaties geformuleerd. Dit leidt weer tot nieuw onderzoek, waarin opnieuw de serie fasen kan worden doorlopen (vandaar 'cyclus'). Deze vijf onderzoeksfasen zullen wij één voor één behandelen.

2.5.1 observatie

In deze fase construeert de onderzoeker een probleem. Dat wil zeggen dat de onderzoeker een idee vorm over de mogelijke relaties tussen verschillende (theoretische) concepten of constructen. Deze veronderstellingen worden later uitgewerkt tot meer algemene hypothesen. Veronderstellingen kunnen op duizenden manieren tot stand komen — maar vereisen altijd nieuwsgierigheid van de onderzoeker. De onderzoeker kan een vreemd fenomeen opmerken dat verklaard moet worden, bv het fenomeen dat het vermogen om absolute toonhoogte te horen ("absoluut gehoor") veel vaker voorkomt bij Chinezen dan bij Amerikanen (Deutsch, 2006). Ook het systematisch doorzoeken van wetenschappelijke publicaties kan leiden tot veronderstellingen. Soms blijkt dan dat de resultaten van verschillende onderzoeken elkaar tegenspreken, of dat er een duidelijke lacune zit in onze kennis.

Veronderstellingen kunnen ook gebaseerd zijn op case-studies: onderzoeken waarbij één of enkele gevallen intensief bestudeerd en extensief beschreven worden. Zo ontwikkelde Piaget zijn theorie over de verstandelijke ontwikkeling van kinderen op basis van observaties van zijn eigen kinderen in de tijd dat hij werkloos was. Deze observaties vormden later, toen Piaget zijn eigen laboratorium had, aanleiding voor vele experimenten op basis waarvan hij zijn theoretische inzichten kon verdiepen en verifiëren.

Het is belangrijk om te beseffen dat puur onbevangen, objectieve waarneming niet mogelijk is. Waarnemingen zijn altijd min of meer theorie-geladen of kennis-geladen. Als we niet weten waarop we moeten letten, kunnen we ook niet goed waarnemen. Zo kunnen wolken-experts veel meer typen van bewolking onderscheiden en interpreteren dan leken. Voordat er observaties gedaan worden en feiten worden geanalyseerd, is het dus verstandig om eerst een expliciet theoretisch kader aan te brengen, ook al is dit nog rudimentair.

Een onderzoeker komt tot veronderstellingen naar aanleiding van opmerkelijke verschijnselen, case-studies, literatuurstudie, e.d. Er zijn echter geen methodologische richtlijnen over hoe dit proces zou moeten verlopen: het is een creatief proces.

2.5.2 inductie

In de inductiefase wordt de in de observatiefase geopperde veronderstelling gegeneraliseerd. Op grond van specifieke observaties wordt nu een hypothese geopperd waarvan de onderzoeker vermoedt dat die algemeen geldig is. (**Inductie** is de logische stap waarbij een algemene bewering of hypothese wordt afgeleid uit specifieke gevallen: mijn kinderen (hebben) leren praten → alle kinderen (kunnen) leren praten.)

Zo kan een onderzoeker uit de observatie dat de vrouwen in zijn/haar omgeving meer praten dan de mannen (meer minuten per etmaal, en meer woorden per etmaal), een algemene hypothese afleiden: H1: vrouwen praten meer dan mannen (zie Voorbeeld 2.2); deze hypothese kan nader ingeperkt worden in tijd en plaats.

De hypothese moet tevens een duidelijk omschreven empirische inhoud hebben, d.w.z. het type of de klasse van observaties moet goed omschreven zijn. Gaat het over alle vrouwen en mannen? Of alleen sprekers van het Nederlands? En hoe zit het met meertalige sprekers? En met kinderen die hun taal nog aan het leren zijn? Die duidelijk omschreven inhoud is nodig om de hypothese te kunnen toetsen (zie subsectie Toetsing hieronder, en zie Hoofdstuk **??**).

Tenslotte moet een hypothese ook logisch coherent zijn, d.w.z. de hypothese moet aansluiten bij andere theorieën of hypothesen. Als een hypothese niet logisch coherent is, dan kan zij per definitie niet eenduidig aan de empirie gerelateerd worden, en is zij dus niet goed toetsbaar. Hieruit volgt dat een hypothese niet multi-interpretabel mag zijn: een hypothese moet op zichzelf één en niet meer dan één uitkomst van een experiment voorspellen.

In het algemeen worden drie typen hypothesen onderscheiden (De Groot, 1961):

- Universeel-deterministische hypothesen.
  Deze hebben als algemene vorm: *alle A's zijn B's.* Bijvoorbeeld: alle zwanen zijn wit, alle (volwassen) mensen kunnen spreken. Als een onderzoeker voor één A kan aantonen dat deze niet B is, dan is de hypothese in beginsel gefalsifieerd. Een universeel deterministische hypothese kan nooit geverifieerd worden; een onderzoeker kan alleen een uitspraak doen over de gevallen die hij geobserveerd, dan wel gemeten heeft. Bij een oneindige verzameling, zoals: alle vogels, of alle mensen, of alle kachels, kan dit tot problemen leiden. De onderzoeker weet niet of er misschien één enkel geval bestaat waarin geldt: A is niet B; er is één vogel die niet kan vliegen, et cetera. Over deze andere gevallen kan dus geen uitspraak gedaan worden, waardoor de universele geldigheid van de hypothese nooit volledig 'bewezen' kan worden.

- Deterministische existentiehypothesen.
  Deze hebben als algemene vorm: *er is tenminste één A die B is.* Bijvoorbeeld: er is tenminste één zwaan die wit is, er is tenminste één mens

die kan praten, er is tenminste één kachel die warmte geeft. Als een onderzoeker kan aantonen dat er één A is die B is, dan is de hypothese geverifieerd. Deterministische existentiehypothesen kunnen echter nooit gefalsifieerd worden. Daarvoor zou het nodig zijn om van een oneindige verzameling alle eenheden of individuen te onderzoeken op het al dan niet B zijn, en dat is door de oneindigheid van de verzameling nu juist uitgesloten. Hieruit blijkt tegelijk dat dit type hypothesen geen algemene uitspraken doen, en dat het wetenschappelijk belang ervan niet zo duidelijk is. Je kunt het ook zo zeggen: voor elk specifiek geval A doet een dergelijke hypothese helemaal geen duidelijke voorspelling; een gegeven A zou de gezochte B kunnen zijn, maar dat hoeft helemaal niet. In deze zin voldoet een deterministische existentiehypothese dan ook niet aan ons criterium van falsificatie.

- Probabilistische hypothesen.
  Deze hebben als algemene vorm: *er zijn relatief meer A's die B zijn, dan niet-A's die B zijn.* In de gedragswetenschappen (inclusief taal en communicatie) is dit verreweg het meest voorkomende type hypothese.
  Bijvoorbeeld: er zijn relatief meer vrouwen die veelpratend zijn dan mannen die veelpratend zijn. Of: er zijn relatief meer hoog-presterende leerlingen bij de nieuwe methode dan bij de oude methode. Of: versprekingen treden relatief vaker op bij het begin dan bij het einde van een woord. Daarmee wordt nog niet aangegeven dat alle vrouwen meer praten dan alle mannen, en evenmin wordt aangegeven dat alle leerlingen met de nieuwe methode beter presteren dan alle leerlingen van de oude methode.

2.5.3 deductie

In deze fase van de empirische cyclus worden specifieke voorspellingen afgeleid uit de algemeen geformuleerde hypothese die is opgezet in de inductiefase. (**Deductie** is de logische stap waarbij een specifieke bewering of voorspelling wordt afgeleid uit een meer algemene bewering: alle kinderen leren praten → mijn kinderen (zullen) leren praten.)

Laten we veronderstellen (H1) dat "vrouwen meer praten dan mannen". Uit deze hypothese doen we in deze fase specifieke voorspellingen voor specifieke steekproeven. Wanneer we bijvoorbeeld 40 vrouwelijke en 40 mannelijke docenten Nederlands zouden interviewen, zonder tijdsbeperking, dan luidt de voorspelling op grond van deze H1 dat de vrouwelijke docenten in deze steekproef meer zullen zeggen dan de mannelijke docenten in de steekproef (en dus ook, dat ze een groter aantal lettergrepen zullen spreken in het interview).

Zoals hierboven uitgelegd (§2.4), wordt in het meeste wetenschappelijk onderzoek echter niet de H1 getoetst, maar de logische tegenhanger daarvan, die met H0 wordt aangeduid. Voor de toetsing (in de volgende fase van de empirische cyclus) is het dus gebruikelijk om voorspellingen te toetsen die zijn afgeleid uit

de H0 (!), bijvoorbeeld "vrouwen en mannen produceren *even veel* lettergrepen in een vergelijkbaar interview".

In de praktijk worden de termen 'hypothese' en 'voorspelling' vaak door elkaar gebruikt, en spreken we vaak over het toetsen van hypothesen. Volgens bovenstaande terminologie toetsen we echter niet de hypothesen, maar leiden we uit de hypothesen voorspellingen af (via deductie), en toetsen we daarna die voorspellingen aan de data.

2.5.4 toetsing

In deze fase verzamelen we empirische observaties en vergelijken we die met de uitgewerkte voorspellingen "onder H0", d.w.z. de voorspellingen als H0 waar zou zijn. In Hoofdstuk **??** zullen we nader ingaan op deze toetsing. Hier introduceren we alleen het algemene principe om nulhypotheses te toetsen. (Naast het hier beschreven conventionele "frequentistische" principe kunnen we ook hypotheses toetsen of vergelijken op een nieuwere "Bayesiaanse" wijze; we bespreken die in §**??**).

Als de observaties buitengewoon onwaarschijnlijk zijn onder H0, dan zijn er twee logische mogelijkheden. (i) De observaties deugen niet, we hebben fout geobserveerd. Maar als de onderzoeker zijn werk goed gecontroleerd heeft en zichzelf serieus neemt, dan is dat niet waarschijnlijk. (ii) De voorspelling was onjuist, H0 is wellicht niet juist, en moet dus verworpen worden, ten gunste van H1.

In ons voorbeeld hierboven (in de voorgaande subsectie over deductie) hebben we uit H0 (!) de voorspelling afgeleid dat in een steekproef van 40 mannelijke en 40 vrouwelijke docenten, de leden van de twee groepen even veel lettergrepen gebruiken in een gestandaardiseerd interview. We vinden echter dat de mannen meer lettergrepen gebruiken (gemiddeld 4210 lettergrepen) dan de vrouwen (gemiddeld 3926 lettergrepen) (Quené, 2008, p.1112). Hoe waarschijnlijk is dit verschil als de observaties kloppen, en als H0 waar zou zijn? Die kans is zodanig klein dat de onderzoeker H0 verwerpt (zie optie (ii) hierboven), en concludeert dat vrouwen en mannen *niet even veel* praten, althans in dit onderzoek.

In het bovenstaande voorbeeld worden in de toetsingsfase twee groepen vergeleken, hier mannen en vrouwen. Eén van die twee groepen is vaak een neutrale groep of controle-groep, zoals we al zagen in het eerdere voorbeeld van de nieuwe en oude lesmethode. Waarom maken onderzoekers vaak gebruik van zo'n controle-groep? Stel je eens voor dat we alleen de nieuwe-methode-groep zouden onderzoeken. In de toetsingsfase meten we de prestaties van de leerlingen, en die is ruim voldoende: gemiddeld een 7. Is de nieuwe methode dan een succes? Misschien niet: als de leerlingen volgens de oude methode een 8 zouden behalen, dan zou de nieuwe methode eigenlijk slechter zijn, en zouden we de nieuwe methode beter niet kunnen invoeren. Om daar een zinnige conclusie over te kunnen trekken, is het essentieel om de nieuwe en

oude methoden onderling te vergelijken. Vandaar dat in veel onderzoek een neutrale conditie, nul-conditie, controle-groep, placebo-behandeling, o.i.d, is opgenomen.

Hoe kunnen we nu de kans bepalen op de gevonden observaties, als H0 waar zou zijn? Dat is vaak wat complex, maar we illustreren het hier met een eenvoudig voorbeeld. We gooien kop of munt met een munt. We veronderstellen (H0): de munt is eerlijk, de kans op kop is 1/2 per worp. We gooien 10× met dezelfde munt, en wonderbaarlijk genoeg observeren we alle 10× een kop als uitkomst. De kans dat dit gebeurt, als H0 waar is, is $P = (1/2)^{10} = 1/1024$. Als H0 waar zou zijn is deze uitkomst uiterst onwaarschijnlijk (al is de uitkomst niet onmogelijk, want $P > 0$), en daarom verwerpen we H0. We concluderen dus dat de munt waarschijnlijk niet eerlijk is.

Dit roept een belangrijk punt op: wanneer is een uitkomst zò onwaarschijnlijk dat we H0 verwerpen? Welk criterium hanteren we voor de kans op de gevonden observaties als H0 waar zou zijn? Dit is de vraag naar het significantieniveau, d.w.z. het kansniveau waarbij we besluiten de H0 te verwerpen. Dit wordt aangeduid met symbool α. Als in een onderzoek een significantieniveau gehanteerd wordt van $\alpha = 0.05$, dan wordt de H0 verworpen als de kans om deze resultaten te vinden als H0 waar is[2], kleiner is dan 5%. De uitkomst is dan zo onwaarschijnlijk, dat we ervoor kiezen om H0 te verwerpen (optie (ii) hierboven), d.w.z. we concluderen dat H0 waarschijnlijk *niet* waar is.

Als we H0 aldus verwerpen, dan lopen we wel een kleine kans dat we eigenlijk met optie (i) te maken hebben: H0 is waar, maar de observaties wijken *toevallig* sterk af van de voorspelling op basis van H0, en H0 wordt dan ten onrechte verworpen. Dit wordt een Type-I-fout genoemd. Deze fout is vergelijkbaar met de onterechte veroordeling van een onschuldig persoon, of met de onjuiste classificatie van een onschuldig email-bericht als 'spam'. Meestal wordt $\alpha = .05$ gebruikt, maar ook andere significantie-niveau's zijn mogelijk en soms verstandiger.

Merk op dat de significantie betrekking heeft op de kans om de gevonden extreme gegevens (of meer extreme gegevens) te vinden, indien H0 waar is:

$$\text{significantie} = P(\text{data}|\text{H0})$$

De significantie is dus *niet* de kans dat H0 waar is als je deze gegevens gevonden hebt, $P(\text{H0}|\text{data})$, hoewel we deze denkfout vaak tegenkomen.

Bij iedere vorm van toetsing is er ook een kans op de omgekeerde fout, nl. dat we H0 ten onrechte niet verwerpen. Dat wordt een Type-II-fout genoemd: H0 is eigenlijk niet waar (dus H1 is waar) maar H0 wordt desalniettemin niet verworpen. Deze fout is vergelijkbaar met de onterechte vrijspraak van een schuldig persoon, of met het onterecht goedkeuren van een *spam* email-bericht (zie Tabel 2.2).

[2]Vollediger: Als de kans om deze resultaten te vinden, of resultaten die nog meer verschillen van de door H0 voorspelde resultaten, kleiner is dan 5%, dan wordt H0 verworpen.

Table 2.2: Mogelijke uitkomsten van beslissingsprocedure.

werkelijkheid		
	H0 verworpen	**H0 niet verworpen**
H0 is waar (H1 onwaar)	Type-I-fout (α)	correct
H0 is onwaar (H1 waar)	correct	Type-II-fout (β)
	verdachte veroordeeld	**verdachte vrijgesproken**
verdachte is onschuldig (H0)	Type-I-fout	correct
verdachte is schuldig	correct	Type-II-fout
	bericht weggegooid	**bericht doorgestuurd**
bericht is OK (H0)	Type-I-fout	correct
bericht is spam	correct	Type-II-fout

Als we het significantieniveau hoger instellen, bv. $\alpha = .20$, dan is de kans om de H0 te verwerpen dus ook veel groter. In de toetsingsfase verwerpen we immers H0 al indien de kans op deze gegevens (of meer extreme gegevens) kleiner is dan 20%. Een uitkomst van 8× kop in 10 worpen is dan al voldoende om H0 te verwerpen (d.i. om de munt als onzuiver te beoordelen). Er zijn dus meer uitkomsten mogelijk waarbij we H0 zullen verwerpen. Dat hogere significantieniveau houdt dus een groter risico in op een Type-I-fout, en tegelijk een kleiner risico op een Type-II-fout. De afweging tussen de twee typen fouten hangt af van de precieze omstandigheden van het onderzoek, en van de consequenties van de twee typen van fouten. Welke fout is ernstiger: een goed bericht weggooien, of een *spam*-bericht doorsturen? De kans op een Type-I-fout (significantieniveau) heeft een onderzoeker in eigen hand. De kans op een Type-II-fout is afhankelijk van drie factoren, en is lastig te bepalen. We zullen dat nader bespreken in Hoofdstuk **??**.

2.5.5 evaluatie

Aan het einde van het onderzoek moet de onderzoeker de onderzoeksresultaten evalueren: wat is het nu allemaal waard? Het draait hier niet slechts om de vraag of de onderzoeksresultaten al dan niet ten gunste van de getoetste theorie uitgevallen zijn. Het gaat om een kritische beschouwing van de wijze waarop de data zijn verzameld, de denkstappen, de operationalisatie, de mogelijke alternatieve verklaringen, alsmede de consequenties van de resultaten. De resultaten moeten in een bredere context geplaatst en besproken worden. Wellicht leiden de conclusies ook tot aanbevelingen, bijvoorbeeld voor klinische toepassingen of voor de onderwijspraktijk. Dit is ook het moment om suggesties voor ander of vervolgonderzoek te doen.

In deze fase gaat het primair om de interpretatie van de resultaten, waarbij de onderzoeker als interpretator een belangrijke en persoonlijke rol speelt. Verschillende onderzoekers kunnen dezelfde uitkomsten geheel anders interpreteren. En

soms zijn de resultaten in tegenspraak met wat was voorspeld of gewenst.

2.6 Keuzemomenten

Onderzoek bestaat uit een reeks van keuze-momenten: van de inspirerende observaties in de eerste fase, via de operationele beslissingen in de uitvoering van het onderzoek, tot de interpretatie van de resultaten in de laatste fase. Zelden zal een onderzoeker in staat zijn om altijd de beste keuze te maken, maar hij of zij moet er voor waken dat ergens een slechte keuze gemaakt zou worden. Het hele onderzoek is net zo sterk als de zwakste schakel: de waarde van het hele onderzoek hangt af van de slechtste keuze in de reeks van keuzes. Ter illustratie geven we een beeld van de keuzes die een onderzoeker moet maken tijdens de gehele empirische cyclus.

De eerste keuze die gemaakt moet worden betreft de probleemstelling. Relevante vragen die de onderzoeker op dit moment moet beantwoorden zijn: hoe herken ik een bepaalde onderzoeksvraag, is onderzoek hier het aangewezen middel, is dit idee onderzoekbaar? De beantwoording van dergelijke vragen is van allerlei factoren afhankelijk, zoals mens- en maatschappijvisie, wensen van de opdrachtgever, financiële en praktische mogelijkheden, enz.

De onderzoeksvraag moet wel te beantwoorden zijn met de beschikbare methoden en middelen. Maar binnen die beperking kan de onderzoeksvraag elk aspect van de werkelijkheid betreffen, ongeacht of dit aspect nu irrelevant of belangrijk wordt geacht. Er zijn vele voorbeelden van onderzoek dat aanvankelijk werd afgedaan als irrelevant, maar dat desondanks wel degelijk van wetenschappelijke waarde bleek te zijn, bijvoorbeeld een studie over de vraag “is ‘Huh?’ a universal word?” (Dingemanse et al., 2013) (Voorbeeld 1.1). Ook bleken ideeën die eerst als onjuist werden afgedaan later toch te kloppen met de werkelijkheid. Zo beweerde Galilei, zogenaamd ‘ten onrechte’, dat de aarde om de zon draaide. Kortom, onderzoeksvragen moeten niet te snel verworpen worden omdat zij ‘nutteloos’, een ‘open deur’, ‘irrelevant’ of ‘triviaal’ zouden zijn.

Als de onderzoeker besluit om verder te gaan met het onderzoek, dan is de volgende stap doorgaans literatuurstudie. In de meeste handboeken wordt aanbevolen veel te lezen, maar hoe wordt de literatuur verzameld? Uiteraard moet de relevante onderzoeksliteratuur over het probleemgebied doorgenomen worden. Gelukkig bestaan er tegenwoordig allerlei hulpmiddelen om relevante wetenschappelijke publicaties te vinden. Het is raadzaam om daarvoor de aanwijzingen en zgn. “libguides” te bestuderen die de Universiteitsbibliotheek aanbiedt (zie http://www.uu.nl/bibliotheek, en http://libguides.library.uu.nl). Ook de gids van (Sanders, 2011) bevelen we ten zeerste aan: de gids bevat vele uiterst nuttige aanwijzingen over het opsporen van relevante onderzoeksliteratuur.

In de fase daarna doemen de eerste methodologische problemen op. De onderzoeker moet namelijk de probleemstelling exacter formuleren. Een belangrijke

afweging die hier gemaakt dient te worden is of de probleemstelling wel onderzoekbaar is (§2.4). Een vraag als "wat is het effect van beginleeftijd van leren op de taalvaardigheid in een vreemde taal?" is bijvoorbeeld niet zonder meer onderzoekbaar. Deze vraag moet nader gespecificeerd worden. Cruciale concepten moeten ge(her)definieerd worden: wat is de beginleeftijd van het leren van een vreemde taal? Wat is taalvaardigheid? Wat is een effect? En wat is eigenlijk een vreemde taal? Hoe definieer ik de populatie? De onderzoeker wordt geconfronteerd met allerlei vragen over definities en operationalisatie: Worden begrippen theoretisch of empirisch of pragmatisch gedefinieerd? Welke instrumenten worden gebruikt om de verschillende constructen te meten? Maar ook: hoe ingewikkeld moet het onderzoek worden? Kan het hele onderzoek dan wel tot een goed einde worden gebracht? Op welke wijze moeten de gegevens verzameld worden? Kunnen de gewenste gegevens wel verzameld worden, of zullen respondenten dergelijke vragen nooit (kunnen) beantwoorden? Is de voorgestelde manipulatie ethisch verantwoord? Wat is de afstand tussen het theoretische construct en de wijze waarop dat zal worden gemeten? Wanneer in deze fase iets fout gaat, dan heeft dat direct weerslag op de rest van het onderzoek.

Als er met succes een probleemstelling is geformuleerd en geoperationaliseerd, dan volgt een nadere literatuurverkenning. Dit tweede literatuuronderzoek is veel meer toegespitst op de inmiddels uitgewerkte onderzoeksvraag dan de eerder genoemde, brede literatuurverkenning. Op grond van eerdere publicaties kan de onderzoeker zijn of haar oorspronkelijke probleemstelling heroverwegen. Niet alleen moet de literatuur nu doorgenomen worden met het oog op inhoudelijk theoretische overwegingen, maar ook moet aandacht worden besteed aan voorbeelden van operationalisering van de kernbegrippen. Zijn deze begrippen wel goed geoperationaliseerd, en als er verschillende manieren van operationalisering zijn, wat is dan de ratio achter deze verschillen? En, kunnen de kernbegrippen zo geoperationaliseerd worden dat de afstand tussen het begrip-zoals-bedoeld en het begrip-zoals-bepaald (nog) kleiner is (§1.3)? De aanwijzingen hierboven voor het zoeken in wetenschappelijke literatuur zijn hier wederom van nut. De onderzoeker dient zich vervolgens (nogmaals) te beraden op het nut van het onderzoek. Afhankelijk van de probleemstelling moeten vragen gesteld worden als: draagt het onderzoek bij aan de kennis op een bepaald gebied, worden door het onderzoek oplossingen gecreëerd voor ervaren knelpunten of problemen, of draagt het onderzoek bij aan te creëren oplossingen? Voldoet de vraagstelling nog aan het oorspronkelijke probleem (of de oorspronkelijke vraagstelling) van de opdrachtgevers? Zijn er voldoende (technische, financiële, praktische) mogelijkheden om het onderzoek uit te voeren?

In de volgende stap moet worden gespecificeerd hoe de gegevens worden verzameld. Dit is een essentiële stap die van invloed is op de rest van het onderzoek; we wijden er daarom een apart hoofdstuk aan (Hoofdstuk **??**). Waaruit bestaat de populatie: uit taalgebruikers? leerlingen? tweetalige babies? versprekingen van medeklinkers? zinnen? En hoe moet je een representatieve steekproef of steekproeven trekken uit deze populatie(s)? Hoe groot moet die steekproef dan zijn? Ook moet er in deze fase gekozen worden voor een analysemethode. Het

is zelfs aan te bevelen om in deze fase al een analyseplan te ontwerpen. Welke analyses zullen worden uitgevoerd, welke exploraties van de gegevens worden voorzien?

Met al deze keuzes zijn de voorbereidingen nog niet afgerond. Ook de instrumenten moeten worden gekozen: welke apparatuur, opname-gereedschap, vragenlijsten, enz., worden gebruikt om waarnemingen mee te doen? Bestaan er al geschikte instrumenten? Zo ja, zijn deze dan makkelijk toegankelijk en mogen zij gebruikt worden? Zo nee, dan moeten instrumenten ontwikkeld worden (§1.3). Maar in dat geval neemt de onderzoeker ook de taak op zich om deze instrumenten eerst te beproeven, om na te gaan of de gegevens die met deze instrumenten verkregen worden, voldoen aan de kwaliteitseisen die de onderzoeker zich gesteld heeft, of die in het algemeen aan de instrumenten in wetenschappelijk onderzoek gesteld mogen worden (in termen van betrouwbaarheid en validiteit, zie Hoofdstukken 5 en **??**.

Pas nadat ook de instrumenten in gereedheid gebracht zijn begint het eigenlijke empirische onderzoek: de gekozen gegevens van de gekozen steekproef worden verzameld op de gekozen wijze met behulp van de gekozen instrumenten. Ook hierbij zijn er allerlei, vaak praktische problemen waar de onderzoeker tegenaan loopt. Een waar gebeurd voorbeeld: drie dagen nadat een onderzoeker zijn vragenlijst verstuurd had begon een poststaking die twee weken duurde. Helaas had de onderzoeker de respondenten ook twee weken de tijd gegeven om te reageren. Dus toen de poststaking voorbij was, was de inzendtermijn verlopen. Wat moest hij toen? Bij gebrek aan alternatieven besloot de onderzoeker alle 1020 respondenten telefonisch te benaderen met het verzoek de vragenlijst alsnog in te vullen en te retourneren.

Voor de onderzoeker die zich de moeite getroost heeft van te voren een analyseplan op te stellen breekt nu de tijd aan om te oogsten. Eindelijk kunnen de geplande analyses ook uitgevoerd worden. Helaas blijkt de werkelijkheid meestal veel weerbarstiger dan de onderzoeker van te voren had bedacht. Proefpersonen geven onverwachte responsies, of houden zich niet aan de instructies, veronderstelde verbanden blijken niet aanwezig, en onverwachte (en ongewenste) verbanden blijken in sterke mate aanwezig. In latere hoofdstukken zullen we dieper ingaan op analysemethoden en problemen daarbij.

Tenslotte moet de onderzoeker ook rapporteren over het onderzoek. Zonder (adequaat) onderzoeksverslag zijn de gegevens niet toegankelijk en had het onderzoek net zo goed *niet* uitgevoerd kunnen worden. Dit is een essentiële stap, waarbij onder meer de vraag gesteld dient te worden of het onderzoek op basis van de verslaglegging controleerbaar en repliceerbaar is. Meestal wordt van onderzoeksactiviteiten verslag gedaan in de vorm van een werkstuk, een onderzoeksrapport of een artikel in een wetenschappelijk tijdschrift. Soms wordt van een onderzoek ook verslag gedaan in een meer populair tijdschrift, dat voor een bredere doelgroep bedoeld is dan alleen collega-onderzoekers.

Tot zover een beknopt overzicht van de keuzen die onderzoekers moeten maken

tijdens hun onderzoek. Ieder empirisch onderzoek bestaat uit een aaneenschakeling van problemen, keuzes en beslissingen. De belangrijkste keuzes zijn al gemaakt voordat de onderzoeker begint met gegevens verzamelen.

Chapter 3

Integriteit

3.1 Inleiding

Wetenschappelijk onderzoek heeft de mensheid onmetelijk grote baten opgeleverd, zoals betrouwbare computer-technologie, goede medische zorg, en begrip van andere talen en culturen. Al deze verworvenheden zijn gebaseerd op wetenschappelijk onderbouwde kennis. Onderzoekers produceren kennis, en de vooruitgang en groei van kennis ontstaat omdat onderzoekers voortbouwen op de ervaringen en inzichten van hun voorgangers.

Voorbeeld 3.1: Sir Isaac Newton schreef over zijn wetenschappelijke werk: "If I have seen further it is by standing on (the) shoulders of Giants" (in een brief aan Robert Hooke d.d. 5 Feb 1676[1]). Dit beeld is te herleiden tot de middeleeuwse geleerde Bernard de Chartres: "…nos esse quasi nanos gigantum umeris insidentes" (dat wij zijn als dwergen gezeten op de schouders van reuzen) in vergelijking tot geleerden uit de Oudheid. Newton's uitspraak is ook het motto van Google Scholar (scholar.google.com), een zoekmachine voor wetenschappelijke publicaties.

In dit hoofdstuk bespreken we de ethische en morele aspecten van wetenschappelijk onderzoek. Wetenschap is mensenwerk, en het vereist

[1] Een kopie van de brief is te lezen via http://digitallibrary.hsp.org/index.php/Detail/Object/Show/object_id/9285; voor achtergrond-informatie zie http://www.bbc.co.uk/worldservice/learningenglish/movingwords/shortlist/newton.shtml.

een goed ontwikkeld beoordelingsvermogen van de onderzoekers. De *Nederlandse Gedragscode Wetenschappelijke Integriteit* (VSNU, 2018) (http://www.vsnu.nl/wetenschappelijke_integriteit) beschrijft hoe wetenschappelijke onderzoekers (en studenten) zich dienen te gedragen. Volgens deze gedragscode dient wetenschappelijk onderzoek en onderwijs gebaseerd te zijn op de volgende principes:

- eerlijkheid,
- zorgvuldigheid,
- transparantie,
- onafhankelijkheid, en
- verantwoordelijkheid

In de volgende paragrafen zullen we nagaan hoe we volgens deze principes dienen te handelen bij de verschillende fasen van wetenschappelijk onderzoek. Hoe moeten we op eerlijke, zorgvuldige, transparante, onafhankelijke en verantwoordelijke wijze een onderzoek opzetten, de gegevens verzamelen en verwerken, en verslag doen van het onderzoek? We moeten daarover nadenken nog voor het onderzoek begint, en daarom bespreken we deze onderwerpen aan het begin van deze syllabus, hoewel we ook vooruit zullen wijzen naar termen en begrippen die worden uitgewerkt in volgende hoofdstukken.

3.2 Ontwerp

Weliswaar levert wetenschappelijk onderzoek ons onmetelijk grote baten op, maar daar staan ook aanzienlijke kosten tegenover. De directe kosten zijn o.a. de inrichting en onderhoud van laboratoria, apparatuur en technische ondersteuning, maar ook de loonkosten van de onderzoekers, vergoedingen voor informanten en proefpersonen, reiskosten voor toegang tot bibliotheken, archieven, informanten en proefpersonen, e.d. Deze directe kosten worden doorgaans gefinancierd uit publieke middelen van universiteiten en andere wetenschappelijke instellingen. Daarnaast zijn er indirecte kosten, die voor een deel ten laste komen van de informanten en proefpersonen: tijd en moeite die niet aan iets anders besteed kan worden, verlies van privacy, en mogelijke andere risico's die we nog niet kennen. Een vaak vergeten kostenpost is het verlies van onbevangenheid: een proefpersoon die heeft meegedaan aan een experiment leert daarvan, en reageert daarna misschien anders in een volgend experiment (zie §5.4, onder Geschiedenis). De resultaten uit zo'n volgend experiment zijn daardoor minder goed generaliseerbaar naar andere personen die een andere geschiedenis hebben, en *niet* eerder aan een onderzoek hebben meegedaan.

Gezien de grote kosten moet onderzoek zodanig zijn doordacht en ontworpen, dat de verwachte baten redelijkerwijs opwegen tegen de verwachte kosten (Rosenthal and Rosnow, 2008, Ch.3). Als de kans op valide conclusies uit een onderzoek erg klein is, dan is het beter om dat onderzoek *niet* uit te voeren, en zo de directe en indirecte kosten te besparen.

Voorbeeld 3.2: Stel dat we willen onderzoeken of tweetalige kinderen van 4 jaar oud een cognitief voordeel hebben boven eentalige leeftijdsgenoten. Op grond van eerder onderzoek verwachten we een verschil van tenminste 2 punten (op een 10-punts-schaal) tussen beide groepen (met "pooled standard deviation" $s_p = 4$, dus $d = 0.5$, zie § **??** en § **??**).

We vergelijken twee groepen van elk $n = 4$ kinderen. Zelfs als er inderdaad een verschil is van 2 punten tussen de twee groepen (dus als de onderzoekshypothese waar is), dan nog is er in dit onderzoek slechts 51% kans om een significant verschil te vinden: de power is slechts .51 (Hoofdstuk **??**), omdat de twee groepen zo weinig proefpersonen bevatten. De vierjarige kinderen en hun ouders kunnen beter andere dingen doen (school, thuis, werk) dan meedoen aan dit onderzoek.

Als er echter $n = 30$ kinderen in elk van de twee groepen zouden meedoen, en als er inderdaad een verschil is van 2 punten tussen de twee groepen (dus als de onderzoekshypothese waar is) dan zou de power .90 zijn. Met grotere groepen hebben we dus een veel betere kans om onze onderzoekshypothese te bevestigen. Dit uitgebreide ontwerp van het onderzoek zal meer kosten (voor de onderzoekers en de kinderen en hun ouders), maar levert vermoedelijk ook veel meer op: een valide conclusie met grote maatschappelijke impact.

Het ontwerp van een onderzoek (zie Hoofdstuk **??**) moet zo efficiënt mogelijk zijn, en de onderzoeker moet daarover al in een vroeg stadium nadenken. De efficiëntie hangt ten eerste af van keuzes over hoe de onafhankelijke variabelen worden gevarieerd. Is er een aparte groep proefpersonen voor iedere conditie van de onafhankelijke variabele (condities zijn "between subjects", zoals in voorbeeld 3.2 hierboven? Bij een between-subjects ontwerp met twee groepen zijn er ca $n = (5.6/d)^2$ nodig in elke groep (Gelman and Hill, 2007) (zie §**??**). Of doen alle proefpersonen mee aan alle condities (condities zijn "within subjects")? Bij een within-subjects ontwerp met twee condities zijn er dan slechts $n = (2.8/d)^2$

proefpersonen nodig in elke conditie, en het onderzoek heeft dan dus minder directe en indirecte kosten voor veel minder proefpersonen. In het algemeen is het daarom beter om indien mogelijk, onafhankelijke variabelen te variëren binnen proefpersonen, en niet tussen proefpersonen. Toch is dat niet altijd mogelijk, ten eerste omdat individuele kenmerken nu eenmaal alleen verschillen tussen proefpersonen (denk aan: mannelijk/vrouwelijk geslacht, wel/niet meertalige jeugd, wel/niet afasie, enz.). Ten tweede moeten we terdege rekening houden met effecten van 'transfer' tussen condities, die de validiteit bedreigen (denk aan: ervaring, leren, vermoeidheid, rijping). We keren hierop terug in §5.3.

Meertaligheid en geslacht zijn kenmerken die alleen tussen personen kunnen variëren. Maar andere condities kunnen ook variëren binnen personen, bijvoorbeeld de dag waarop een cognitieve meting wordt afgenomen. Stel dat we een verschil verwachten van $D = 2$ punten tussen cognitieve metingen afgenomen op maandag of op vrijdag (met $s = 4$ en $d = 0.5$, zie voorbeeld 3.2. Als we de dag van de meting variëren tussen proefpersonen, en dus aparte groepen maken voor de maandag-kinderen en de vrijdag-kinderen, dan zijn er $n = (5.6/0.5)^2 = 126$ kinderen nodig in iedere groep, dus $N = 252$ kinderen in totaal. Als we de dag van de meting echter variëren binnen proefpersonen, en iedere proefpersoon dus observeren zowel op maandag als op vrijdag, dan zijn er in totaal slechts $N = (2.8/0.5)^2 = 32$ kinderen nodig. Met het within-subjects ontwerp hoeven we dus veel minder kinderen lastig te vallen met onze cognitieve meting. Wel moeten we terdege rekening houden met leereffecten tussen de eerste en de tweede meting, en daarvoor gepaste maatregelen treffen. We kunnen bijvoorbeeld niet meer dezelfde vragenlijsten afnemen in beide condities.

De efficiëntie van een onderzoek hangt ook af van de afhankelijke variabele, en met name van het meetniveau (Hoofdstuk **??**), de nauwkeurigheid, en de betrouwbaarheid van de observaties (Hoofdstuk **??**). Hoe lager het meetniveau, des te lager ook de efficiëntie van het onderzoek. En hoe lager de nauwkeurigheid, des te lager ook de efficiëntie van het onderzoek, en des te meer proefpersonen en observaties zijn er nodig om valide conclusies te kunnen trekken.

Stel dat we een verschil willen onderzoeken tussen twee condities binnen proefpersonen, en stel dat het verschil in werkelijkheid 2 punten bedraagt (met $s_D = 4$ en $d = 0.5$, zie voorbeeld 3.2). We kijken nu echter niet naar de richting en de grootte van het verschil, maar alleen naar de *richting* van het verschil tussen de twee observaties per proefpersoon: heeft die proefpersoon een positief of een negatief verschil tussen de eerste en de tweede conditie? Deze binomiale afhankelijke variabele bevat minder informatie dan de oorspronkelijke puntenscore (nl. alleen de richting, en niet de grootte van het verschil), en het onderzoek is daardoor dus minder efficiënt. We hebben daarom in dit specifieke voorbeeld niet 34 maar tenminste 59 proefpersonen nodig.

Onderzoekers zijn dus verantwoordelijk om de kosten en baten van hun onderzoek zorgvuldig en eerlijk af te wegen en te beoordelen, en zij dienen te beschikken over voldoende methodologische bagage om een goed onderzoekson-

twerp (design) te kiezen gezien het tijdsbestek, de mogelijk beschikbare proefpersonen, de meetinstrumenten, enz.

3.3 Proefpersonen en informanten

Wetenschappelijk onderzoek is mensenwerk: onderzoekers zijn ook mensen. Op het gebied van de geesteswetenschappen bestuderen die onderzoekers weer het gedrag en de geestelijke producten van (andere) mensen. Daarvoor gelden wetten, regels, richtlijnen en gedragscodes waaraan onderzoekers (en studenten!) zich dienen te houden, vanuit de eerder genoemde principes van zorgvuldigheid en verantwoordelijkheid. Het onderzoek zelf, en de verzamelde gegevens, mogen geen schade of groot verlies van privacy opleveren voor de deelnemers.

Voor geesteswetenschappelijk onderzoek zijn twee wetten relevant:

- Algemene Verordening Gegevensbescherming (AVG),
  zie https://autoriteitpersoonsgegevens.nl/nl/onderwerpen/avg-europese-privacywetgeving
- Wet Medisch-wetenschappelijk Onderzoek met mensen (WMO),
  zie http://www.wetten.nl

Het is verplicht om proefpersonen (of hun wettelijke vertegenwoordigers) te vragen om expliciete “informed consent”. Dat houdt in dat de proefpersonen eerlijk geïnformeerd worden over het onderzoek, over de baten en kosten daarvan, en over hun beloning, en dat zij daarna (d.i. “informed”) expliciet toestemmen in hun deelname (“consent”). Voorbeelden van informed consent (informatiebrieven en toestemmingsverklaringen) zijn te vinden op de website van de Facultaire Ethische Toetsingscommissie (FETC, hieronder nader besproken), via https://fetc-gw.wp.hum.uu.nl/.

Alle gegevens waaruit een individuele persoon te herleiden is, worden beschouwd als “persoonsgegevens”, en deze persoonsgegevens mogen alleen worden verzameld en verwerkt conform de AVG. Het is raadzaam om de onderzoeksgegevens zo snel mogelijk los te koppelen van de persoonsgegevens, d.w.z. dat je de gegevens anonimiseert. De koppeling tussen persoonsgegevens en en onderzoeksgegevens (bijv. een lijst met namen van proefpersonen en hun bijbehorende anonieme persoonlijke code) is zelf weer vertrouwelijke informatie die je zorgvuldig moet bewaren en opslaan. Bewaar de persoonsgegevens niet langer dan nodig. De onderzoeksgegevens mag je alleen gebruiken voor het (wetenschappelijke) doel waarmee ze zijn verzameld. Zorg ook dat de proefpersonen niet herkenbaar zijn (gebruik anonieme codes) in verslagen en publicaties over het onderzoek.

Foto’s en opnames van personen (audio, video, fysiologische gegevens, EEG) vallen onder het zgn. portretrecht. Foto’s en andere identificerende

opnames worden dus als portretten beschouwd. Bij publicatie kan de afgebeelde/weergegeven persoon zich beroepen op het portretrecht, en een schadevergoeding eisen voor het letsel dat hem of haar door die publicatie wordt aangedaan. Als je een herkenbare opname zou willen publiceren, dan moet je dus vooraf expliciete toestemming daarvoor vragen van de opgenomen persoon of zijn wettelijke vertegenwoordiger (zie het bovengenoemde voorbeeld van "informed consent"). Dat geldt ook als je een fragment van zo'n opname laat zien of horen tijdens een conferentie of op een website.

In de wet WMO is vastgelegd dat onderzoek met mensen eerst moet worden goedgekeurd door een speciale commissie; voor de Faculteit Geesteswetenschappen van de Universiteit Utrecht is dat de Medisch-Ethische Toetsingscommissie die valt onder het Universitair Medisch Centrum Utrecht (METC). Die commissie weegt af of de mogelijke baten van het onderzoek redelijkerwijs opwegen tegen de kosten en mogelijke schade voor de proefpersonen.

Het meeste onderzoek op het gebied van talen en communicatie bij de Universiteit Utrecht is vrijgesteld van de tijdrovende toetsing door de METC, maar moet wel verplicht worden voorgelegd aan de **Facultaire Ethische Toetsingscommissie** (FETC), en wel aan de kamer Linguïstiek daarvan. Dat geldt echter niet voor onderzoek door studenten! Op de website van de FETC is meer informatie te vinden: https://fetc-gw.wp.hum.uu.nl/. Overleg bij twijfel altijd met je begeleider of docent. Ethische toetsing is ook verplicht voor studenten en onderzoekers uit andere domeinen (literatuur, geschiedenis, media & cultuur) die van plan zijn onderzoek te verrichten met mensen.

3.4 Gegevens

De verzamelde data of gegevens vormen de onderbouwing voor de conclusies uit wetenschappelijk onderzoek. Die gegevens zijn daarmee van essentieel belang: zonder gegevens geen valide conclusies. Zoals we hierboven zagen (§3.2) zijn die gegevens bovendien zeer kostbaar (in tijd, geld, privacy, enz). We moeten er dus uiterst zorgvuldig mee omgaan. We moeten anderen kunnen overtuigen van de validiteit van onze conclusies op basis van die gegevens, en we moeten de onderliggende gegevens desgevraagd kunnen delen met andere onderzoekers.

Die zorgvuldigheid vereist dus in ieder geval dat we zo snel mogelijk voldoende reservekopieën maken, en die bewaren op verschillende veilige plaatsen. Bedenk eens wat er zou gebeuren als een brand of overstroming je werkplek of woning zou vernietigen, of als tijdens je scriptie-project je laptop wordt gestolen (waar gebeurd!). Heb je dan goede en recente kopieën van de gegevens elders opgeslagen? Voor kopieën en 'backups' kun je goed gebruik maken van een afdoende beveiligde "cloud service"[2].

[2]Medewerkers van de UU kunnen SurfDrive (https://www.surfdrive.nl) gebruiken om gegevens veilig en makkelijk te bewaren op een beveiligde netwerk-schijf.

De zorgvuldigheid vereist ook dat we goed bijhouden wat de gegevens voorstellen, en hoe ze zijn verzameld. Gegevens zonder bijbehorende beschrijving zijn nagenoeg waardeloos voor wetenschappelijk onderzoek. Charles Darwin noteerde nauwkeurig welke vogel op welk van de Galapagos-eilanden welke vorm van snavel had, en deze observaties vormden later (deel van) de onderbouwing van zijn evolutie-theorie. Houd dus een logboek bij (op papier of digitaal) waarin je alle stappen van je onderzoek beschrijft, en eventueel motiveert. Noteer ook merk en type en instellingen van de gebruikte apparatuur, en noteer versie-nummer en instellingen van de gebruikte software. Houd bij welke bewerkingen je op de gegevens hebt toegepast, en waarom, en in welk bestand welke gegevens zijn opgeslagen.

Als je werkt met geditigaliseerde data (bv in Excel of SPSS of R), houd dan ook zorgvuldig bij welke variabelen in welke kolom is opgeslagen, in welke eenheden, en met welke codes.

Voorbeeld 3.3: Het bestand http://tinyurl.com/nj4pjaq bevat gegevens van 80 sprekers van het Nederlands, ten dele ontleend aan het Corpus Gesproken Nederlands (CGN). De eerste regel bevat de namen van de variabelen. Iedere volgende regel correspondeert met één spreker. De gegevens op iedere regel zijn gescheiden door spaties. De eerste kolom bevat de anonieme identificatie-code van de spreker volgens het CGN. In de vijfde kolom staat de regio van herkomst van de spreker gecodeerd, als één letterteken, met de codes `W` Randstad, `M` Midden-Nederland, `N` Noord-Nederland, `S` Zuid-Nederland) (Quené, 2008). Door de zorgvuldige annotatie zijn deze gegevens nog goed bruikbaar, ook al zijn ze ruim 20 jaar geleden verzameld door collega-onderzoekers.

Gegevens blijven het intellectuele eigendom van degene die ze heeft verzameld. Gebruik van andermans data zonder bronvermelding kan beschouwd worden als diefstal, of als plagiaat.

Fraude met gegevens (gegevens fabriceren of verzinnen, in plaats van observeren) is uiteraard strijdig met meerdere principes uit de bovengenoemde gedragscode (VSNU, 2018). Fraude schaadt het wederzijds vertrouwen waarop wetenschap is gebaseerd. Het misleidt andere onderzoekers die voortbouwen op de fictieve resultaten, en onderzoeksgeld voor die frauduleuze onderzoekslijn wordt weggezogen uit ander, niet frauduleus onderzoek — kortom, een wetenschappelijke doodzonde. Als je wilt overleggen over vragen of dilemma's hierover, neem dan contact op met prof.dr. Josine Blok, vertrouwenspersoon wetenschappelijke integriteit van de Faculteit Geesteswetenschappen (`j.h.blok@uu.nl`).

3.5 Teksten

Wetenschappelijk onderzoek wordt pas echt nuttig, als de resultaten ervan verspreid worden. Onderzoek dat niet wordt gerapporteerd, zou net zo goed *niet* kunnen zijn uitgevoerd, en de kosten van dat onderzoek zijn dan feitelijk tevergeefs geweest. Een belangrijk deel van het wetenschappelijk werk bestaat daarom uit verslaglegging ervan. Publicaties (en octrooien) vormen een zeer belangrijk deel van de "output" van wetenschappelijk onderzoek. Onderzoekers worden gemeten naar het aantal publicaties, en naar de "impact" daarvan (het aantal malen dat die publicaties weer geciteerd worden door anderen die erop voortbouwen). Mede gezien de grote belangen dienen we dus zorgvuldig om te gaan met teksten van anderen en van onszelf.

De onderzoekers die betrokken zijn bij een onderzoek, moeten met elkaar overleggen wie de auteurs van het verslag of van de publicatie zullen zijn, en in welke volgorde. Mede-auteurs van een wetenschappelijk verslag moeten voldoen aan drie voorwaarden (Office of Research Integrity, 2012, Ch.10). Ten eerste moeten zij een substantiële wetenschappelijke bijdrage hebben geleverd aan één of meer fasen in het onderzoek: het oorspronkelijke idee bedenken, het onderzoek opzetten en ontwerpen, de gegevens verzamelen, en de gegevens analyseren en interpreteren. Ten tweede moeten ze hebben meegewerkt aan het verslag, als schrijver en/of als commentator. Ten derde moeten ze instemmen met de definitieve tekst van het verslag (meestal impliciet, soms expliciet), en tevens instemmen met hun mede-auteurschap daarvan. De auteurs doen er goed aan om af te spreken in welke volgorde hun namen vermeld worden. Meestal correspondeert die volgorde met het afnemend belang en de afnemende omvang van de respectievelijke bijdragen van de auteurs. Als de eindverantwoordelijke hoofd-onderzoeker tevens mede-auteur is, dan wordt deze vaak als laatste genoemd.

Voorbeeld 3.4: Student-assistent A heeft geholpen bij het verzamelen van de gegevens, maar deze assistent heeft geen andere bijdragen geleverd, en weet niet goed waar het onderzoek eigenlijk over gaat. A hoeft geen mede-auteur te worden van het verslag, maar de auteurs dienen de bijdrage van A wel te beschrijven en te erkennen in hun verslag.

Student B heeft één van de delen van een onderzoeksproject uitgevoerd onder begeleiding van onderzoeker C. Deze begeleider C heeft het hele project bedacht, maar B heeft literatuur verzameld, een deelonderzoek opgezet en uitgevoerd, data verzameld, geanalyseerd en geïnterpreteerd, en daarvan verslag gedaan in een werkstuk. Student B en begeleider C zijn daarom beiden mede-auteurs van een

> publicatie over B's deel van het onderzoeksproject. Zij spreken af in welke volgorde de auteurs genoemd worden. Omdat student B het belangrijkste was voor dit werk, en C de eindverantwoordelijke was, spreken zij af dat B de eerste auteur wordt en C de tweede en laatste.

Onderzoekers bouwen voort op het werk van hun voorgangers (zie voorbeeld 3.1). Dat kan ook gelden voor hun redeneringen, en zelfs hun teksten, maar daarbij moeten we dan altijd correct verwijzen naar de juiste bron, d.w.z. naar het werk van die voorgangers. Anders is immers niet meer te onderscheiden wie verantwoordelijk is voor welke gedachte of tekstfragment. Plagiaat is "het overnemen van stukken, gedachten, redeneringen van anderen en deze laten doorgaan voor eigen werk" (*Van Dale*, 12e druk). Ook deze vorm van fraude is een wetenschappelijke doodzonde, waar krachtige sancties op kunnen volgen. De Faculteit Geesteswetenschappen van de UU zegt daarover het volgende:

> "Van plagiaat is sprake bij het in een scriptie of ander werkstuk gegevens of tekstgedeelten van anderen overnemen zonder bronvermelding. Onder plagiaat valt onder meer:
>
> - het knippen en plakken van tekst van digitale bronnen zoals encyclopedieën en digitale tijdschriften zonder aanhalingstekens en verwijzing;
> - het knippen en plakken van teksten van het internet zonder aanhalingstekens en verwijzing;
> - het overnemen van gedrukt materiaal zoals boeken, tijdschriften en encyclopedieën zonder aanhalingstekens en verwijzing;
> - het opnemen van een vertaling van bovengenoemde teksten zonder aanhalingstekens en verwijzing;
> - het parafraseren van bovengenoemde teksten zonder (deugdelijke) verwijzing: parafrasen moeten als zodanig gemarkeerd zijn (door de tekst uitdrukkelijk te verbinden met de oorspronkelijke auteur in tekst of noot), zodat niet de indruk wordt gewekt dat het gaat om eigen gedachtegoed van de student;
> - het overnemen van beeld-, geluids- of testmateriaal van anderen zonder verwijzing en zodoende laten doorgaan voor eigen werk;
> - het zonder bronvermelding opnieuw inleveren van eerder door de student gemaakt eigen werk en dit laten doorgaan voor in het kader van de cursus vervaardigd oorspronkelijk werk, tenzij dit in de cursus of door de docent uitdrukkelijk is toegestaan;

- het overnemen van werk van andere studenten en dit laten doorgaan voor eigen werk. Indien dit gebeurt met toestemming van de andere student is de laatste medeplichtig aan plagiaat;
- ook wanneer in een gezamenlijk werkstuk door een van de auteurs plagiaat wordt gepleegd, zijn de andere auteurs medeplichtig aan plagiaat, indien zij hadden kunnen of moeten weten dat de ander plagiaat pleegde;
- het indienen van werkstukken die verworven zijn van een commerciële instelling (zoals een internetsite met uittreksels of papers) of die tegen betaling door iemand anders zijn geschreven."

http://students.uu.nl/praktische-zaken/regelingen-en-procedures/fraude-en-plagiaat

Bij plagiaat van eigen werk worden de teksten of gedachten niet overgenomen van anderen maar van één van de auteurs. Over dit zelf-plagiaat wordt verschillend gedacht; het is echter raadzaam om in voorkomende gevallen wel de bron te vermelden, vanuit de principes van zorgvuldigheid, betrouwbaarheid, controleerbaarheid, en verantwoordelijkheid.

Een verwijzing of citatie of referentie is een verkorte bronvermelding in de tekst; in dit boek ben je er al vele tegengekomen. Aan het einde van de tekst volgt dan de volledige lijst van bronnen, meestal aangeduid als bronvermeldingen, geraadpleegde bronnen, referenties, literatuur, of bibliografie ("boekbeschrijving"). Een foutieve bronvermelding kan worden beschouwd als plagiaat (Universiteitsbibliotheek, Vrije Universiteit Amsterdam, 2015) omdat de lezer niet verwezen wordt naar de juiste bron. Onderzoekers dienen hun bronnen daarom op correcte wijze te vermelden. Daarvoor bestaan verschillende conventies, afhankelijk van het vakgebied. Meestal zal een docent aangeven volgens welke stijl of conventie je je bronnen moet vermelden. In dit boek volgen we zoveel mogelijk de stijl van de (American Psychological Association, 2010), die gebruikelijk is in de sociale wetenschappen en een deel van de geesteswetenschappen.

De regels voor bronvermelding zijn soms ingewikkeld. Bovendien moeten de auteurs zorgen dat de citaties in de tekst overeenkomen met de lijst van referenties. Deze taken kunnen beter worden bijgehouden door een zgn."reference manager", een programma dat referenties of bronvermeldingen verzamelt en op de juiste wijze invoegt in de tekst. Een overzicht van zulke programma's is te vinden via https://en.wikipedia.org/wiki/Comparison_of_reference_management_software. Voor dit tekstboek is gebruik gemaakt van Zotero, gecombineerd met BibTeX.

Chapter 4

Meetniveau

4.1 Inleiding

In Hoofdstuk 2 maakten we al kennis met variabelen: eigenschappen die verschillende waarden kunnen aannemen. De waarde van een variabele is dus een aanduiding van een eigenschap, of kwaliteit, of hoedanigheid, van een object of persoon. Als het gaat om een afhankelijke variabele, dan wordt die waarde ook aangeduid als *score* of *responsie*, vaak aangeduid met symbool Y. De wijze waarop een kenmerk wordt uitgedrukt in een (gemeten) waarde, noemen we het *meetniveau* van de variabele; het meetniveau is dus een eigenschap of kenmerk van de variabele zelf! We onderscheiden vier meetniveau's, in toenemende niveau's van informativiteit: nominaal, ordinaal, interval, ratio. Bij de eerste twee meetniveau's worden alleen discrete categorieën onderscheiden, zonder of met ordening. Bij de laatste twee meetniveau's worden getalswaarden gebruikt, zonder of met nulpunt. We zullen de meetniveau's hieronder nader bespreken. Inzicht in het meetniveau van een variabele is van belang voor de interpretatie van de scores op een variabele en — zoals we later zullen zien — voor de keuze van de juiste statistische toets om een onderzoeksvraag te beantwoorden.

4.2 Nominaal

We spreken van een nominale variabele (of nominaal meetniveau) als een kenmerk gecategoriseerd wordt in afzonderlijke (discrete) categorieën, waarbij er *niet* een ordening is tussen de categorieën. Bekende voorbeelden zijn o.a. de nationaliteit van een proefpersoon, het merk van een auto, de kleur van iemands ogen, de smaak van een bak schepijs, je woonsituatie (bij gezin van herkomst, op kamers, zelfstandig, samenwonend, anders), enz. De scores kunnen alleen gebruikt worden om de categorieën te onderscheiden (de uitspraak "vanille is

anders dan aardbei" is wel zinnig). We kunnen wel tellen hoe vaak iedere categorie voorkomt, maar er is geen interpreteerbare rangorde (de uitspraak "vanille is groter dan aardbei" is onzinnig), en we kunnen ook niet rekenen met de gemeten waarden van een nominale variabele. We kunnen dus wel de meest voorkomende nationaliteit vaststellen, maar we kunnen niet de gemiddelde nationaliteit uitrekenen.

4.3 Ordinaal

Er is sprake van een ordinale variabele (of van een ordinaal meetniveau) als een kenmerk gecategoriseerd wordt in afzonderlijke categorieën, waarbij er *wel* een rangorde is tussen de categorieën. Bij een ordinale variabele weten we echter niets over de afstand tussen de verschillende categorieën. Bekende voorbeelden zijn o.a. schooltype (VMBO, HAVO, VWO, …), antwoord op een schaalvraag (*mee eens, neutraal, niet mee eens*), positie op een ranglijst, volgorde van afvallen bij een talentenjacht, kledingmaat (XS, S, M, L, XL, …), of militaire rang (soldaat, majoor, generaal, …). Ook hier kunnen we wel tellen hoe vaak iedere categorie voorkomt, en we kunnen ook de rangorde zinnig interpreteren (wie als laatste afvalt presteert beter dan wie als eerste afvalt, maat L is groter dan M, een generaal is de baas van een majoor). We kunnen echter niet rekenen met de gemeten waarden van een ordinale variabele. We kunnen wel de meest verkochte kledingmaat vaststellen, maar we kunnen niet de gemiddelde verkochte kledingmaat uitrekenen[1].

4.4 Interval

Er is sprake van een interval-variabele (of van een interval-meetniveau) als een kenmerk uitgedrukt wordt in een getal op een continue schaal, waarbij deze schaal *niet* een nulpunt heeft. Door de schaal weten we bij een interval-variabele ook wat de afstanden of intervallen zijn tussen de verschillende waarden. Bekende voorbeelden zijn o.a. temperatuur in graden Celcius (het nulpunt is arbitrair), of jaartal (idem). We kunnen tellen hoe vaak iedere categorie voorkomt, we kunnen de rangorde zinnig interpreteren (het jaar 1999 in onze gregoriaanse kalender ging vooraf aan het jaar 2000), en we kunnen ook de intervallen zinnig interpreteren (van 1918 tot 1939 is net zo lang als van 1989 tot 2010). We kunnen wel rekenen met de waarden van een interval-variabele, maar de enige zinnige bewerkingen zijn optellen en aftrekken. Daarmee kunnen we wel een gemiddelde berekenen, bijv. het gemiddelde jaar waarin de personen in een steekproef hun eerste mobiele telefoon begonnen te gebruiken.

[1] Als de helft van de respondenten *mee eens* antwoordt, en de andere helft *niet mee eens*, dan kunnen we niet zinnig concluderen dat de responsies gemiddeld *neutraal* zouden zijn.

4.5 Ratio

Het vierde en hoogste meetniveau is het ratio-niveau. Er is sprake van een ratio-variabele (of van een ratio-meetniveau) als een kenmerk uitgedrukt wordt in een getal op een continue schaal, waarbij deze schaal *wel* een nulpunt heeft. Door de schaal weten we bij een ratio-variabele wat de afstanden of intervallen zijn tussen de verschillende waarden. Bovendien weten we door het nulpunt wat de verhoudingen of ratio's zijn tussen de verschillende waarden. Bekende voorbeelden zijn o.a. temperatuur in graden Kelvin (vanaf het absolute nulpunt), de responsietijd[2] in duizendsten van een seconde (ms), je lengte in cm, je leeftijd in jaren, het aantal gemaakte fouten in een toets, enz. Bij een ratio-variabele kunnen we tellen hoe vaak iedere categorie voorkomt, we kunnen de rangorde zinnig interpreteren (iemand van 180 cm is langer dan iemand van 179 cm), we kunnen intervallen zinnig interpreteren (de toename in leeftijd van 12 naar 18 is tweemaal zo groot als de toename van 9 naar 12), en we kunnen ook verhoudingen tussen de waarden zelf zinnig interpreteren (een leeftijd van 24 is *tweemaal* zo oud als een leeftijd van 12). We kunnen rekenen met de waarden van een interval-variabele, en daarbij kunnen we niet alleen optellen en aftrekken maar ook delen en vermenigvuldigen. Ook hier is het mogelijk om een gemiddelde te berekenen, bijv. de gemiddelde leeftijd waarop de personen in een steekproef hun eerste mobiele telefoon begonnen te gebruiken.

4.6 Ordening van meetniveaus

De meetniveaus zijn hierboven besproken in toenemende informativiteit of sterkte. Een nominale variabele bevat het minste informatie en geldt als het laagste meetniveau, en een ratio-variabele bevat het meeste informatie en geldt als het hoogste meetniveau.

Het is altijd mogelijk om gegevens gemeten op een hoger meetniveau te interpreteren alsof ze op een lager niveau zijn gemeten. Als we bijvoorbeeld het maandinkomen van de personen in een steekproef hebben gemeten op ratio-niveau (in €), dan kunnen we daar probleemloos een ordinale variabele van maken (*minder dan modaal, van modaal tot tweemaal modaal, meer dan tweemaal modaal*). We gooien daarbij informatie weg: de oorspronkelijke meting in € bevat meer informatie dan de daaruit afgeleide classificatie in drie geordende categorieën.

Natuurlijk is het omgekeerde niet mogelijk: een variabele van een laag meetniveau kunnen we niet interpreteren op een hoger niveau. We zouden dan informatie achteraf moeten toevoegen die we niet hebben verzameld bij de oorspronkelijke meting van die variabele. Het is dus zaak om de relevante variabelen te meten of te observeren op het juiste meetniveau. Stel je voor dat we de

[2]Het nulpunt is het moment van de gebeurtenis waarop de proefpersoon moet reageren.

lichaamslengte van volwassen mannen en vrouwen willen vergelijken. Als we de lichaamslengte meten op ordinaal meetniveau (met drie categorieën *kort*, *middelmatig* en *lang* gelijkelijk gedefinieerd voor alle personen), dan kunnen we dus niet de gemiddelde lichaamslengte uitrekenen, en we kunnen ook niet een statistische toets gebruiken die refereert aan het gemiddelde van de lichaamslengte. Dat hoeft geen probleem te zijn, maar het is wel goed om vooraf te doordenken wat de consequenties zijn van de keuze voor een bepaald meetniveau.

Chapter 5

Validiteit

5.1 Inleiding

Experimenteel onderzoek heeft tot doel om hypotheses te toetsen. Ook in ander, niet-experimenteel onderzoek kunnen hypotheses worden getoetst, maar we beperken ons hier voor de helderheid tot experimenteel onderzoek, d.w.z. onderzoek waarin het experiment als methode wordt gebruikt. In experimenteel onderzoek wordt getracht causale verbanden aannemelijk te maken. Als de resultaten van een experimenteel onderzoek de onderzoekshypothese bevestigen (d.w.z. de nulhypothese wordt verworpen), dan is het aannemelijk dat een verandering in de onafhankelijke variabele de oorzaak (Latijn: *causa*) is voor een verandering of *effect* in de afhankelijke variabele. Zo kunnen we na experimenteel onderzoek met enige zekerheid concluderen, bijvoorbeeld, dat een verschil in behandeling na een herseninfarct de oorzaak is, of een belangrijke oorzaak is, van een verschil in taalvaardigheid van een patiënt zoals geobserveerd 6 maanden na een herseninfarct. Het experiment heeft aannemelijk gemaakt dat er een causaal of oorzakelijk verband is tussen de behandelingsmethode (onafhankelijke variabele) en de resulterende taalvaardigheid (afhankelijke variabele).

5.2 Causaliteit

Een causaal of oorzakelijk verband tussen twee variabelen is iets anders dan een 'gewoon' verband of samenhang tussen twee variabelen. Als twee verschijnselen met elkaar samenhangen, hoeft het ene niet de oorzaak van het andere te zijn. Een eerste voorbeeld zien we bij de samenhang tussen de lengte van personen en hun gewicht: lange mensen zijn over het algemeen zwaarder dan korte mensen (en omgekeerd: korte mensen zijn over het algemeen lichter

dan lange mensen). Is er nu sprake van een causale relatie tussen lengte en gewicht? Wordt het ene kenmerk (deels) veroorzaakt door het andere? Nee, in dit voorbeeld is er wel samenhang maar geen causaal verband tussen de kenmerken: zowel lengte als gewicht worden "veroorzaakt" door andere variabelen, o.a. genetische eigenschappen en voedingspatronen. Een tweede voorbeeld is de samenhang tussen motivatie en taalvaardigheid van iemand die een vreemde taal leert: hoog gemotiveerde studenten leren een nieuwe vreemde taal beter en vlotter dan laag gemotiveerde studenten, maar ook hier is niet duidelijk wat de oorzaak en wat het gevolg is.

Een causaal verband is een speciale vorm van samenhang. Een causaal verband is een verband tussen twee twee verschijnselen of kenmerken, waarbij bovendien voldaan moet zijn aan een aantal extra voorwaarden (Shadish et al., 2002). Ten eerste moet de oorzaak aan het gevolg vooraf gaan (na behandeling treedt verbetering op). Ten tweede moet het gevolg niet optreden als de oorzaak niet aanwezig was (zonder behandeling geen verbetering). Bovendien moet het gevolg — althans in theorie — altijd optreden als de oorzaak aanwezig is (behandeling resulteert altijd in verbetering). Ten derde kunnen we geen andere plausibele verklaring vinden voor het optreden van het gevolg, behalve de mogelijke oorzaak. Als we het causale mechanisme kennen (we snappen waarom behandeling de oorzaak is van verbetering), dan zijn we beter in staat om mogelijke andere plausibele verklaringen uit te sluiten. Helaas is dat bij de gedragswetenschappen, inclusief de taalwetenschap, echter zelden het geval. We constateren wel dat een behandeling resulteert in verbetering, maar de theorie die oorzaak (behandeling) en gevolg (verbetering) verbindt is zelden compleet en vertoont belangrijke lacunes. Dat betekent dat we goede voorzorgen moeten treffen in onze onderzoeksmethoden, teneinde mogelijke alternatieve plausibele verklaringen voor de gevonden effecten uit te sluiten.

5.3 Validiteit

Een bewering of conclusie is *valide* als de bewering *waar* (true) en *gerechtvaardigd* (justified) is. Een ware uitspraak correspondeert met de werkelijkheid: de bewering *ieder kind leert ten minste een taal* is waar, omdat de bewering de werkelijkheid goed weergeeft. Een gerechtvaardigde bewering ontleent geldigheid aan de empirische evidentie waarop die bewering is gebaseerd: ieder kind dat wij hebben geobserveerd, of dat anderen hebben geobserveerd, leert een taal of heeft een taal geleerd (behalve bijzondere gevallen voor wie een aparte verklaring nodig is). De rechtvaardiging van een bewering is sterker naarmate de methode van (directe of indirecte) observatie sterker is en meer zekerheid biedt. Dit houdt ook in dat de validiteit van een bewering niet een categoriale eigenschap is (wel/niet valide) maar een gradueel kenmerk: een bewering kan meer of minder valide zijn.

Aan de validiteit van een bewering kunnen drie verschillende aspecten worden

onderscheiden.

1. In hoeverre zijn de conclusies over de relaties tussen de afhankelijke en de onafhankelijke variabele geldig? Deze vraag heeft betrekking op de *interne validiteit*.
2. In hoeverre zijn de uitwerkingen, operationaliseringen, van de afhankelijke en onafhankelijke variabele adequaat? Deze vraag heeft betrekking op de *constructvaliditeit*.
3. In hoeverre kunnen de conclusies gegeneraliseerd worden naar andere proefpersonen, stimuli, condities, situaties, observaties? Deze vraag heeft betrekking op de *externe validiteit*.

Deze drie vormen van validiteit zullen wij in de navolgende paragrafen toelichten.

5.4 Interne validiteit

Het is vanzelfsprekend de bedoeling om in een experimenteel onderzoek zoveel mogelijk alternatieve verklaringen voor de onderzoeksresultaten uit te sluiten. Er moet immers aangetoond worden dat er een causaal verband is tussen twee variabelen X en Y, en daarbij moeten storende factoren zoveel mogelijk onder controle gehouden worden. Laten we eens kijken naar voorbeeld 5.1 hieronder.

Voorbeeld 5.1: (Verhoeven et al., 2004) onderzochten o.a. de hypothese dat ouderen (boven de 45 jaar) langzamer spreken dan jongeren (onder de 40 jaar). Om dat te onderzoeken werd spraak opgenomen van 160 sprekers, gelijk verdeeld over de twee leeftijdsgroepen, in een interview van ongeveer 15 minuten. Na fonetische analyse van de articulatiesnelheid blijkt dat de "jongeren" relatief snel spreken met 4.78 lettergrepen per seconde, en de "ouderen" aanzienlijk langzamer, met 4.52 lettergrepen per seconde (Verhoeven et al., 2004, p.302). We concluderen dat de hogere leeftijd de *oorzaak* is van het lagere spreektempo bij de oudere sprekers — maar is die conclusie terecht?

Deze vraag naar de rechtvaardiging van de conclusie is een vraag naar de interne validiteit van het onderzoek. De interne validiteit heeft betrekking op de relaties tussen gemeten of gemanipuleerde variabelen, en is onafhankelijk van

de (theoretische) constructen die de verschillende variabelen representeren (vandaar de term 'interne validiteit'). Of, anders gezegd: de vraag naar de interne validiteit is een vraag naar mogelijke alternatieve verklaringen voor de gevonden onderzoeksresultaten. Veel van de mogelijke alternatieve verklaringen kunnen worden ondervangen door de manier waarop de gegevens worden verzameld. We bespreken hieronder de meest in het oog lopende bedreigingen van de interne validiteit (Shadish et al., 2002).

1. **Geschiedenis** is een bedreiging van de interne validiteit. Het begrip 'geschiedenis' omvat o.a. gebeurtenissen die hebben plaatsgevonden tussen of tijdens een voormeting en een nameting; het gaat dan om gebeurtenissen die geen deel uitmaken van de experimentele manipulatie (de onafhankelijke variabele), maar die wel van invloed zouden kunnen zijn op de afhankelijke variabele. Een hittegolf, bijvoorbeeld, kan van invloed zijn op het gedrag van de proefpersonen tijdens een onderzoek.

In een laboratorium wordt de 'geschiedenis' onder controle gehouden door de proefpersonen af te sluiten van invloeden van buitenaf (zoals een hittegolf), of door afhankelijke variabelen te kiezen die nauwelijks beïnvloed kunnen worden door externe factoren. In onderzoek buiten het laboratorium, waaronder veldonderzoek, is het veel lastiger en vaak zelfs onmogelijk om invloeden van buitenaf onder controle te houden. In het volgende voorbeeld wordt dit duidelijk.

Voorbeeld 5.2: In een onderzoek worden twee methoden vergeleken om leerlingen een vreemde taal te leren spreken, i.c. Nieuw-Grieks. De eerste groep moet Griekse woordjes en grammatica leren in een klaslokaal, gedurende enkele weken. De tweede groep gaat in diezelfde periode op een studiereis naar Griekenland, waar leerlingen moeten converseren in de doeltaal. De totale tijd besteed aan het taalvaardigheidsonderwijs is voor beide groepen gelijk. Na afloop blijkt de taalvaardigheid van de tweede groep groter dan die van de eerste groep. Wordt dat verschil in de afhankelijke variabele inderdaad veroorzaakt door de lesmethode (onafhankelijke variabele)?

---

2. **Rijping** is de natuurlijke veroudering of rijping van proefpersonen tijdens een onderzoek. Als de proefpersonen gedurende een onderzoek ouder worden, zich ontwikkelen, meer ervaren of sterker worden, èn als deze rijping niet is opgenomen in de onderzoeksvraag, dan vormt rijping een bedreiging van de interne validiteit. In experimenten waarin reactietijden worden gemeten, bijvoorbeeld, zien we meestal dat de reactietijden

van een proefpersoon sneller worden gedurende het experiment, als gevolg van training en oefening. We kunnen de interne validiteit dan beschermen tegen dit leer-effect, door de stimuli voor iedere proefpersoon in een andere willekeurige volgorde aan te bieden.

Meestal is er sprake van rijping doordat de proefpersonen vele malen achtereen dezelfde taak uitvoeren of dezelfde vragen beantwoorden. Rijping kan echter ook optreden wanneer proefpersonen hun antwoorden kenbaar moeten maken op een juist niet gebruikelijke manier, bv. door een ongewone vraagstelling, of in een ongebruikelijke vorm van meerkeuze-vragen. Bij de eerste paar keer dat een proefpersoon dan vragen beantwoordt, kan de wijze van beantwoorden interfereren met het antwoord zelf. Achteraf kunnen we een vergelijking maken tussen bv. het eerste kwart en het laatste kwart van de antwoorden, om zo te bekijken of er een mogelijk effect was van ervaring, d.w.z. van rijping.

3. Ook de **instrumentatie** of instrumenten die voor een onderzoek gebruikt worden, kunnen een bedreiging vormen voor de interne validiteit. Verschillende instrumenten die worden geacht hetzelfde construct te meten, moeten ook gelijke metingen produceren. En hetzelfde instrument moet ook gelijke metingen produceren onder verschillende omstandigheden. Voor computer-gestuurde experimenten is dat meestal geen probleem. Maar bij vragenlijsten, of bij de beoordeling van schrijfopdrachten, kan de interne validiteit wel worden bedreigd.

Bij veel onderzoek worden observaties gedaan zowel voorafgaand aan een behandeling, als na afloop daarvan. Daarbij kan dezelfde toets gebruikt worden, maar dan kan er een leer-effect optreden (zie hierboven). Onderzoekers gebruiken daarom vaak verschillende toetsen bij de voormeting en de nameting, maar daarbij kan er wel een instrumentatie-effect optreden. De onderzoeker moet de mogelijke voor- en nadelen tegen elkaar afwegen.

Voorbeeld 5.3: (Rijlaarsdam, 1986) onderzocht het effect van 'peer evaluation' op de kwaliteit van schrijfproducten. De opzet van zijn onderzoek was (enigszins vereenvoudigd) als volgt: eerst schrijven de leerlingen een opstel over onderwerp A, dan volgt het schrijfonderwijs inclusief 'peer evaluation', waarna nogmaals een opstel geschreven wordt over onderwerp B. De schrijfproducten van voormeting en nameting worden beoordeeld, waarna getoetst wordt of de gemiddelde prestaties verschillen tussen voormeting en nameting.

In dit onderzoek vormt niet alleen de interventie (schrijfonderwijs) een duidelijk verschil tussen de voormeting en de nameting, maar

ook het onderwerp van de schrijfopdracht (A of B) vormt een verschil. Het is twijfelachtig of met beide schrijfopdrachten wel precies hetzelfde wordt gemeten. Dit verschil in instrumentatie bedreigt de interne validiteit, omdat er op verschillende momenten misschien een (deels) verschillend aspect van de schrijfvaardigheid is gemeten. De instrumentatie (hier: het verschil in onderwerpen van de schrijfopdrachten) geeft een plausibele alternatieve verklaring voor een verschil in schrijfvaardigheid, naast of in plaats van de onafhankelijke variabele (hier: het tussentijdse schrijfonderwijs).

---

4. Een volgende bedreiging van de interne validiteit staat bekend als het effect van **regressie naar het gemiddelde**. Regressie naar het gemiddelde kan een rol spelen zodra het onderzoek gericht is op speciale groepen, bijvoorbeeld slechte lezers, slechte schrijvers, maar evenzo: goede lezers, goede schrijvers, etc. We geven eerst een voorbeeld, omdat het verschijnsel niet direct intuïtief duidelijk is.

Voorbeeld 5.4: Er is enige controverse over het gebruik van illustraties in kinderboeken. Sommigen menen dat in boeken waarmee kinderen leren lezen geen (of zo min mogelijk) illustraties mogen voorkomen: illustraties leiden de aandacht af van te leren kenmerken van woorden. Anderen menen dat in illustraties wezenlijke informatie weergegeven kan worden: illustraties dienen als extra informatiebron.

(Donald, 1983) onderzocht de invloed van illustraties bij een tekst op het begrip van die tekst. De onderzoeker selecteerde 120 leerlingen (uit 1868 leerlingen) uit de derde en zesde groep van het basisonderwijs; 60 uit elk van beide groepen. Volgens de prestaties op een eerder afgenomen leestoets bleken van de 60 leerlingen per klas er 30 als slechte en 30 als goede lezers geclassificeerd te kunnen worden. Elke leerling kreeg dezelfde tekst te zien, aangeboden met of zonder illustraties (onafhankelijke variabele), zie Tabel 5.1.

De resultaten bleken goeddeels de tweede hypothese te ondersteunen: illustraties bevorderen het begrip van de tekst, ook bij onervaren lezers. De slechtere lezers begrepen de tekst met illustraties beter, en ook jongere lezers ondervonden voordeel van de illustraties.

Table 5.1: Aanbiedingscondities in het onderzoek van Donald (1983).

groep	leesvaardigheid	conditie	n
3	slecht	zonder	15
3	slecht	met	15
3	goed	zonder	15
3	goed	met	15
6	slecht	zonder	15
6	slecht	met	15
6	goed	zonder	15
6	goed	met	15

Wat is er nu mis met dit onderzoek? Het antwoord is gelegen in de manier waarop leerlingen zijn geselecteerd. Lezers werden ingedeeld als 'slecht' of 'goed' op basis van een leesvaardigheidstoets, maar hun prestaties op die toets worden altijd beïnvloed door toevallige factoren, die niets met leesvaardigheid te maken hebben: Tom voelde zich niet lekker, daarom heeft hij deze toets slecht gemaakt, Sarah was met haar gedachten elders, Niels had last van zijn knie, Julie was enorm gemotiveerd en heeft zichzelf overtroffen. Met andere woorden: de leesvaardigheid is niet geheel betrouwbaar gemeten. Dit betekent (1) dat de slechte lezers die toevallig boven hun niveau gepresteerd hebben, ten onrechte niet bij de slechte lezers ingedeeld werden, maar deel uitmaakten van de groep goede lezers; en omgekeerd (2) dat goede lezers die bij deze toets toevallig onder hun niveau gepresteerd hebben, ten onrechte als slechte lezers bestempeld werden. Onder de slechte lezers zitten dus altijd ook een paar lezers die helemaal zo slecht nog niet zijn, en onder de goede lezers zitten ook een paar lezers die eigenlijk niet zo goed zijn.

Wanneer de eigenlijk-goede lezers, die ten onrechte geclassificeerd zijn als niet-goede lezers, een tweede leestoets maken (nadat zij een tekst met of zonder illustraties bestudeerd hebben), dan zullen zij meestal weer op hun gewone hoge niveau presteren. Een hogere score op de tweede toets (de nameting) kan dus een artefact zijn van de selectiemethode. Hetzelfde geldt, mutatis mutandis, voor de eigenlijk-slechte lezers die ten onrechte geselecteerd zijn als niet-slechte lezers. Wanneer deze leerlingen een tweede leestoets maken, dan zullen ook zij meestal weer op hun gewone (lage) niveau presteren. De score op de nameting ligt voor hen dus lager dan de score op de voormeting.

Voor het aangehaalde onderzoek van betekent dit dat het geconstateerde verschil tussen slechte en goede lezers deels toevallig is. Ook als de onafhankelijke variabele geen effect heeft, zal de groep 'goede' lezers bij de tweede leestoets gemiddeld slechter presteren, en zal de groep 'slechte' lezers bij de tweede leestoets gemiddeld beter presteren. Met andere woorden: het verschil tussen de twee groepen is bij de nameting minder groot dan bij de voormeting, als gevolg van

toevallige variatie: regressie naar het gemiddelde. Het zal duidelijk zijn dat onderzoeksresultaten getroebleerd kunnen worden door dit verschijnsel. Zoals we hierboven zagen kan een experimenteel effect afgezwakt worden of verdwijnen als gevolg van regressie naar het gemiddelde; omgekeerd kan regressie naar het gemiddelde ten onrechte aangezien worden als een experimenteel effect (Retraction Watch, 2018).

In het algemeen kan regressie naar het gemiddelde optreden als er een classificatie gemaakt wordt op basis van een voormeting, waarvan de scores samenhang vertonen met de scores van de nameting (zie Hoofdstuk **??**). Als er geen enkele correlatie is tussen voormeting en nameting, dan speelt regressie naar het gemiddelde zelfs de hoofdrol: een verschil tussen voormeting en nameting is dan alleen het gevolg van regressie naar het gemiddelde. Als er perfecte correlatie is, dan speelt regressie naar het gemiddelde geen enkele rol, maar dan is ook de voormeting niet informatief, want immers geheel (achteraf) te voorspellen uit de nameting.

Regressie naar het gemiddelde kan een alternatieve verklaring bieden voor de vermeende grote toename van scores tussen voormeting en nameting voor een lage prestatiegroep (bv. slechte lezers), ten opzichte van een kleinere toename voor een hoge prestatiegroep (bv. goede lezers). Omgekeerd kan het ook een alternatieve verklaring bieden voor de vermeende afname van scores tussen voormeting en nameting voor een hoge prestatiegroep (bv. goede lezers), ten opzichte van een lage prestatiegroep (bv. slechte lezers).

Het is beter om de groepen *niet* samen te stellen op basis van een van de uitkomsten van een van de metingen (voormeting of nameting), maar op basis van een ander, onafhankelijk criterium. De proefpersonen van beide groepen zullen dan zullen bij de voormeting ongeveer gemiddeld scoren, en het effect van regressie naar het gemiddelde is dan klein. In alle groepen zullen dan ongeveer evenveel proefpersonen zitten met een door het toeval iets te hoge als met een iets te lage uitgevallen score, zowel bij de voormeting als bij de nameting.

5. Een vijfde bedreiging van de interne validiteit is **selectie**. Hiermee doelen we (voornamelijk) op een zodanige verdeling van proefpersonen over verschillende condities dat deze bij aanvang van het onderzoek niet gelijkwaardig zijn. Wanneer bijvoorbeeld in de experimentele conditie alle slimme proefpersonen zitten, terwijl in de controleconditie alleen de domme leerlingen terecht gekomen zijn, dan kan een effect niet zonder meer aan de manipulatie van de onafhankelijke variabele toegeschreven worden. Het verschil in aanvangsniveau (hier: in intelligentie) levert dan een plausibele alternatieve verklaring die de interne validiteit bedreigt.

Voorbeeld 5.5: Voor een eerlijke vergelijking tussen scholen van hetzelfde schooltype (VMBO, HAVO, VWO, etc) moeten we rekening

houden met verschillen tussen scholen in hun ingangsniveau van de leerlingen. Stel dat school A leerlingen heeft met ingangsnivo 50, en eindexamennivo 100 (op een willekeurige schaal). School B heeft leerlingen met ingangsnivo 30, en eindexamennivo 90 (op dezelfde schaal). Is school B slechter dan A (want lager eindnivo), of is school B beter dan A (want kleiner verschil in eindnivo)?

In veel onderwijskundig onderzoek is het onmogelijk om leerlingen van verschillende klassen op basis van het toeval aan condities toe te wijzen — dit wordt wel aselecte toewijzing genoemd. Dit kan namelijk onoverkomelijke organisatorische problemen met zich meebrengen. Deze organisatorische problemen omvatten meer dan alleen het (aselect) splitsen van de klas, hoewel dit vaak al lastig te realiseren is. Ook moet de onderzoeker rekenschap afleggen van mogelijke overdrachtseffecten tussen de condities: de leerlingen praten met elkaar, leren elkaar misschien zelfs wel de essentialia van de experimentele conditie(s). Het uitblijven van een effect zou dan op tenminste één alternatieve manier verklaard kunnen worden. Vanwege de geschetste problematiek worden vaak complete schoolklassen toegewezen aan een van de condities. Maar klassen bestaan uit een aantal leerlingen van dezelfde school. Bij de keuze van leerlingen, en hun ouders, voor een school treedt zelf-selectie op (in het Nederlandse onderwijssysteem), waardoor er verschillen zijn in uitgangspositie tussen condities (d.w.z. tussen klassen binnen condities). Eventuele gevonden verschillen tussen condities zouden dus ook door zelfselectie van leerlingen naar scholen veroorzaakt kunnen zijn.

Hierboven is al de meest eenvoudige manier aangegeven om verschillende condities een gelijk aanvangsniveau te geven: wijs de leerlingen aselect, volgens toeval, 'at random', toe aan de condities. Deze methode staat bekend als *randomisatie* (Shadish et al., 2002, p.294 ff). We kunnen bijvoorbeeld randomiseren door leerlingen een willekeurig (random) nummer te geven (zie Appendix **??**) en daarna de 'even leerlingen' aan de ene conditie en de 'oneven leerlingen' aan de andere conditie toe te wijzen. Bij aselecte toewijzing van proefpersonen aan condities berusten alle verschillen tussen de proefpersonen in de verschillende condities op toeval, en worden die verschillen dus uitgemiddeld. Naar alle waarschijnlijkheid zijn er dan geen systematische verschillen tussen de onderscheiden groepen of condities. Dit geldt echter alleen indien de groepen groot genoeg zijn.

Randomisatie, de aselecte toewijzing van proefpersonen aan condities, moet onderscheiden worden van de aselecte steekproeftrekking uit een populatie (zie §**??**). Bij aselecte steekproeftrekking gaat het om de willekeurige selectie van proefpersonen uit de populatie van mogelijke proefpersonen naar de steekproef; we streven er dan naar dat de steekproef of steekproeven lijken op de populatie waaruit die getrokken is/zijn. Bij randomisatie gaat het om de willekeurige toewijzing van de proefpersonen uit de steekproef aan de verschillende condities van het onderzoek; we streven er dan naar dat de steekproeven lijken op elkaar.

Een tweede methode om twee gelijke groepen te creëren is *matching*. Bij matching worden proefpersonen eerst gemeten op een aantal relevante variabelen. Daarna worden koppels gevormd die een gelijke score op deze variabelen hebben. Van deze koppels wordt er één aan de ene conditie en één aan de andere conditie toegewezen. Matching heeft echter verschillende bezwaren. Ten eerste kan regressie naar het gemiddelde een rol gaan spelen. Ten tweede is matching, wanneer de proefpersonen op meerdere variabelen gematcht moeten worden, zeer bewerkelijk, en is een grote groep potentiële proefpersonen vereist. Ten derde wordt bij matching alleen rekening gehouden met variabelen die de onderzoeker relevant acht, en niet met andere onbekende variabelen. Bij randomisatie wordt niet alleen gerandomiseerd naar die relevante variabelen, maar ook naar andere eigenschappen die mogelijk een rol zouden kunnen spelen zonder dat de onderzoeker zich dat realiseert. Kortom, de relatief eenvoudige randomisatie is verre te prefereren boven matching.

6. **Uitval** van respondenten is de laatste bedreiging van interne validiteit. In sommige gevallen begint een onderzoeker met veel proefpersonen. Gedurende het onderzoek vallen echter proefpersonen uit. Zolang het percentage uitvallers beperkt blijft, is er geen probleem. Maar er ontstaat wel een probleem, als de uitval selectief is voor één van de onderscheiden condities. Is dat laatste wel het geval, dan kan er over die conditie niet veel meer gezegd worden. Het probleem van uitval speelt vooral een rol bij longitudinaal onderzoek. Dit is onderzoek waarbij een beperkte groep respondenten gedurende een langere periode gevolgd wordt. Men heeft daarbij echter te maken met mensen die verhuizen, of overlijden gedurende het experiment, of participanten die niet meer willen meewerken, enz. Dit kan een enorme reductie van het aantal respondenten teweeg brengen.

Hierboven hebben we een aantal veel voorkomende problemen besproken die de interne validiteit van een onderzoek kunnen bedreigen. De lijst is echter niet uitputtend! Elk type onderzoek heeft zo z'n eigen problemen, en het is de taak van de onderzoeker om alert te zijn op mogelijke bedreigingen van de interne validiteit. Probeer daartoe plausibele verklaringen te bedenken die een eventueel effect ook, of zelfs beter zouden kunnen verklaren dan de te onderzoeken oorzaak. De onderzoeker moet dus denken als zijn eigen scepticus, die geenszins overtuigd is dat de onderzochte factor werkelijk de oorzaak is van het gevonden effect. Welke mogelijke alternatieve verklaringen zijn er volgens die scepticus, en hoe zou de onderzoeker die bedreigingen voor de validiteit kunnen wegnemen door de opzet van het onderzoek? Dat vereist goed inzicht in de logische relaties tussen de onderzoeksvragen, de onderzochte variabelen, de resultaten, en de conclusie.

5.5 Constructvaliditeit

In een experimenteel onderzoek wordt een onafhankelijke variabele gemanipuleerd. Dit kan, afhankelijk van de vraagstelling, op vele manieren. Evenzo kan de wijze waarop de afhankelijke variabele(n) gemeten wordt op verschillende manieren vorm gegeven worden. De manier waarop de onafhankelijke en de afhankelijke variabelen vorm gegeven worden noemen we de *operationalisatie* van deze variabelen. De leesvaardigheid van leerlingen kan bijvoorbeeld geoperationaliseerd worden als (a) hun score op een tekstbegriptoets met open vragen; (b) hun score op een tekstbegriptoets met meerkeuzevragen; (c) hun score op een zgn. cloze-toets (ontbrekend woord invullen); of (d) als de mate waarin geschreven instructies uitgevoerd kunnen worden. Meestal zijn er heel veel manieren om een variabele te operationaliseren, en zelden volgt uit een theorie één dwingende beschrijving voor de wijze van operationalisatie van de onafhankelijke of de afhankelijke variabelen. *Constructvaliditeit*, of *begripsvaliditeit*, heeft betrekking op de mate waarin de operationalisatie van zowel de afhankelijke variabele(n) als de onafhankelijke variabele(n) een adequate afspiegeling is (zijn) van de theoretische constructen waar het onderzoek zich op richt. Met andere woorden: zijn de onafhankelijke en de afhankelijke variabelen goed gerelateerd aan de theoretische concepten waar het onderzoek op gericht is?

Voorbeeld 5.6: De *taalontwikkeling* van babies en peuters is lastig te observeren, en al helemaal als het gaat om de auditieve en perceptieve ontwikkeling van deze proefpersonen die nog niet of nauwelijks zelf spreken. Een veel gebruikte methode is het Head Turn Preference Paradigm (Johnson and Zamuner, 2010). Bij deze methode kijkt de baby eerst naar een groen knipperend licht recht voor zich. Als de aandacht van het kind zo is gevangen, dooft vervolgens het groene licht en begint een rood licht te knipperen, aan de linker of rechter zijde van de proefpersoon. Het kind draait dan zijn of haar hoofd om het knipperende licht te zien. Vervolgens wordt er een spraakgeluidsbestand afgespeeld, via een luidspreker vlak bij het knipperende licht aan de zijkant. De afhankelijke variabele is de periode waarin het kind zijwaarts blijft kijken (met minder dan 2 s onderbreking). Daarna begint een nieuwe aanbiedingscyclus. De kijktijd wordt opgevat als een indicatie voor de mate van voorkeur van het kind voor de gesproken stimulus.

De interpretatie van de verkregen kijktijden is echter lastig, omdat kinderen nu eens voorkeur hebben voor nieuwe geluidsstimuli (bv zinnen in een onbekende taal), en dan weer juist aan bekende stimuli (bv grammaticale vs ongrammaticale zinnen). Zelfs als de stimuli

nauwkeurig zijn afgestemd op het ontwikkelingsniveau van de proefpersoon, is het lastig om de afhankelijke variabele (kijktijd) goed te relateren aan het bedoelde theoretische construct (voorkeur van kind).

Voorbeeld 5.7: Zoals hierboven aangegeven kan het begrip *leesvaardigheid* op allerlei manieren worden geoperationaliseerd. Volgens sommigen kan leesvaardigheid niet goed gemeten worden met behulp van meerkeuzevragen (Houtman 1986, Shohamy 1984). Bij meerkeuzevragen worden de antwoorden zeer sterk beïnvloed door andere zaken zoals algemene ontwikkeling, gokvaardigheid, ervaring met eerdere toetsen, en door de vraagstelling zelf, zoals geïllustreerd in deze vraag:

Wie van de volgende personen heeft de afgelopen jaren een autobiografie gepubliceerd?
a. Jeanne d'Arc (algemene ontwikkeling)
b. mijn buurvrouw (vraagstelling, ervaring)
c. Malala Yousafzai
d. Alexander Graham Bell (algemene ontwikkeling)**

Deze vraag is duidelijk niet construct-valide voor het meten van kennis over autobiografieën.

Uiteraard gelden bovengenoemde problemen met de constructvaliditeit niet alleen voor schriftelijke vragen of meerkeuzevragen, maar ook voor mondelinge vragen aan proefpersonen.

Voorbeeld 5.8: Als we ouders mondeling de vraag stellen *Hoe vaak leest U uw kind eigenlijk voor?* dan wekken we met die vraag al de suggestie dat voorlezen wenselijk is. De ouders zouden hun voorleesgedrag wel eens kunnen overschatten. We meten dus niet alleen het construct 'voorleesgedrag', maar ook het construct 'neiging tot sociaal wenselijke antwoorden' (zie hierna).

Een notoir lastig construct om te operationaliseren is *schrijfvaardigheid*. Wat is een goed en wat is een slecht schrijfproduct? En wat is dan eigenlijk schrijfvaardigheid? Kan schrijfvaardigheid gemeten worden door een telling van relevante inhoudselementen in een tekst, moeten er zinnen of woorden geteld worden, of misschien vooral connectieven (*dus, want, omdat, hoewel*, enz), moeten er oordelen van *lezers* verzameld worden over de geschreven tekst (t.a.v. doelgerichtheid, publiekgerichtheid, stijl), of moet er één oordeel van lezers verzameld worden over de globale kwaliteit, moeten er spelfouten geteld worden, etc? De problemen bij de operationalisatie komen voort uit een gebrek aan theorie over schrijfvaardigheid, waaruit een definitie voor de kwaliteit van schrijfproducten afgeleid zou kunnen worden (Van den Bergh and Meuffels, 1993). Kritiek op onderzoek naar schrijfvaardigheid is daarom makkelijk, maar alternatieve operationalisaties van het construct zijn moeilijk.

Een ander lastig construct is de *verstaanbaarheid* van gesproken zinnen. Verstaanbaarheid ('intelligibility') kan op diverse manieren worden geoperationaliseerd. De eerste mogelijkheid is dat de onderzoeker de woorden of zinnen uitspreekt en dat de proefpersoon die naspreekt, waarbij fouten in de reproductie geteld worden; een nadeel hierbij is dat er nauwelijks controle is over de model-uitspraak van de onderzoeker. Een tweede mogelijkheid is dat de woorden of zinnen vooraf worden opgenomen en verder dezelfde procedure wordt gevolgd; een nadeel blijft dat de responsies worden beïnvloed door kennis van de wereld, grammaticale verwachtingen, bekendheid met de spreker of zijn taalgebruik, enz. De meest betrouwbare methode is die van de zgn. 'speech reception threshold' (Plomp and Mimpen, 1979) beschreven in het volgende voorbeeld. Deze methode heeft echter als nadeel dat ze tijdrovend is, niet goed automatisch afgenomen kan worden, en dat er veel stimulusmateriaal (spraakopnamen) nodig is (zijn) voor een enkele meting.

Voorbeeld 5.19: We laten een lijst van 13 gesproken zinnen horen, gemaskeerd met ruisgeluid. De verhouding tussen spraak en ruis (speech-to-noise ratio, SNR) wordt uitgedrukt in dB. Bij 0 dB SNR zijn spraak en ruis even luid, bij +3 dB SNR is de spraak 3 dB luider dan de ruis, bij -2 dB SNR is de spraak 2 dB *zachter* dan de ruis, etc. Na iedere zin moet de luisteraar de aangeboden zin naspreken. Als dat foutloos gebeurt, dan wordt voor de volgende zin de SNR met 2 dB verlaagd (minder spraak of meer ruis); als de responsie fout was, dan wordt voor de volgende zin de SNR met 2 dB verhoogd (meer spraak of minder ruis). Na een paar zinnen is er weinig variatie meer in SNR, en schommelt de SNR rond een optimum. De gemiddelde SNR over de laatste 10 aangeboden zinnen vormt de 'speech reception threshold' (SRT). Deze SRT is ook op te vatten als de SNR waarbij de helft van de zinnen goed wordt verstaan.

Tot nog toe hebben we het gehad over problemen met betrekking tot de constructvaliditeit van de afhankelijke variabelen. Maar ook de operationalisatie van de *on*afhankelijke variabele staat vaak ter discussie. De onderzoeker heeft immers vele keuzes moeten maken tijdens de operationalisering van zijn onafhankelijke variabele (zie §2.6), en de gemaakte keuzes zijn vaak wel aanvechtbaar.

Een onderzoek is niet constructvalide, of niet begripsvalide, als de operationalisaties van de afhankelijke variabelen de toets der kritiek niet kunnen doorstaan. Een onderzoek is ook niet constructvalide, als de onafhankelijke variabele niet een valide operationalisatie is van het-theoretische-begrip-zoals-bedoeld. Als die operationalisatie niet valide is, dan wordt er dus eigenlijk iets anders gemanipuleerd dan de bedoeling was. In dat geval is de relatie tussen de afhankelijke variabele en de gemanipuleerde onafhankelijke variabele zoals bedoeld niet eenduidig meer. Eventuele geobserveerde verschillen in de afhankelijke variabele hoeven niet alleen veroorzaakt te worden door de onafhankelijke variabele zoals bedoeld, maar kunnen ook beïnvloed zijn door andere factoren. Een bekend effect in dit opzicht is het zogenaamde Hawthorne-effect.

> *Voorbeeld 5.10*: De directie van de Hawthorne Works Factory (Western Electric Company) in Cicero (Illinois), USA, was gealarmeerd door slechte bedrijfsresultaten. Een team onderzoekers nam de gang van zaken onder de loep, waarbij ongeveer alles werd onderzocht: werktijden, beloning, pauzes, verlichting, verwarming, werkoverleg, management, enz. De resultaten van dat onderzoek (uit 1927) wezen uit dat de productiviteit enorm was gestegen – maar dat er geen verband was met een van de onafhankelijke variabelen. De toename van productiviteit werd uiteindelijk toegeschreven aan de grotere aandacht voor de werknemers.

Het Hawthorne-effect houdt dus in dat een verandering in gedrag niet samenhangt met de manipulatie van enige onafhankelijke variabele, maar dat die verandering van gedrag het gevolg is van een psychologisch verschijnsel: proefpersonen die weten dat ze worden geobserveerd, doen meer hun best om gewenst gedrag te vertonen.

Voorbeeld 5.11: (Richardson et al., 1978) vergeleken de effectiviteit van twee methoden ter verbetering van de leesvaardigheid van slechte lezers. De leerlingen werden geselecteerd op basis van hun scores op drie toetsen. De 72 geselecteerde leerlingen werden aselect toegewezen aan één van de twee methode-condities (gestructureerd leesonderwijs versus geprogrammeerde instructie). In de eerste conditie werd het gestructureerde leesonderwijs verzorgd door vier docenten, die aan een klein groepje (van vier leerlingen) les gaven. In feite is de leerling-docent-ratio dus 1 : 1. In de tweede conditie (geprogrammeerde instructie) bemoeiden de docenten zich zo min mogelijk met de leerlingen. Het experiment nam 75 sessies van 45 minuten in beslag. Na de tweede observatie bleek dat de leerlingen die volgens de eerste (gestructureerde) methode les gekregen hadden, beter vooruit waren gegaan dan de leerlingen die met behulp van de tweede methode (geprogrammeerde instructie) les gekregen hadden.

Tot zover is er geen probleem met dit onderzoek. Er ontstaat pas een probleem als we zouden concluderen dat de gestructureerde methode beter is dan de geprogrammeerde instructie. Een alternatieve verklaring, die in dit onderzoek niet uitgesloten kan worden, is dat het gevonden effect niet (alleen) het gevolg is van de methode, maar (ook) een gevolg is van de grotere individuele aandacht in de eerste conditie (gestructureerd leesonderwijs).

Net zoals bij de interne validiteit kan ook bij de construct- of begripsvaliditeit een aantal validiteitbedreigende factoren genoemd worden.

1. Een eerste bedreiging van de begripsvaliditeit is *mono-operationalisatie*. In veel onderzoeken wordt de afhankelijke variabele slechts op één manier geoperationaliseerd. De proefpersonen hoeven slechts één taak uit te voeren, bv. één auditieve taak met reactietijdmetingen (over meerdere aanbiedingen), of één vragenlijst (met meerdere vragen). Het onderzoek staat of valt dan met deze specifieke operationalisatie van de afhankelijke variabele. Over de validiteit van deze specifieke operationalisatie zijn dan geen verdere gegevens voorhanden. De onderzoeker laat in zo'n geval ruimte voor twijfel. Strikt genomen moeten we de onderzoeker immers op zijn woord geloven omtrent de validiteit van zijn operationalisering. Dergelijk onderzoek kan veel beter worden uitgevoerd. De onderzoeker moet dan het te meten construct op verschillende manieren operationaliseren, bv. door meerdere auditieve taken te laten uitvoeren, met niet alleen reactietijdmetingen maar ook met tellingen van foutieve responsies. Of de onderzoeker laat niet alleen een vragenlijst invullen, maar observeert

het bedoelde construct ook d.m.v. andere taken en observatiemethoden. Wanneer de prestaties op de verschillende typen responsies in hoge mate samenhangen, kan daarmee aangetoond worden dat al deze toetsen hetzelfde construct vertegenwoordigen. We noemen dit *convergente validiteit*. Er is sprake van convergente validiteit als de prestaties op instrumenten die hetzelfde theoretische construct vertegenwoordigen, in hoge mate samenhangen (convergeren).

Het is echter niet voldoende om te laten zien dat toetsen die hetzelfde concept of construct beogen te meten, inderdaad convergent valide zijn. Daarmee is immers nog niet aangetoond wat dit construct is, en evenmin of het gemeten construct wel het bedoelde construct is. Hebben we wel echt 'vloeiendheid' van de spreker gemeten, met meerdere methoden, of hebben we eigenlijk steeds het construct 'aandacht' of 'spreeksnelheid' gemeten? En hebben we wel echt 'mate van tekstbegrip' gemeten, met verschillende convergente methoden, of hebben we eigenlijk steeds het construct 'faalangst' gemeten? Om de construct-validiteit te waarborgen moet eigenlijk ook worden aangetoond dat de operationalisaties *divergent valide* zijn ten opzichte van operationalisaties die een ánder aspect of een ándere (verwante) vaardigheid beogen te meten. Kortom de onderzoeker moet kunnen aantonen dat de prestaties op instrumenten (operationalisaties) die één vaardigheid (construct) vertegenwoordigen in hoge mate samenhangen (convergeren), terwijl de prestaties op instrumenten die verschillende vaardigheden vertegenwoordigen juist lage samenhang vertonen (divergeren). Pas dan heeft de onderzoeker aannemelijk gemaakt dat de specifieke operationalisaties inderdaad constructvalide zijn.

2. Ook de verwachtingen van de onderzoeker — die zich uiten in bewust èn onbewust gedrag — kunnen de constructvaliditeit van een onderzoek bedreigen. De onderzoeker is ook een mens, en is dus niet immuun voor de invloed van zijn of haar eigen verwachtingen op de uitkomsten van het onderzoek. Na afloop van het experiment is de invloed van de onderzoeker helaas moeilijk te achterhalen.

Voorbeeld 5.12:Kluger Hans was een paard dat kon rekenen. Als aan Kluger Hans gevraagd werd *hoeveel is* $4+4$*?*, dan stampte het paard 8 maal met zijn rechter voorhoef, als gevraagd werd *hoeveel is* $3-1$*?*, dan stampte Hans twee maal met zijn voorhoef. Kluger Hans baarde veel opzien en werd onderwerp van verschillende studies. Een commissie stelde in 1904 vast dat Kluger Hans inderdaad kon rekenen (en communiceren met mensen). Later constateerde een lid van de onderzoekscommissie, Carl Stumpf, samen met zijn assistent Oskar Pfungst, echter: "…het paard laat verstek gaan, als de oplossing van de gestelde opgave aan geen van de aanwezigen bekend is" (Pfungst,

> 1907, p.185, vert. HQ), of als het de persoon die de oplossing weet niet kan zien. “Es bedarf also optischer Hilfen” (idem). Na nauwkeurige observaties bleek dat de baas van Kluger Hans (en andere aanwezigen) zich een heel klein beetje ontspande zodra Hans het juiste aantal malen met zijn rechter voorpoot gestampt had. Dit onopzettelijke teken was voor Kluger Hans voldoende aanleiding om het stampen te staken (d.i. om zijn rechter voorhoef op de grond te houden), teneinde daarna zijn beloning van wortels en brood in ontvangst te nemen (Pfungst, 1907) (Watzlawick, 1977, p.38–47).
>
> Een misschien vergelijkbaar, recenter geval is dat van Alex, een papegaai met bijzondere cognitieve gaven, zie o.a. (Boswall, zj) en (Ale, 2015).

Het beroemde voorbeeld van Kluger Hans illustreert hoe subtiel de invloed van een onderzoeker of proefleider op het te onderzoeken object kan zijn. Deze invloed bedreigt natuurlijk de constructvaliditeit. Het is daarom beter als de onderzoeker niet ook zelf fungeert als experimentator[1] of proefleider. Studies waarin de onderzoeker zelf optreedt als behandelaar of docent of beoordelaar, kunnen worden bekritiseerd omdat de (verwachtingen van de) onderzoeker de uitkomsten kunnen beïnvloeden, waardoor de constructvaliditeit van de onafhankelijke variabele wordt bedreigd. Onderzoekers kunnen zich wel verweren tegen deze ‘experimenter bias’. In het Head Turn Preference Paradigm (voorbeeld 5.6), bijvoorbeeld, is het gebruikelijk dat de experimentator niet weet uit welke groep een proefpersoon afkomstig is, en dat de experimentator niet hoort welk geluidsbestand wordt aangeboden (Johnson and Zamuner, 2010, p.74).

3. Een derde bedreiging van de constructvaliditeit kan samengevat worden onder de term *motivatie*. Aan de bedreiging van de validiteit door motivatie zitten tenminste twee kanten. Als (ten minste) één van de condities in een onderzoek erg belastend of vervelend is, dan kunnen de proefpersonen gedemotiveerd raken en zich minder inspannen bij hun taken. Ze presteren dan minder, maar dit is een effect van (gebrek aan) motivatie, en niet een direct effect van de onafhankelijke variabele (hier: conditie). Het effect hoeft dan niet veroorzaakt te worden door de manipulatie van het bedoelde construct, maar door de onbedoelde manipulatie van de *motivatie* van de proefpersonen. Ook het omgekeerde kan natuurlijk een bedreiging van de constructvaliditeit vormen. Indien van één van de condities een extra motiverende werking op de proefpersonen heeft, dan kan een eventueel effect toegeschreven worden aan motivationele aspecten. Ook dan kan er sprake zijn van een effect van een onbedoeld gemanipuleerde variabele.

[1] De experimentator is degene die een experiment afneemt bij een proefpersoon. De experimentator kan een andere persoon zijn dan de onderzoekers die de onderzoekshypothesen hebben opgesteld en/of proefpersonen hebben gerecruteerd.

4. Een vierde bedreiging van de validiteit heeft te maken met de keuze uit de vele mogelijke waarden van een onafhankelijke variabele, d.w.z. de *'dosering'* ervan. Als de onafhankelijke variabele is 'het aantal keren dat een gedicht ter voorbereiding mag worden doorgelezen', moet de onderzoeker bepalen hoeveel keer de proefpersonen het gedicht mogen doorlezen: één, twee, drie of meer keren? Als de onafhankelijke variabele is 'de tijd die de proefpersonen mogen studeren', dan moet de onderzoeker kiezen hoe lang de proefpersonen mogen leren: vijf minuten, een kwartier, twee uur? De onderzoeker maakt een keuze uit de dosering van de onafhankelijke variabele 'leertijd'. Op grond van deze dosering kan de onderzoeker concluderen dat de afhankelijke variabele niet beïnvloed wordt door de onafhankelijke variabele. In feite moet de onderzoeker echter concluderen dat er geen verband lijkt tussen de *gekozen dosering van* de onafhankelijke variabele, en de afhankelijke variabele. Een mogelijk effect wordt verhuld door de keuze van de dosering (waarden) van de onafhankelijke variabele.

Voorbeeld 5.13: Als een personenauto en een voetganger botsen, loopt de voetganger een risico te overlijden. Dat overlijdensrisico is relatief gering (kleiner dan 20%) bij botsingssnelheden tot ca 50 km/u. Als we ons onderzoek naar het verband tussen botsingssnelheid en overlijdensrisico zouden beperken tot deze lage 'doseringen' van botsingssnelheden, dan zouden we wellicht concluderen dat de botsingssnelheid géén invloed heeft op het overlijdensrisico voor de voetganger. Dat zou een foutieve conclusie zijn (van welk type?), want bij hogere botsingssnelheden neemt het overlijdensrisico voor de voetganger toe tot bijna 100% (Rosén et al., 2011; SWOV, 2012).

---

5. Een vijfde bedreiging van de constructvaliditeit wordt veroorzaakt door de *sturende werking van de voormeting*. In veel studies wordt de afhankelijke variabele herhaaldelijk gemeten, zowel voor als na manipulatie van de afhankelijke variabele: de zgn. voormeting en nameting. De aard en inhoud van de voormeting kunnen echter sporen nalaten bij de proefpersoon. Zo kan de proefpersoon zijn onbevangenheid verliezen, waardoor het effect van de onafhankelijke variabele (bv. behandeling) wordt verkleind. Een eventueel verschil in scores tussen de experimentele condities kan dus op meerdere manieren worden verklaard. De verklaring kan immers liggen in een effect van alleen de onafhankelijke variabele, maar kan ook liggen in een effect van *de combinatie van voormeting en onafhankelijke variabele*. Bovendien kan de afwezigheid van een effect soms worden verklaard door het feit dat een voormeting is verricht (zie het Solomon vier-groepen-ontwerp, in Hoofdstuk **??**, voor een onderzoeksontwerp dat hiermee rekening houdt).

Voorbeeld 5.14: We kunnen de effecten van twee behandelingen vergelijken in een experiment waarin de deelnemers volgens het toeval in twee groepen worden ingedeeld. De eerste groep (E) krijgt eerst een voormeting, dan een behandeling, en dan een nameting. De tweede groep (C) krijgt geen voormeting, en ook geen behandeling, maar alleen een nameting, die voor deze groep de enige meting is.

Als we bij de nameting een verschil vinden tussen de twee groepen, dan is dat niet zonder meer toe te schrijven aan het verschil in behandeling. Het verschil zou ook, of mede, veroorzaakt kunnen zijn door de sturende werking van de voormeting, bv als gevolg van de sturende woordkeuze of zinsbouw van de vragen of opdrachten in de voormeting. Misschien hebben de deelnemers in groep E iets geleerd in de voormeting, d.w.z. *niet* in de behandeling, waardoor ze beter of anders presteren in de nameting dan de deelnemers in groep C.

---

6. Een ander probleem dat van invloed kan zijn op de constructvaliditeit is *sociaal wenselijk antwoorden*. Dat is niets anders dan dat mensen geneigd zijn een antwoord geven, dat in de gegeven sociale situatie wenselijk is, en dat hen dus niet in de problemen brengt of tot gezichtsverlies leidt. Een voorbeeld kan dit verduidelijken.

Voorbeeld 5.15: Bij peilingen voor verkiezingen zijn respondenten geneigd om sociaal wenselijk te antwoorden, en dat geldt ook voor de vraag of de respondent überhaupt zal gaan stemmen (Karp and Brockington, 2005). De neiging tot het sociaal wenselijke antwoord ("ja, ik ga stemmen") is sterker naarmate respondenten hoger zijn opgeleid, en dus is de overschatting van het opkomst-percentage groter voor hoger-opgeleiden dan voor lager-opgeleiden. Dat heeft weer gevolgen voor de uitslagen van de peilingen van de verschillende partijen, omdat de populariteit van de politieke partijen verschillend is voor kiezers van verschillend opleidingsniveau.

Dit effect heeft mede gezorgd voor de overschatting van het aantal Clinton-stemmers, en onderschatting van het aantal Trump-stemmers, bij de peilingen voorafgaand aan de Amerikaanse presidentsverkiezing in 2016.

7. Een laatste probleem met betrekking tot de constructvaliditeit kan aangeduid worden als: een *beperkte generaliseerbaarheid* over constructen. Bij de presentatie van onderzoeksresultaten worden regelmatig opmerkingen gemaakt als: 'Ja, ik ben het eens met uw conclusie dat X van invloed is op Y, maar hoe zit het met…'. Op de puntjes kan dan van alles ingevuld worden: de toepasbaarheid bij andere doelgroepen, of in andere genres, of in andere talen, etc. Deze aspecten zijn weliswaar van belang, maar spelen in het onderzoek zelf niet direct een rol: we hebben het onderzoek immers uitgevoerd met een bepaalde selectie van doelgroep, genre, talen, etc.

Toch bevelen we wel aan om zulke vragen over generaliseerbaarheid onder ogen te zien. Zijn de conclusies eveneens van toepassing op een andere doelgroep of taal? Waarom wel of niet? Welke andere factoren zouden de generalisatie kunnen beïnvloeden? Zou een gunstig effect voor de ene groep of taal ook kunnen uitpakken als een ongunstig effect voor een andere groep of taal die buiten het onderzoek is gevallen?

5.6 Externe validiteit

Op basis van de gegevens die zijn verzameld kan een onderzoeker — als het goed is — de conclusie trekken: *in dit onderzoek geldt dat…*. Het is echter zelden de bedoeling van een onderzoeker om conclusies te trekken die alleen gelden voor één onderzoek. Een onderzoeker wil niet aantonen dat tweetaligheid een gunstige invloed heeft op de taalontwikkeling *van de steekproef van onderzochte kinderen*. Een onderzoeker wil conclusies trekken als: tweetaligheid heeft een gunstige invloed op de taalontwikkeling *van kinderen*. De onderzoeker wil generaliseren. In het dagelijks leven doen we hetzelfde: we proeven één hapje soep uit een hele pan, en op grond daarvan doen we een uitspraak over die hele pan soep. We gaan er van uit dat onze bevindingen op basis van dat ene hapje gegeneraliseerd mogen worden naar de hele pan, en dat het niet nodig is om de hele pan leeg te eten voordat we er een uitspraak over kunnen doen.

De vraag of een onderzoeker de resultaten kan en mag generaliseren is de vraag naar de *externe validiteit* van een onderzoek (Shadish et al., 2002). Generalisatie heeft betrekking op o.a.

- eenheden: zijn de resultaten ook geldig voor andere elementen (bv. scholen, personen, teksten) uit de populatie, die niet aan het onderzoek deelnamen?
- behandelingen: zijn de resultaten ook geldig voor andere behandelingen die lijken op de specifieke condities in dit onderzoek?
- situaties: zijn de resultaten ook geldig buiten de specifieke context van dit onderzoek?

- tijden: zijn de resultaten van dit onderzoek ook geldig op andere tijdstippen?

Bij externe validiteit maken we een onderscheid tussen (1) de generalisatie *naar* een beoogde specifieke doelgroep, situatie en tijd, en (2) de generalisatie *over* andere doelgroepen, situaties en tijden. Het generaliseren *naar* en *over* zijn twee aspecten van de externe validiteit die goed uit elkaar gehouden moeten worden. Het generaliseren *naar* een doelgroep of populatie, van personen en vaak ook van taalmateriaal, heeft te maken met de representativiteit van de gebruikte steekproef; in hoeverre is de steekproef een goede afspiegeling van de populatie (van personen, van woorden, van relevante mogelijke zinnen)? Het generaliseren *naar* is dus direct verbonden met het onderzoeksdoel; pas als er gegeneraliseerd kan worden naar gedefinieerde populaties kan een onderzoeksdoel bereikt zijn. Het generaliseren *over* doelgroepen heeft te maken met de mate waarin de geformuleerde conclusies geldig zijn voor te onderscheiden deelpopulaties. We illustreren dit met een voorbeeld.

Voorbeeld 5.16: (Lev-Ari and Keysar, 2010) onderzochten of luisteraars minder geloof hechten aan sprekers met een vreemd buitenlands accent in de uitspraak van het Engels. Voor de stimuli lieten ze zinnen uitspreken (bv. *A giraffe can hold more water than a camel*) door verschillende sprekers zonder enig accent, met licht accent, of met sterk accent. Luisteraars (moedertaal-sprekers van het Engels) gaven aan in welke mate ze dachten dat de gesproken zin waar was. De resultaten lieten zien dat de luisteraars de zinnen beoordeelden als minder waar, als de zin was gesproken door een spreker met een vreemd buitenlands accent.

We mogen aannemen dat deze uitkomst gegeneraliseerd kan worden *naar* de beoogde doelgroep, nl. alle moedertaal-luisteraars van het Amerikaans Engels. Deze generalisatie kan worden gemaakt ondanks de mogelijkheid dat verschillende luisteraars misschien in verschillende mate beïnvloed werden door het buitenlandse accent van de spreker.

Wellicht zou een latere analyse kunnen laten zien dat er verschil is tussen vrouwelijke en mannelijke luisteraars. Het is denkbaar dat vrouwen en mannen verschillen in hun gevoeligheid voor het accent van de spreker. Zo'n (denkbeeldige) uitkomst zou laten zien dat er niet gegeneraliseerd mag worden *over* deel-populaties binnen de doelgroep, hoewel er wel gegeneraliseerd kon worden *naar* de doelgroep.

In het (toegepast) taalwetenschappelijk onderzoek proberen onderzoekers doorgaans om *tegelijkertijd* te generaliseren naar *twee* populaties van eenheden, nl. van personen (c.q. scholen of families) *en stimuli* (woorden, zinnen, teksten, enz). We willen aannemelijk maken dat de resultaten niet alleen geldig zijn voor de onderzochte taalgebruikers, maar ook voor andere taalgebruikers. Tegelijkertijd willen we ook aannemelijk maken dat de resultaten niet alleen geldig zijn voor de onderzochte stimuli, maar ook voor andere vergelijkbaar taalmateriaal waaruit de steekproef van stimuli is getrokken. Die gelijktijdige generalisatie vereist een complex onderzoeksontwerp, doordat er herhaalde observaties zijn binnen proefpersonen (meerdere oordelen van eenzelfde proefpersoon) en binnen stimuli (meerdere oordelen over dezelfde stimulus). Stimuli, proefpersonen en condities worden vervolgens slim gecombineerd om de interne validiteit zo goed mogelijk te beschermen. Uiteraard vereist de generalisatie naar ander taalmateriaal wel, dat de stimuli willekeurig zijn geselecteerd uit de (soms oneindig grote) populatie van al het mogelijke taalmateriaal (zie Hoofdstuk **??**).

Bibliography

(2015). Alex Foundation.

American Psychological Association (2010). *Publication Manual of the American Psychological Association*. American Psychological Association, Washington, D.C., 6th edition.

Boswall, J. (z.j.). Alex, the talking parrot.

De Groot, A. (1961). *Methodologie: Grondslagen van onderzoek en denken in de gedragswetenschappen*. Mouton, 's-Gravenhage.

Deutsch, D. (2006). The enigma of absolute pitch. *Acoustics Today*, 2:11–19.

Dingemanse, M., Torreira, F., and Enfield, N. (2013). Is "huh?" a universal word? conversational infrastructure and the convergent evolution of linguistic items. *PLOS One*, 8(11):e78273.

Donald, D. (1983). The use and value of illustrations as contextual information for readers at different progress and developmental levels. *British Journal of Educational Psychology*, 53(2):175–185.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge.

Hume, D. (1739). *A Treatise on Human Nature*.

Johnson, E. K. and Zamuner, T. (2010). *Using infant and toddler testing methods in language acquisition research*, chapter 4, pages 73–93. John Benjamins, Amsterdam.

Karp, J. A. and Brockington, D. (2005). Social desirability and response validity: A comparative analysis of overreporting voter turnout in five countries. *Journal of Politics*, 67(3):825–840.

Kerlinger, F. N. and Lee, H. B. (2000). *Foundations of Behavioral Research*. Harcourt College Publishers, Fort Worth, 4th edition.

Koring, L., Mak, P., and Reuland, E. (2012). The time course of argument reactivation revealed: Using the visual world paradigm. *Cognition*, 123(3):361–379.

Lev-Ari, S. and Keysar, B. (2010). Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology*, 46(6):1093–1096.

Morton, A. (2003). *A Guide through the Theory of Knowledge*. Blackwell, Malden, MA, 3e edition.

Office of Research Integrity (2012). Responsible conduct of research training.

Pfungst, O. (1907). *Das Pferd des Herrn von Osten (Der kluge Hans): Ein Beitrag zur experimentellen Tier- und Menschen-Psychologie*. J. A. Barth, Leipzig.

Plomp, R. and Mimpen, A. M. (1979). Improving the reliability of testing the speech reception threshold for sentences. *International Journal of Audiology*, 18(1):43–52.

Popper, K. (1935). *Logik der Forschung. Zur Erkentnistheorie der modernen Naturwissenschaft*. Julius Springer, Wien.

Popper, K. (1959). *The logic of scientific discovery*. Routledge, London.

Popper, K. (1963). *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge and Kegan Paul, London.

Quené, H. (2008). Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. *Journal of the Acoustical Society of America*, 123(2):1104–1113.

Quené, H., Semin, G. R., and Foroni, F. (2012). Audible smiles and frowns affect speech comprehension. *Speech Communication*, 54(7):917–922.

Retraction Watch (2018). The "regression to the mean project:" what researchers should know about a mistake many make. Technical report.

Richardson, E., DiBenedetto, B., Christ, A., Press, M., and Winsberg, B. G. (1978). An assessment of two methods for remediating reading deficiencies. *Reading Improvement*, 15(2):82.

Rijlaarsdam, G. (1986). *Effecten van leerlingrespons op aspecten van stelvaardigheid*. PhD thesis.

Rosenthal, R. and Rosnow, R. L. (2008). *Essentials of Behavioral Research: Methods and Data Analysis*. McGraw Hill, Boston, 3e edition.

Rosén, E., Stigson, H., and Sander, U. (2011). Literature review of pedestrian fatality risk as a function of car impact speed. *Accident Analysis and Prevention*, 43(1):25–33.

Sanders, E. (2011). *Eerste Hulp bij e-Onderzoek voor studenten in de geesteswetenschappen: Slimmer zoeken, slimmer documenteren*. Early Dutch Books Online.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Wadsworth, Belmont, CA.

SWOV (2012). De relatie tussen snelheid en ongevallen.

Universiteitsbibliotheek, Vrije Universiteit Amsterdam (2015). Webcursus informatievaardigheden - algemeen - niveau b.

Van den Berg, M., Amuzu, E. K., Essizewa, K., Yevudey, E., and Tagba, K. (2017). Crosslinguistic effects in adjectivization strategies in Suriname, Ghana and Togo. In Cutler, C., Vrzić, Z., and Angermeyer, P., editors, *Language Contact in Africa and the African Diaspora in the Americas: in honor of John V. Singler*, pages 343–362. Benjamins, s.l.

Van den Bergh, H. and Meuffels, B. (1993). Schrijfvaardigheid. In Braet, A. and Van de Gein, J., editors, *Taalbeheersing als tekstwetenschap: terreinen en trends*. ICG, Dordrecht.

Verhoeven, J., De Pauw, G., and Kloots, H. (2004). Speech rate in a pluricentric language: A comparison between Dutch in Belgium and the Netherlands. *Language and Speech*, 47(3):297–308.

VSNU (2018). Nederlandse gedragscode wetenschappelijke integriteit. Technical report, VSNU.

Watzlawick, P. (1977). *Is 'werkelijk' waar? Spraakverwarring, zinsbegoocheling en onvoorstelbare werkelijkheid*. Van Loghum Slaterus, Deventer.

Xie, Y. (2020). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.18.