WAARHEIDSVINDING ALS WETENSCHAPPELIJK EXPERIMENT

De psychologie van strafrechtelijk bewijs uitgelegd aan de hand van wetenschappelijke publicaties uit Nederland

WAARHEIDSVINDING ALS WETENSCHAPPELIJK EXPERIMENT

De psychologie van strafrechtelijk bewijs uitgelegd aan de hand van wetenschappelijke publicaties uit Nederland

Eric Rassin

www.bravenewbooks.nl/ericrassin

ISBN: 9789465389110

INHOUD

INLEIDING

Dit boek gaat over de psychologie van bewijs, ook wel rechtspsychologie genoemd. Er zijn inmiddels meerdere boeken over dat onderwerp op de markt. Het indelen van de hoofdstukken kan op verschillende manieren, bijvoorbeeld per deelonderwerp (met hoofdstuktitels zoals getuigen, verdachten en gezichtsherkenning; zie Rassin, 2009), of aan de hand van casus (Rassin, 2023). In dit boek heb ik een andere ordening gekozen. Ik heb een aantal publicaties van Nederlandse rechtspsychologen geselecteerd, aan de hand waarvan ik enkele rechtspsychologische thema's behandel. Rode draad is wetenschap: wat maakt (rechtspsychologische) kennis en inzichten wetenschappelijk?

Elk hoofdstuk begint met de bespreking van een publicatie van een Nederlandse rechtspsycholoog. De inhoud van die publicatie wordt besproken zodat de lezer kennismaakt met dat onderwerp. Daarnaast wordt ingegaan op de methodologie van het onderzoek dat aan de betreffende publicatie ten grondslag ligt. Daardoor krijgt de lezer ook inzicht in wat wetenschap eigenlijk is en met dat inzicht wordt het gemakkelijker om de waarde van uitspraken met wetenschappelijke pretentie op hun merites te beoordelen. Het doel van dit boek is dus tweeledig: kennisnemen van rechtspsychologische thema's en inzicht verwerven in gedragswetenschappelijke methodologie. Het boek heeft niet tot doel om de lezer *up to date* te brengen van de *state of the art*. Daarvoor leent een boek zich heden ten dage niet meer goed. De ontwikkelingen gaan snel en voor het laatste nieuws kan men zich beter wenden tot tijdschriften en andere media.

Wat is een wetenschappelijk experiment?

Wie in de zoekmachine van Google "experimenteel" intikt, krijgt als antwoord "Experimenteel betekent dat iets gebaseerd is op of betrekking heeft op een experiment, een

methode om iets te onderzoeken door het te testen en te observeren. Het kan ook duiden op iets wat nog niet volledig is bewezen of algemeen geaccepteerd, maar wordt gebruikt als onderdeel van een test of onderzoek". De gratis online Van Dale schrijft: "1) Proefondervindelijk, 2) Het karakter van een experiment hebbend, 3) zich kenmerkend door experimenten". Uit deze omschrijvingen valt af te leiden dat het begrip "experimenteel" zowel een positieve (dat wil zeggen kenmerkend voor wetenschap) als een ietwat negatieve (nog niet helemaal bewezen) connotatie kan hebben. In dit boek wordt het gebruikt in de eerste betekenis: kennis die is gebaseerd op experimenteel onderzoek is de empirisch-wetenschappelijk hoogst haalbare kennis.

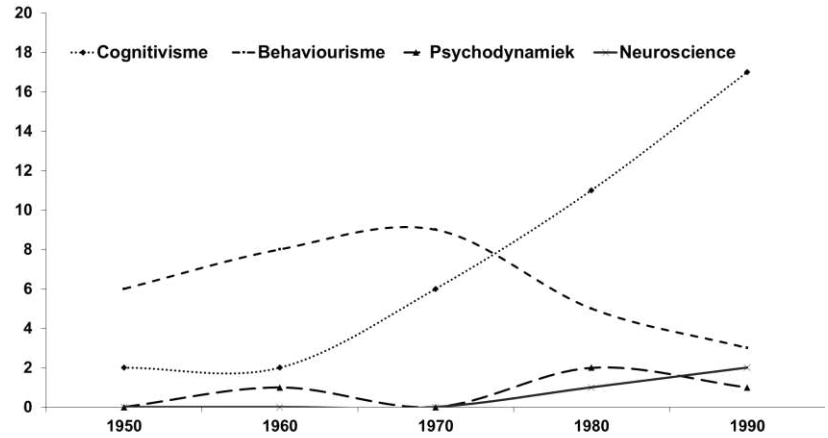
Een goed experiment kent tenminste de volgende aspecten: een theoretisch uitgangspunt, een hypothese op basis van die theorie, een goed design voor een systematische waarneming, een statistische analyse, een interpretatie en discussie van de resultaten, en de publicatie van het onderzoek. In het navolgende worden deze aspecten behandeld voor de psychologie als academische discipline.

Op het kenmerkt van het theoretisch uitgangspunt scoort de psychologie matig, omdat er welbeschouwd niet één coherente, maar vier concurrente theoretische stromingen zijn. De wellicht bekendste is de psychodynamiek, waarvan Sigmund Freud (1856-1939) de voornaamste agonist, zo niet de grondlegger is. Freud's werk is omvangrijk en het laat zich moeilijk samenvatten. Centraal staat dat de menselijke psyche uit verschillende compartimenten bestaat waarvan sommige (bijvoorbeeld het Es) grotendeels onbewust opereren. De psychodynamiek wordt daarom ook wel dieptepsychologie genoemd. De tweede stroming staat hier juist haaks op. In het *behaviourisme* gaat het in het geheel niet om wat zich binnenin afspeelt, maar uitsluitend om het observeerbare (overte) gedrag, vandaar dat deze benadering ook bekend

is als de gedragswetenschappen. In deze benadering wordt de mens niet aangestuurd door onbewuste impulsen (zoals bij Freud), maar door interacties met de omgeving. Het draait om de mens (en het dier) als lerend organisme. Bekende concepten zijn klassieke conditionering (leren voorspellen) en operante conditionering (leren interacteren met de omgeving). Bekende behaviouristen zijn Ivan Pavlov (1849-1936) en Burrhus Skinner (1904-1990). De derde school, het cognitivisme, houdt enigszins het midden tussen de psychodynamiek en het behaviourisme. Deze stroming heeft als uitgangspunt dat er voor psychologen toch meer te bestuderen is dan alleen het observeerbare gedrag, maar anders dan bij de psychodynamiek, gaat het cognitivisme er niet vanuit dat de meeste psychologische processen onbewust (en dus onkenbaar) zijn. Tot slot is er de neuropsychologie, of *neuroscience*, waarin het draait om het leggen van verbanden tussen gedrag en fysiologische processen. Cesare Lombroso (1835-1909) zou een neuropsycholoog *avant-la-lettre* kunnen worden genoemd. En waar Aleksandr Luria (1902-1977) in de Tweede Wereldoorlog, noodgedwongen onderzoek deed bij soldaten met hersenletsel, staan de neurowetenschapper tegenwoordig minder invasieve methoden ter beschikking zoals *functionele magnetische resonantie imaging* (fMRI's).

Robins et al. (1999) probeerden in kaart te brengen wat de relatieve bijdrage van de verschillende theoretische stromingen aan de wetenschappelijke psychologie is. Zo berekenden ze de impact factor van vier toonaangevende tijdschriften per school. Stel dat er in een bepaald jaar 100 artikelen verschijnen in een wetenschappelijk tijdschrift en dat in datzelfde jaar 1.000 keer wordt verwezen naar artikelen in dat tijdschrift, dan is de impact factor (1.000 : 100 =) 10. In andere woorden: een gemiddeld artikel in dit tijdschrift is een inspiratiebron voor tien nieuwe onderzoeken. De onderzoekers vonden dat de gemiddelde impact factor van vier psychodynamische tijdschriften in 1996, 0,85 was, die van vier behaviouristische tijdschriften

1,98, die van vier cognitieve tijdschriften 2,46 en die van vier neuroscience tijdschriften 15,8. Robins et al. (1999) keken ook hoeveel artikelen er zijn verschenen tussen 1950 en 1997, in toonaangevende stromingsloze tijdschriften zoals *Psychological Bulletin* en *Psychological Review*. Het resultaat van die analyse is weergegeven in Figuur 1.



Figuur 1. Publicaties per stroming in toonaangevende tijdschriften als percentage van alle publicaties (Robins et al., 1999).

Een tweede kenmerk van wetenschap is dat er, op basis van theorieën, toetsbare hypothesen worden geformuleerd. Omdat hypothesen toetsbaar moeten zijn, moeten er enkele eisen worden gesteld aan stellingen voordat zij als wetenschappelijke hypothese kunnen worden gezien. Karl Popper (1902-1994) heeft een grote invloed gehad op de manier waarop wij tegenwoordig naar hypothesen en de toetsing daarvan kijken. Om te beginnen dacht Popper (1959, 1963) dat hypothesen niet kunnen worden bevestigd door te zoeken naar bevestiging. Om dit te begrijpen, moeten we onderscheid maken tussen universele hypothesen en existentiële hypothesen. Een voorbeeld van de eerste soort is "alle goudvissen zijn oranje". Een voorbeeld van de tweede soort is "er bestaat een paarse goudvis". Als Popper schrijft over de

(on)toetsbaarheid van hypothesen, gaat het over universele uitspraken. Bij die uitspraken is het onmogelijk om te beslissen hoeveel bevestigende observaties er moeten worden gedaan om de hypothese te kunnen bevestigen. Dat is het zogenoemde inductieprobleem. Stel dat we gaan zoeken en 100 of zelfs 1.000 goudvissen vinden, die allemaal oranje zijn. Klopt de stelling dat alle goudvissen oranje zijn dan? Niet per se. Stel echter dat we toevallig een goudvis zien die niet oranje is, dan weten we wel meteen zeker dat de hypothese incorrect is. Universele uitspraken zijn informatief, omdat ze een groot domein betreffen, niet verifieerbaar vanwege het inductieprobleem, maar wel falsifieerbaar. Sinds Popper's geschriften, worden universele uitspraken ideaal geacht als wetenschappelijke hypothesen, in tegenstelling tot existentiële uitspraken die in veel opzichten het tegenovergestelde zijn.

In Popper's gedachtegang betekent wetenschappelijke toetsbaarheid vooral weerlegbaarheid ofwel falsifieerbaarheid. Sterker, een hypothese is wetenschappelijker naarmate hij vergezochter is en ogenschijnlijk smeekt om weerlegging. Als er vervolgens door systematische waarneming blijkt dat de hypothese inderdaad niet klopt, moeten wetenschappers die negatieve uitkomst accepteren en niet gaan zoeken naar alternatieve verklaringen voor de nulbevinding. In deze wijdverspreide visie treedt wetenschappelijke vooruitgang op door falsificatie. Daardoor worden veel niet-kloppende hypothesen ontmaskerd en wordt er incidenteel, door een mislukte falsificatiepoging, iets nieuws ontdekt. Een en ander neemt niet weg dat er tal van interessante hypothesen zijn te bedenken die zich niet kwalificeren als wetenschappelijke hypothese (zie Tabel 1).

Na het ontwikkelen van een deugdelijke theorie en het formuleren van toetsbare hypothesen, is een derde kenmerk van wetenschap dat er systematische observaties plaatsvinden ter toetsing van de hypothese. Er zijn verschillende onderzoeks*designs*, waarvan het experiment

bovenaan de hiërarchie staat. In een experiment wordt de onderzochte variabele gemanipuleerd in een deel van de observaties (de experimentele conditie) en vervolgens wordt het effect van die manipulatie vergeleken met observaties van overigens zo identiek mogelijke gevallen waarin de manipulatie niet heeft plaatsgevonden (de controleconditie). In deze vergelijking zit de falsificatie, want als er geen verschil is tussen beide condities, dan wordt geconcludeerd dat de hypothese (luidend dat de variabele wel een effect heeft) verworpen.

Tabel 1. Voorbeelden van onweerlegbare hypothesen.

1.	Sommige mensen kunnen met overledenen communiceren.
2.	Men kan geluk hebben.
3.	Het regent, of niet.
4.	De meeste psychische processen spelen zich af zonder dat we dat weten.
5.	De doodstraf is slecht.
6.	Er is leven na de dood.
7.	Mensen worden zeven keer gereïncarneerd.
8.	We worden omgeven door entiteiten die op geen enkele manier zijn waar te nemen.
9.	De meeste misdaden komen aan het licht.
10.	De meeste misdaden blijven onontdekt.

De aard van een experiment en de moeite die we soms hebben om die aard te erkennen en waarderen, laat zich goed illustreren met een onderzoek van Kuhn et al. (1985). Deze onderzoekers legden de volgende casus voor aan 130 jonge mensen in de leeftijdscategorie van 9 tot 31 jaar. “Last summer, I bought a new car. Around the time I bought it, I heard about a new product for cars that is supposed to help your car engine – it’s called EngineHelp. One of the things that the makers of EngineHelp said in their ads was that EngineHelp helps your car run well. I wanted my new car to run well, so I decided to use it. I was curious and decided to see what some other people’s experience was with EngineHelp, so I went out and talked to some people. Six people I talked to said they use EngineHelp and they all said their cars run well”. Vervolgens

vroegen ze of de proefpersoon dacht dat Enginehelp het functioneren van de auto bevordert. Niet minder dan 67% bevestigde deze vraag. Merk op dat het bewijs een serie gevalsbeschrijvingen is. Daarna kregen de deelnemers aanvullende informatie: "Two other people I talked to said that they use EngineHelp and that their cars run poorly". Met deze informatie wordt duidelijk dat 6 van de 8 Enginehelp-gebruikers tevreden is (75%). Opnieuw gevraagd naar hun mening, gaf 40% aan te geloven dat Enginehelp goed is voor de auto. Merk op dat in de informatie nog niet is gecontroleerd voor een mogelijke alternatieve interpretatie, bijvoorbeeld dat de tevredenheid niet wordt veroorzaakt door Enginehelp, maar simpelweg door de nieuwheid van de auto zelf. Daarom kregen de deelnemers de volgende informatie. "Three other people I talked to said that they never use EngineHelp and that their cars run well. And one last person I talked to said that he never uses EngineHelp and that his car runs poorly". Nu blijkt dat van de mensen die een nieuwe auto hebben en geen Enginehelp gebruiken, ook 75% tevreden is. Geconcludeerd moet worden dat Enginehelp toch niets toevoegt aan de tevredenheid. Niettemin geloofde nog steeds 40% van de deelnemers dat Enginehelp toch goed is voor je auto. Een deel van de proefpersonen (n = 66) kreeg alle informatie in een keer. Van hen was 52% overtuigd dat Enginehelp echt werkt.

Een goed design is noodzakelijk maar niet voldoende voor een goed onderzoek. Daarvoor zijn ook nog goed gekozen en valide gemeten variabelen en een deugdelijke statistische analyse vereist. In het algemeen geldt dat de instrumenten waarmee variabelen worden gemeten betrouwbaar (consistent) en valide (waarheidsgetrouw) dienen te zijn. Dat is niet altijd gemakkelijk in de psychologie. Neem het voorbeeld van de klanttevredenheid in het Enginehelp-onderzoek van Kuhn et al. (1985). Hoe meet een onderzoeker of de Enginehelp-gebruiker tevreden is over zijn auto? Kan hij dat eenvoudigweg vragen? Of

loopt hij dan het risico op sociaal wenselijke antwoorden? Zijn mensen consistent in hun tevredenheid, of zijn ze 's ochtends minder tevreden dan 's avonds? Is het antwoord op de vraag naar tevredenheid zuiver, of wordt het oordeel over de motor onbedoeld gekleurd door de waargenomen vriendelijkheid van de verkoper? Is het aantal gereden kilometers een goede maat voor tevredenheid? In een andere context is het de vraag of patiënttevredenheid iets zegt over de kwaliteit van de geneeskundige behandeling als die behandeling invasief is en ernstige bijwerkingen heeft (zoals chemotherapie). Zeggen studentevaluaties echt iets over de kwaliteit van het onderwijs, of zou het vruchtbaarder zijn om studenten na voltooiing van hun studie te vragen welke vakken hun het meest bijstaan?

Inmiddels is er een ongeschreven lijst met criteria waaraan meetinstrumenten idealiter voldoen opdat er deugdelijk mee kan worden gemeten. Deze vereiste psychometrische kwaliteiten zien toe op de betrouwbaarheid en validiteit van het instrument. Daarbij zij opgemerkt dat, hoewel in het dagelijks spraakgebruik begrippen als betrouwbaarheid, validiteit en geloofwaardigheid door elkaar kunnen worden gebruikt, dat niet geldt in psychometrisch jargon. In jargon kan een meetinstrument betrouwbaar zijn (het geeft consistent hetzelfde resultaat), maar toch niet valide (het meet niet wat we willen dat het meet). Wagenaar schreef daarover in een andere context: "Een getuige die steeds beweert dat hij door marsmannetjes is ontvoerd, is voor de psycholoog betrouwbaar, hoewel waarschijnlijk niet geloofwaardig" (1998, p. 50). Het ligt voor de hand dat alleen al dit soort vakjargon voor misverstanden kan zorgen en wellicht zelfs tot fouten kan leiden wanneer psycholoog en jurist samenwerken. Bekende maten voor betrouwbaarheid zijn factorstructuur (het instrument meet blijkens statistische analyse dat aantal concepten dat het geacht wordt te meten), Cronbach's *alpha* (items in het instrument hangen goed samen) en test-hertest betrouwbaarheid (als een

instrument met als meetpretentie een stabiel kenmerk, bijvoorbeeld intelligentie, twee keer wordt ingezet bij dezelfde persoon, komt er beide keren inderdaad hetzelfde resultaat uit). Validiteit vergt idealiter een samenhang (correlatie) met een gouden standaard (convergente validiteit). Tegelijkertijd moet het instrument niet samenhangen met variabelen waarmee men geen samenhang verwacht (divergente validiteit). Als het instrument een variabele meet waarvoor er reeds andere instrumenten zijn, mag men ook incrementele validiteit verwachten, dat wil zeggen dat het nieuwe instrument op de een of andere manier beter is dan de bestaande.

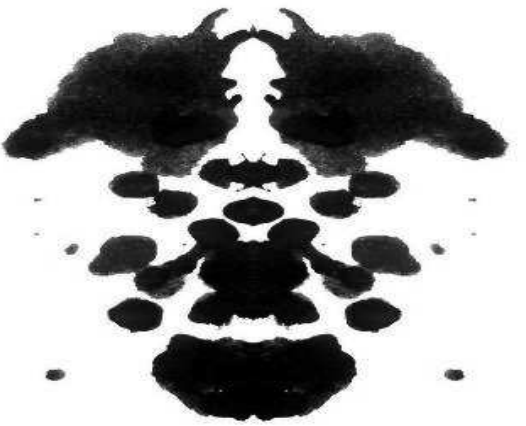
Zelfrapportages zijn een veelgebruikt meetinstrument, maar er kleven, psychometrisch bezien, nadelen en beperkingen aan. Er is dan ook nogal wat onderzoek verricht naar de vermeende onbetrouwbaarheid van zelfrapportages. Zo ontdekten Schwarz et al. (1991) dat de schaal waarop een antwoord wordt gegeven op de vraag “hoe succesvol bent u in het leven?” van invloed is op de verkregen informatie. Vierhonderd en tachtig respondenten beantwoordden deze vraag door een getal te omcirkelen tussen 0 (*helemaal niet succesvol*) en 10 (*extreem succesvol*). Vierendertig procent gaf zichzelf een onvoldoende. Een tweede groep van 552 respondenten kreeg dezelfde vraag, maar nu liep de antwoordschaal van -5 (*helemaal niet succesvol*) tot 5 (*extreem succesvol*). In dit geval durfde slechts 13% zichzelf een cijfer in de lagere regionen te geven. Volgens de onderzoekers duidt dit erop dat mensen zichzelf sneller een onvoldoende geven op een schoolcijferschaal dan op een schaal met negatieve getallen. Een ander voorbeeld. Strack et al. (1985) vroegen hun respondenten om aan te geven op een schaal van 1 tot 11 hoe gelukkig ze waren. Respondenten die voorafgaand aan deze vraag drie recente positieve gebeurtenissen hadden moeten opschrijven, scoorden hoger (8,9) op de geluksschaal dan respondenten die drie negatieve gebeurtenissen hadden moeten rapporteren (7,1). De

beoordeling van geluk verliep blijkbaar congruent met de, door de rapportage van drie emotioneel geladen gebeurtenissen, geïnduceerde stemming. Het rapporteren van drie emotioneel geladen gebeurtenissen uit het verre verleden had echter het tegengestelde effect. Respondenten die schreven over drie prettige gebeurtenissen in hun jeugd, voelden zich nu minder gelukkig (7,5) dan personen die drie negatieve jeugdherinneringen hadden opgehaald (8,5). Waarschijnlijk overtuigden de drie fijne jeugdherinneringen de respondenten ervan dat vroeger alles leuker en beter was en dat ze nu dus wel relatief ongelukkig moesten zijn (een zogenoemd contrasteffect; zie voor een overzicht van zelfrapportageperikelen, Nisbett & Wilson, 1977; Schwarz, 1999).

In de psychologische literatuur zijn er naast zelfrapportages andere soorten variabelen te vinden die worden gemeten met andersoortige instrumenten. De verschillen worden deels ingegeven door het bestaan van concurrerende theoretische scholen. Psychodynamici veronderstellen bijvoorbeeld dat zelfrapportages veelal onbruikbaar zijn, omdat mensen niet weten wat hun onbewuste drijfveren zijn. Daarom nemen zij hun toevlucht tot indirecte metingen (projectieve tests), waarvan de Rorschachtest mogelijk het bekendste voorbeeld is (zie Figuur 2). Bij de Rorschachtest dient de onderzochte persoon te vertellen wat hij ziet in de abstracte inktvlekken. Op basis van dat narratief kan de diagnosticus onbewuste processen blootleggen. Deze test vergt veel interpretatie van de psycholoog en daarom is de test volgens sommigen weinig betrouwbaar en (dus) ook weinig valide (Garb et al., 2005).

Behaviouristen zijn vooral gecharmeerd van observaties. Net als projectieve tests en alle andere psychologische meetinstrumenten, zijn observaties feilbaar. Observatie lost het probleem van de gemankeerde zelfrapportage op, maar brengt daarvoor in de plaats de mogelijke onbetrouwbaarheid van de observatie mee. Er is

nogal wat onderzoek waaruit blijkt dat deskundigen alleen dan zinnige observaties plegen wanneer ze kunnen beschikken over gestandaardiseerde meetinstrumenten. Faust en Ziskin concludeerden dat "professionals often fail to reach reliable or valid conclusions and that the accuracy of their judgments does not necessarily surpass that of laypersons" (1988, p. 31). Vijftien jaar later gold deze conclusie in essentie nog steeds: "One would assume that clinical and counselling psychologists are more accurate than psychology graduate students. However, with few exceptions, this difference has not been found" (Garb & Boyle, 2003, p. 20).



Figuur 2. Voorbeeld van een fictieve Rorschach inktvlek.

Een bekend probleem met observaties is dat mensen geneigd zijn om het gedrag van anderen sneller te wijten aan interne factoren (zoals de persoonlijkheid van de geobserveerde), terwijl men het eigen gedrag (met name het slechte) eerder toeschrijft aan situationele omstandigheden (zie Malle et al., 2007). Er is geen aanwijzing dat deskundigen immuun zijn voor deze zogenaamde *actor-observer bias*. In een overzichtsartikel concludeerden Grove et al. (2000) dat clinici die afgaan op hun eigen observaties het in 6% tot 16% van de gevallen

beter doen dan clinici die vertrouwen op tests. Daar staat tegenover dat dezelfde onderzoeken aantonen dat het gebruik van tests leidt tot betere diagnostiek, vergeleken bij klinische observatie, in 33% tot 47% van de gevallen. Het lijkt er dus op dat tests in het algemeen iets beter werken dan observaties. In elk geval geldt dat de betrouwbaarheid van een observatieschaal vergt dat als deze schaal wordt gebruikt door diagnosticus A dezelfde uitkomst wordt gegenereerd als wanneer diagnosticus B de test hanteert bij dezelfde observandus. Dit kenmerk wordt aangeduid met de term interbeoordelaarsbetrouwbaarheid.

Cognitief psychologen zullen een voorkeur hebben voor prestatietests, waarvan de IQ-test een bekend voorbeeld is. Het verschil met zelfrapportages is dat de respondent niet wordt gevraagd om zichzelf te beschrijven of beoordelen, maar dat hij daadwerkelijk een prestatie moet leveren. Aan de hand van die prestatie wordt een conclusie getrokken over de betreffende variabele (in dit voorbeeld intelligentie).

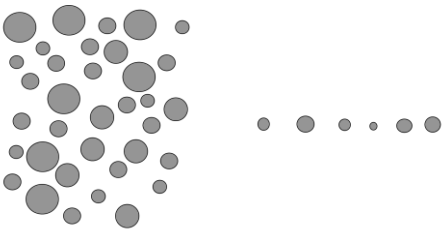
Neurowetenschappers, tot slot, maken gebruik van fysiologische metingen. Maar ook die maten hebben hun beperkingen. Uitdraaien van *event-related potentials* (ERP's) en fMRI-beelden zijn niet alleen gebaseerd op tal van aannames en afspraken, de interpretatie ervan laat eveneens ruimte voor discussie (Vul et al., 2009). Desondanks lijkt het erop dat het gebruik van fysiologische variabelen nogal indruk maakt op de lezer. Weisberg et al. (2008) gaven aan proefpersonen een verklaring voor een bepaald fenomeen inhoudende dat "mensen meer fouten maken als ze de kennis van anderen moeten schatten dan wanneer ze een oordeel vellen over hun eigen kennis" (p. 471). Een andere groep deelnemers kreeg dezelfde verklaring, maar dan met de toevoeging dat een bepaald gebied in de frontaalkwab van belang is voor de beoordeling van kennis. Alle deelnemers dienden hun tevredenheid met deze verklaring weer te geven door een getal te omcirkelen tussen -3 (*zeer onbevredigend*) en 3 (*zeer bevredigend*).

Terwijl mensen in de eerste conditie gemiddeld -0,73 scoorden, was de tevredenheid in de groep met biologisch substraat 0,16. Een andere kanttekening is dat hoewel fysiologische maten op het eerste gezicht wellicht sterker en objectiever lijken dan zelfrapportages, dat welbeschouwd niet zo is. McNally (2001) zegt daarover: "…certain phenomena, such as obsessions, have no outward manifestation other than that revealed in language. Even PET studies of OCD patients require one to confirm that the person is, indeed, obsessing in the scanner" (p. 520).

Kortom, psychologen hebben de beschikking over een breed arsenaal aan meetinstrumenten. Tegelijkertijd moet worden onderkend dat psychologische instrumenten zelden of nooit de accuratesse bezitten van een weegschaal of meetlat. De sensitiviteit is zelden 100% (ofwel: er treden bijna altijd *misses* of vals negatieven op) en de specificiteit/selectiviteit is evenmin 100% (er treden bijna altijd vals positieven op). Het is niet altijd gemakkelijk om de psychometrische kwaliteiten van een psychologisch meetinstrument uit de losse pols te schatten. Forer (1949) beschreef een in dit kader relevant onderzoek waarin hij 39 studenten een test liet invullen en hun vervolgens confronteerde met de uitkomsten. Zo gaf hij hen individueel een persoonlijkheidsprofiel waarin beschrijvingen zaten zoals "Security is one of your main goals in life" en "You have found it unwise to be too frank in revealing yourself to others". In werkelijkheid waren dit algemene en daardoor nietszeggende frases uit een horoscoop. Forer had alle studenten precies dezelfde beschrijving gegeven. Niettemin dachten zijn studenten dat deze persoonsbeschrijvingen zeer accuraat en uniek waren. Dit *Forer-effect*, ook wel *P.T. Barnum-effect* genoemd, onderstreept het belang van validatie van meetinstrumenten (Meehl, 1956). Zo gauw mensen, inclusief psychologen, zich overgeven aan persoonlijke indrukken over de kwaliteit van een meetinstrument, is er geen garantie dat die indrukken

accuraat zijn. Lilienfeld et al. (2006) betogen dat naast het Forer-effect, ook de *alchemist's fantasy* (de overtuiging dat meerdere onbetrouwbare metingen samen toch nog een betrouwbaar eindresultaat kunnen opleveren), de *ad antiquitem fallacy* (de overtuiging dat een test goed is, alleen al omdat hij al lang in gebruik is) en de *ad populum fallacy* (de overtuiging dat een test goed is, alleen al omdat veel mensen hem gebruiken) debet eraan zijn dat er in de psychologische praktijk veel onvoldoende gevalideerde tests circuleren.

Het gebruik van meetinstrumenten van voldoende psychometrisch niveau maakt het mogelijk om de metingen statistisch te analyseren. Stel dat ik de tuinschuur opruim en daarbij twee zakjes met bloemzaadjes tegenkom. Een zakje is behoorlijk gevuld, maar in het andere zakje zitten maar zes zaadjes. Ik overweeg om beide zakjes samen te voegen, maar wil dat alleen doen als het zaadjes van dezelfde plant zijn. Helaas verschillen de zaadjes nauwelijks. Het enige waar ik op kan letten, is de omvang van de zaadjes. Stel dat ik weet dat de zaadjes in het volle zakje (zakje 1) inderdaad van dezelfde plant zijn en dat ik de diameter van elk zaadje goed, betrouwbaar kan meten. Ik neem ook aan dat de zaadjes in zakje 2 van eenzelfde soort zijn, maar de vraag is dus of dat dezelfde plant is als die waarvan de zaadjes in zakje 1 afkomstig zijn. Ik strooi de inhoud van beide zakjes voorzichtig uit op tafel en zie dan iets wat lijkt op Figuur 3.



Figuur 3. Inhoud van fictieve zakjes met bloemzaadjes.

Het valt op dat de zaadjes van eenzelfde soort toch enigszins verschillen in grootte. Er zijn met andere woorden individuele verschillen binnen de twee groepen. Ook valt op dat de zes zaadjes uit zakje 2 iets kleiner lijken. Na meting van alle zaadjes, blijkt de gemiddelde diameter van de zaadjes in zakje 1 vijf millimeter. De gemiddelde diameter in zakje 2 is drie millimeter. De verschillen in grootte tussen alle zaadjes in zakje 1 leveren een standaard deviatie (een soort gemiddelde spreiding rondom het gemiddelde) van 0,5 millimeter op. De spreiding in zakje 2 is kleiner (0,1 millimeter). Er zijn wiskundige formules (en software) om deze metingen statistisch te analyseren. Die analyses maken gebruik van de gemiddelden en de standaard deviaties, en geven een antwoord op de vraag of het verschil (5 mm versus 3 mm) verwaarloosbaar is, of juist statistisch significant is.

Als wetenschapper zou ik nu zeggen dat mijn nulhypothese is dat er geen verschil is tussen de zaadjes in zakje 1 en die in zakje 2. De alternatieve hypothese luidt juist dat er wel een verschil is. Een statistische toets levert een *likelihood* op die weergeeft hoe groot de kans is om deze metingen te doen (5 mm in zakje 1 en 3 mm in zakje 2) als er in werkelijkheid geen verschil is tussen de zaadjes uit beide zakjes. Als de likelihood groot is, concludeer ik dat de zaadjes op een hoop kunnen worden geveegd (er is geen verschil). Als de likelihood klein is (kleiner dan 5%), concludeer ik dat het verschil statistisch significant is en zal ik de zaadjes niet in hetzelfde zakje doen. De likelihood wordt aangeduid met een *probability*-waarde (*p*-waarde). Bij nader inzien is hier iets vreemds aan de hand. De statistische analyse zegt bijvoorbeeld dat het waarschijnlijk is ($p = 0{,}80$) om deze data te vinden, indien er geen verschil is tussen steekproef 1 en steekproef 2. En daarom concludeer ik dat er dus (waarschijnlijk) inderdaad geen verschil is. Het vreemde is dat de likelihood van het vinden van de data niet per se uitsluitsel geeft over de correctheid van de hypothesen. Het gevaar van deze denkstap kan als

volgt worden geïllustreerd. Stel dat de kans dat een hond vier poten heeft, 99% is (sommige honden missen een poot). Gegeven deze stelling is het nog niet (per se) logisch correct om te concluderen dat als we een dier met vier poten tegenkomen, dat met 99% waarschijnlijkheid een hond is.

Er is nog iets opmerkelijks. Stel dat de uitkomst van de analyse is dat *p* 0,80 is. Oftewel: als er geen verschil is tussen de zakjes, is de kans om de zaadjes te vinden zoals ik die vond best groot, namelijk 80%. Maar dit zegt nog niets over hoe groot de kans is om de zaadjes te vinden als er in werkelijkheid sprake is van twee verschillende soorten zaadjes (de alternatieve hypothese is correct). Stel dat als de twee zakjes inderdaad verschillende zaadjes bevatten, de kans om te vinden wat ik vond 90% is, dan is mijn vondst waarschijnlijker onder de alternatieve hypothese dan onder de nulhypothese. En gemakshalve zou ik dan concluderen dat de kans groter is dat de zakjes wel verschillen. In deze gedachtegang moet ik niet alleen rekening houden met de kans op de waarneming onder de nulhypothese (de *p*-waarde), maar ook met de kans onder de alternatieve hypothese. Door alleen op de *p*-waarde af te gaan, doe ik half werk. Door de waarschijnlijkheid van de waarneming te schatten onder de nul- en de alternatieve hypothese, kan ik concluderen bij welke van die twee mijn resultaten het best passen. Deze benadering wordt de likelihood ratio benadering genoemd (Royall, 1997). Die benadering past in de zogenoemde Bayesiaanse benadering.

Het theorema van Bayes is vernoemd naar de Engelse predikant Thomas Bayes (1702-1761). Er is niet veel bekend over Bayes. De enige tekst van hem die bewaard is gebleven is een theoretische verhandeling over kansen. Een vriend van Bayes, Richard Price, vond dit opstel tussen Bayes' spullen enige tijd na diens overlijden. Price werkte de tekst van Bayes verder uit en stuurde hem naar John Canton, die het geheel publiceerde in

Philosophical Transactions of the Royal Society of London. In zijn tekst gaat Bayes (1763) in op de structuur van een kansspel met indirecte waarneming. Stel dat er een bal op een tafel ligt, maar de vraag is waar hij precies ligt. Stel verder dat we de tafel niet kunnen zien, maar dat een assistent in dit gedachte-experiment dat wel kan. We vragen nu aan de assistent om nog een bal op de tafel te gooien en – als die bal eenmaal tot stilstand is gekomen – om ons mede te delen of deze tweede bal links of rechts naast de oorspronkelijke bal ligt. Stel dat de assistent ons mededeelt dat de tweede bal links van de oorspronkelijke bal ligt, dan duidt dat wellicht erop dat de oorspronkelijke bal ietwat rechts van het midden op de tafel ligt. Als de volgende bal weer links van de oorspronkelijke bal ligt, bevestigt dit de hypothese dat de oorspronkelijke bal inderdaad behoorlijk ter rechter zijde ligt. Mocht er na een aantal *trials* ook een bal rechts van de oorspronkelijke bal terechtkomen, dan weten we dat de oorspronkelijke bal niet helemaal rechts op de rand kan liggen. Op deze manier kunnen we door indirecte waarneming conclusies trekken over de locatie van de bal, zonder dat we die kunnen waarnemen.

Dit idee werd jaren later uitgewerkt door Pierre Simon Laplace (1749-1827) tot wat wij nu kennen als het theorema van Bayes. Daarom zou die benaming volgens McGrayne (2011) eerlijkheidshalve beter Bayes-Price-Laplace kunnen luiden. Hoe dan ook leert de bal-op-de-tafel analogie ons enkele cruciale kenmerken van de Bayesiaanse visie. Behalve dat we door indirecte waarneming conclusies kunnen trekken over een onderzoeksobject, kan de dataverzameling gefaseerd plaatsvinden: elke nieuw op tafel gegooide bal levert nieuwe informatie op die onze mening over de locatie van de eerste bal beïnvloedt. De Bayesiaanse aanpak is een *belief updating* systeem. Verder is het zo dat onze mening bij aanvang van het gedachte-experiment een grote invloed heeft op de einduitkomst. Stel bijvoorbeeld dat we om