

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

© Eric Rassin 2026

INHOUD

Inleiding	1
Wat is een wetenschappelijk experiment?	1
Wat mis kan gaan bij wetenschappelijk onderzoek	22
Experimenteren bij juridische besluitvorming	29
The proper seat	39
Het herkennen van Iwan	51
The memory of concentration camp survivors	75
Crashing memories	95
False Confessions	105
Symptom validity testing	129
The big-effect-big-cause heuristic	135
Abducted by a UFO	143
The use-the-best heuristic	155
Conclusie	185
Literatuur	187

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 kennismakt 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: kennismaken 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.

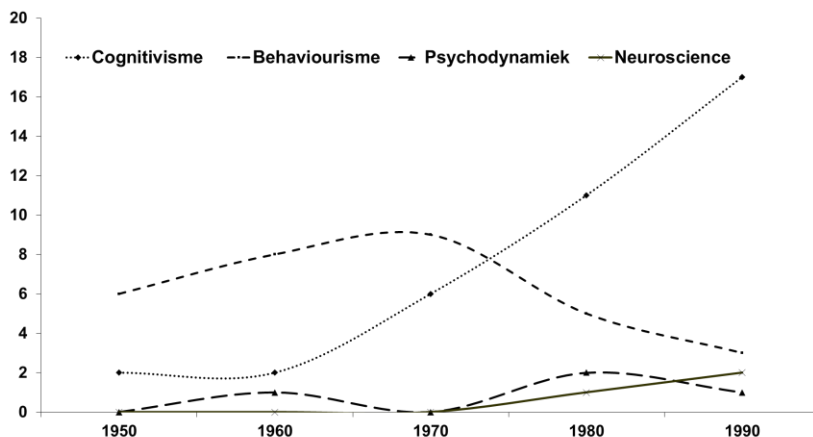
Een goed experiment kent tenminste de volgende aspecten: een theoretisch uitgangspunt, een hypothese op basis van die theorie, een goed design voor een systematische waarneming, een statistische analyse, een interpretatie en discussie van de resultaten, en de publicatie van het onderzoek. In het navolgende 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 gezien, 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 klinici 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).

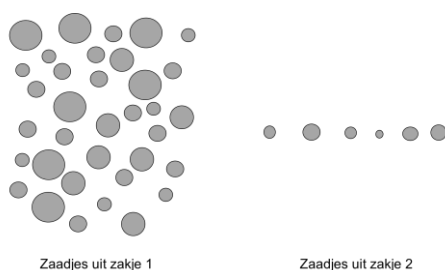
Neurowetenschappers, tot slot, maken gebruik van fysiologische metingen. Maar ook die maten hebben hun beperkingen. Uitdraaien van *event-related potentials* (ERP's) en fMRI-beelden zijn niet alleen gebaseerd op tal van aannames en afspraken, de interpretatie ervan laat eveneens ruimte voor discussie (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 uitsluitel 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

welke reden dan ook vooraf sterk vermoeden dat de bal rechts op de tafel ligt. In dat geval zal de bevinding dat de eerste bal links van de oorspronkelijke bal is gevallen ons sterker doen geloven dat de oorspronkelijke bal inderdaad rechts ligt, dan wanneer we geen uitgesproken mening vooraf over het rechts liggen hadden gehad. De schatting vooraf van de werkelijkheid wordt in de Bayesiaanse benadering *prior odds* genoemd (a priori kansverhouding). Deze a priori kans wordt vervolgens aangepast vanwege de nieuwe informatie (de dataverzameling), en als alle relevante informatie in het model is verwerkt, is er sprake van een *posterior odds* (a posteriori kansverhouding), de uiteindelijke schatting van de werkelijkheid.

Voor ons Nederlanders is het wellicht ietwat onwennig om te denken in termen van kansverhoudingen. Wij zijn meer gewend aan kansen. Er is een manier om terug te rekenen van kansverhouding naar kans. Die luidt als volgt. De kans dat de nulhypothese waar is, is ($\text{odds} * 100$) gedeeld door ($\text{odds} + 1$).

Een voorbeeld waarmee de bruikbaarheid van de regel van Bayes zich goed laat illustreren is het volgende. Er is een ongeval gebeurd waarbij een taxi betrokken was. De taxichauffeur is vervolgens doorgereden. Er is een getuige, mevrouw Prob, die verklaart dat de taxi blauw was. Stel nu dat we ons afvragen of de betreffende taxi inderdaad blauw is. Zonder kwantificatie komt het antwoord dan aan op de keuze om mevrouw Prob te geloven of niet. Stel echter verder dat uit talloze experimenten blijkt dat mevrouw Prob in 80% van de gevallen in staat is om onder vergelijkbare omstandigheden (het was al schemerig) een blauwe taxi te onderscheiden. Nu zou verdedigd kunnen worden dat de kans dat de taxi blauw was 80% is. De likelihood ratio van de verklaring van mevrouw Prob is 4 (80% hits gedeeld door 20% vals positieven). Stel tot slot ook nog dat er in de betreffende stad slechts twee soorten taxi's zijn: blauwe en groene. Van al die taxi's is 15% blauw en 85% groen. Nog steeds is de vraag hoe groot de kans is

dat de bij het ongeval betrokken taxi blauw is. Is die kans nog steeds 80%, of moet de lage *base rate* van 15% worden verdisconteerd? In de Bayesiaanse aanpak is het antwoord op de allerlaatste vraag bevestigend. De lage base rate vormt namelijk de prior odds. Als het percentage blauwe taxi's en de accuratesse van mevrouw Prob de enige relevante bronnen van informatie zijn, wordt de kans dat de taxi blauw is als volgt Bayesiaans berekend. De prior odds is 15% gedeeld door 85%: 0,18. Deze prior odds wordt vermenigvuldigd met de likelihood ratio van Prob's accuratesse (4). De posterior odds is dan 0,72. Als we deze uitkomst terugrekenen naar een kans, dan volgt dat de kans dat de taxi blauw is 42% is - en de kans dat-ie groen is dus 58%. Dit voorbeeld maakt duidelijk dat de prior odds in de Bayesiaanse analyse een cruciale rol spelen, net zoals de likelihood ratio van een bewijsmiddel.

Een ander voorbeeld. Stel dat er een test bestaat voor een ziekte. Die test is behoorlijk betrouwbaar: als je deze ziekte hebt, zal de test dat altijd uitwijzen. Er is echter ook een kleine kans dat als je de ziekte niet hebt, de test abusievelijk toch aangeeft dat je ziek bent. Deze vals positief kans is 5%. Stel verder dat het gaat om een veel voorkomende ziekte: 25% van de mensen heeft hem. Een man gaat naar de dokter en ondergaat deze test. De uitslag is positief. Hoe groot is de kans dat deze man de betreffende ziekte inderdaad heeft? In de Bayesiaanse benadering geldt het volgende. De prior odds is 0,33 (25% : 75%). De likelihood ratio op grond van de testuitslag is 20 (100% : 5%). De posterior odds is 6,67. De kans dat de patiënt de ziekte heeft is dan 87%.

Stel nu dat het gaat om een minder prevalentie ziekte, een ziekte die slechts 1% van de mensen treft. Wat is dan de kans om de ziekte daadwerkelijk te hebben, als je van de dokter te horen krijgt dat de testuitslag positief was? De likelihood ratio op grond van de test is nog steeds 20. De prior odds is nu echter 0,01 (0,01 : 0,99). De posterior

odds is 0,2. Omrekenen naar de kans om de ziekte te hebben, levert een kans van 17% op.

Het klinkt ongeloofwaardig dat de kans op het hebben van een bepaalde ziekte slechts 17% is als je net van de dokter te horen hebt gekregen dat je die ziekte hebt, nota bene op grond van een betrouwbare test. Deze frictie wordt uitsluitend veroorzaakt door de lage a priori kans. In dit voorbeeld is de lage a priori kans het gevolg van de lage base rate. Maar het wordt nog bizarder als de a priori kans voor eenzelfde ziekte gaat verschillen tussen mensen vanwege hun persoonlijke geschiedenis. Als twee mannen vanwege dezelfde test te horen krijgen dat ze HIV-positief zijn, is de kans dat de ene man inderdaad besmet is dan groter, omdat hij meermaals onveilige seks heeft gehad? Mag/kan/moet er onderscheid worden gemaakt in het quarantainebeleid tussen twee mensen die Ebola (b)lijken te hebben (althans volgens de test), terwijl een van hen in een risicogebied is geweest en de andere niet? In een Bayesiaanse analyse wordt deze vraag bevestigend beantwoord. De analyse is dan ook strikt empirisch en dat druist soms in tegen ons gevoel van politieke correctheid. In de eeuwen na Bayes is diens visie niet alleen door navolgers verfijnd, maar zijn er juist ook tegenstanders opgestaan. Zo zijn sommigen van mening dat wetenschappers principieel geen a priori overtuiging moeten hebben wanneer ze data verzamelen. Een a priori kans is bovendien in sommige gevallen moeilijk te bepalen en dus subjectief. Verder vinden sommigen een conclusie van het type "Hypothese 1 is 24 keer zo waarschijnlijk als hypothese 2" onduidelijk en onpraktisch. Vandaar dat mensen als Gosset, Pearson en Fisher hebben toegewerkt naar frequentistische methoden, zoals ANOVAs en *t*-tests met bijbehorende *p*-waarden, die sommigen nu de "klassieke" statistiek noemen (Hordijk & Sjerps, 2005).

Uit bovenstaande uiteenzetting volgt dat mijn probleem met de zakjes zaadjes niet zo complex is. Stel dat ik geen a prior menig heb (de prior odds is dan 1). De

likelihood ratio van mijn observatie is 80% gedeeld door 90%. De likelihood ratio is dus 0,89. De likelihood ratio wordt soms dan weer de BF_{01} genoemd: de Bayes factor die aangeeft hoeveel waarschijnlijker de data zijn onder de nulhypothese vergeleken bij de alternatieve hypothese. Omdat die in dit geval kleiner is dan 1, passen de data dus beter bij de alternatieve hypothese. De BF_{01} kan ook worden gespiegeld tot een BF_{10} . Die geeft aan hoeveel beter de data passen bij de alternatieve hypothese. In dit geval is dat 90% gedeeld door 80% is 1,13. In alle gevallen ligt de uitkomst ronde de 1. Omdat ik geen rekening houd met een prior odds (die stelde ik ook op 1), is de conclusie dat de data even waarschijnlijk zijn onder beide hypothesen. Wellicht valt op dat de Bayesiaanse formulering weliswaar genuanceerder maar inderdaad ook minder duidelijk is dan de frequentistische (de enkelvoudige p -waarde). Los daarvan is niet zo duidelijk waar de kritische grens ligt om te concluderen of de nulhypothese of de alternatieve hypothese juist is. Bij de p -waarde is dat gemakkelijker. Daar ligt de grens veelal bij 0,05. Bij het gebruik van likelihoods is er meer nuance. In Tabel 2 staat een overzicht van mogelijke kritische waarden.

Tabel 2. Likelihood ratio marges (Jarosz & Wiley, 2014).

Verbale omschrijving	Likelihood ratio
Weak/anecdotal support	1 - 3
Positive/substantial support	3 - 10
Positive/strong support	10 - 30
Strong/very strong support	30 - 100
Very strong/decisive support	100 <

Een laatste kenmerk van wetenschap is gelegen in de manier waarop wetenschappelijke bevindingen worden gepubliceerd. Wetenschappers publiceren hun resultaten bij voorkeur in een van de vele internationale *peer reviewed journals*. Dat zijn tijdschriften die een streng selectiebeleid voeren ten aanzien van artikelen die zij publiceren. Als de redactie een manuscript van een wetenschapper ontvangt,

zal de redactie eerst zelf dat stuk lezen, om te bezien of het geschikt is voor publicatie. Maar na deze eerste globale check, sturen ze het manuscript naar twee of drie wetenschappers waarvan bekend is dat ze verstand hebben van het onderwerp van het manuscript. Deze wetenschappers (*peers* van de auteurs) wordt verzocht streng kritiek te leveren op het manuscript. Om te bewerkstelligen dat de beoordelaar inderdaad onbevooroordeeld en streng zijn, vindt de beoordeling liefst dubbelblind plaats. Dat betekent dat de beoordelaars niet weten wie de auteurs van het manuscript zijn en dat de auteurs wel het commentaar van de beoordelaars zullen zien, maar niet hun identiteit zullen weten. De review van de peers blijkt in de praktijk vaak inderdaad heel streng. Bij sommige tijdschriften wordt slechts 10% van alle ter publicatie aangeboden manuscripten daadwerkelijk gepubliceerd. De rest wordt afgewezen. Dan moet de onderzoeker opnieuw beginnen.

Perfect uitgevoerde wetenschap is een proces van ontdekken van empirische wetmatigheden waarbij de kans op een vals-positieve bevinding geminimaliseerd is. Maar die kans is niet nul. Wetenschap is feilbaar en er zullen altijd nieuwe ontdekkingen zijn die eerdere tegenspreken. Los daarvan wordt in de praktijk van de wetenschap regelmatig een korte bocht genomen, waardoor de kracht van veel bevindingen niet maximaal is. In zoverre dat wetenschap mensenwerk is, gaat het bij alle cruciale aspecten mis.

Wat mis kan gaan bij wetenschappelijk onderzoek

Als wetenschap wordt gedefinieerd als een gedegen conservatief proces van waarheidsvinding, kunnen er wetenschappelijke fouten worden gemaakt in dat proces, op meerdere punten. Omdat er een groot volume aan wetenschappelijke output is, zijn er van allerlei soorten fouten voorbeelden te vinden. Om te beginnen wagen wetenschappers zich soms aan hypothesen die niet toetsbaar, althans niet falsificeerbaar zijn. Zo is er

onderzoek gedaan naar de hypothese dat er leven na de dood is (Van Lommel et al., 2001), maar ook naar de hypothese dat Sigmund Freud als kind is misbruikt (De Vries, 1993).

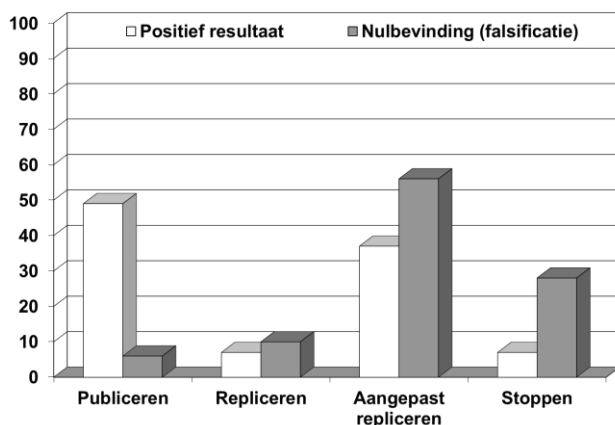
Met enige regelmaat verschijnt er een publicatie over een onderzoek waarin is vergeten om een adequate controlegroep te includeren. Daardoor worden verbanden gedetecteerd die er in werkelijkheid niet per se zijn (Lewis et al., 1997; Reinders et al., 2003). Dat honderd mensen zich beter voelen na een bepaalde therapie zegt niet dat die therapie werkt. Als honderd moordenaars allemaal zelf zijn mishandeld in hun jeugd, betekent dat niet dat mishandeling een voorspeller van moord is. Als duizend verkrachters allemaal psychopaat zijn, betekent dat niet dat psychopathie zedendelinquentie voorspelt. Dit soort conclusies wordt pas mogelijk als er controlegroepen zijn bestudeerd van mensen die een andere of geen therapie ontvangen, van mishandelde mensen die geen moord plegen, en van psychopaten onder niet-zedendelinquenten.

Wat de statistische analyse betreft, moet allereerst de meest passende techniek worden gebruikt. Daarnaast moet worden onderkend dat er altijd een (kleine) kans is op een vals-positieve uitkomst. Dat wil zeggen dat de onderzoeker concludeert dat de alternatieve hypothese klopt (er is een effect), terwijl er in werkelijkheid geen effect is. De kans op deze type 1 fout wordt groter naarmate de onderzoeker de analyse vaker herhaalt, wellicht in de hoop om het veronderstelde effect inderdaad te vinden. Door vaak eenzelfde analyse te doen, met steeds een kleine kans op een vals-positieve uitkomst, wordt die kans op een gegeven moment bewaarheid. Op deze manier zijn er potentieel vals-positieve resultaten geboekt, zoals de bevinding dat mensen paranormale gaven bezitten (Bem, 2011), dat bidden voor naasten helpt bij hun herstel na een medische ingreep (Harris et al., 1999) en dat er geheime boodschappen in de Bijbel zitten waarin de toekomst wordt voorspeld (Witztum et al., 1994; Thomas, 1997).

Opgemerkt zij dat het vissen naar positieve resultaten niet altijd onwetend gebeurt. Soms doen onderzoekers dat doelbewust en wordt er gesproken van *p-hacking*. Positieve resultaten kunnen bijvoorbeeld ook worden behaald door data te manipuleren (Chambers, 2017; Simmons et al., 2011).

De drang tot het vinden van positieve resultaten en daarmee het ontdekken van nieuwe wetmatigheden, heeft ook invloed op wat er uiteindelijk in de wetenschappelijke tijdschriften wordt gepubliceerd. Er lijkt namelijk een bias te zijn in het voordeel van positieve resultaten ten opzichte van nulbevindingen. Het peer review systeem beschermt niet tegen deze publicatiebias. Gevreesd moet worden dat onderzoekers niet hun best doen om al hun bevindingen te publiceren. Greenwald (1975) zette een vragenlijst uit onder 75 onderzoekers. De eerste vraag informeerde naar de persoonlijke overtuiging van de respondenten met betrekking tot onderzoek. Waren de respondenten ervan overtuigd dat ze een verschil zouden vinden, of dat er juist geen verschil zou worden gevonden, wanneer ze aan een nieuw onderzoek begonnen? Voor 80% gold dat ze een verschil verwachtten. Blijkbaar is men meer gemotiveerd om verschillen te vinden dan overeenkomsten. Vervolgens vroeg Greenwald wat ze zouden doen als een onderzoek inderdaad het verwachte verschil aan het licht zou brengen. Opties waren: opschrijven en ter publicatie aanbieden, eerst het onderzoek precies proberen te repliceren, eerst een aangepaste versie van het onderzoek doen, of stoppen met dit onderzoeksonderwerp. Tot slot stelde hij dezelfde vraag, maar dan met betrekking tot de vervolgactie in geval van een negatieve bevinding. De percentages aangevinkte antwoorden worden weergegeven in Figuur 4. Zoals te zien, hadden respondenten inderdaad meer zin om hun bevinding op te schrijven als "er iets uit hun onderzoek was gekomen". Als er "niets" uit was gekomen, was de kans groter dat het probleem in de koelkast belandde. Dit is faliekant in strijd met het falsificatieprincipe zoals Popper

(1959, 1963) erover schreef. Volgens dat principe is falsificatie (dat wil zeggen geen effect vinden en de nulhypothese dus handhaven) juist cruciaal. Als de nulhypothese wordt bevestigd, is dat dus eigenlijk ook geen "nul"bevinding. Het onderzoek is dan niet mislukt. Integendeel, een statistisch niet-significante uitkomst is een teken dat het verwachte effect er niet is, en dat de verwachting dus incorrect was. Toch lijken negatieve uitkomsten velerlei speculaties uit te lokken over waarom het verwachte effect niet is gevonden. Die speculaties betreffen dan veelal tekortkomingen in het onderzoek.



Figuur 4. Reacties van onderzoekers in geval van bevestiging en weerlegging (Greenwald, 1975).

Redacteuren en reviewers lijken ook aan ten prooi te vallen aan de weerzin tegen negatieve bevindingen. Coursol en Wagner (1986) vroegen aan 1.000 leden van de *American Psychological Association* of ze wel eens een therapie-effectstudie hadden gedaan. In totaal bleken 99 respondenten tezamen 194 van dergelijke evaluatieonderzoeken te hebben verricht. Vervolgens vroegen de auteurs of de respondenten hun bevinding ter publicatie hadden aangeboden. Van de 129 evaluaties waarin bleek dat de patiënt na behandeling beter was

geworden, werden er 106 (82%) ter publicatie aangeboden. Van de 65 onderzoeken waarin de behandeling geen effect bleek te hebben, werden er slechts 28 (43%) aangeboden. Blijkbaar is de schaarste van artikelen over negatieve bevindingen inderdaad deels te wijten aan de onderzoekers. Maar vervolgens werd gevraagd hoe de tijdschriften hadden gereageerd op de betreffende manuscripten. Van de 106 significante manuscripten werden er 85 (80%) geaccepteerd. Van de 28 non-bevindingen werden er slechts 14 (50%) geaccepteerd. Dit duidt inderdaad op een beleid van tijdschriften gericht op de publicatie van nieuwe ontdekkingen. Het gevolg van dit alles is dat het beeld in de literatuur niet strookt met dat in het lab. Immers, de verhouding tussen positieve resultaten en non-bevindingen was volgens de respondenten 129 : 65 (ofwel: 66% van de onderzoeken leverde een significant resultaat op). In de gepubliceerde literatuur was de verhouding echter 85 : 14 (ofwel: 86% van de publicaties duidde op een significant effect). Blijkbaar worden negatieve bevindingen in zekere zin gediscrimineerd. Deze publicatiebias wordt ook wel aangeduid met de term *file drawer bias*, omdat een deel van de onderzoeken, met name de negatieve bevindingen, achterblijft in de bureaulades van de onderzoekers (Rosenthal, 1979).

Los van de publicatiebias functioneert het peer review systeem volgens sommigen niet altijd zoals het aanvankelijk bedoeld was. Hoewel reviewers niet weten wiens manuscript ze beoordelen, gebeurt het in de praktijk toch dat ze kunnen achterhalen wie de auteurs zijn. Bijvoorbeeld omdat ze de schrijfstijl herkennen, of omdat ze weten wie de afgelopen tijd bezig is geweest met dit onderzoek. Dan kan het gebeuren dat een reviewer bij zijn beoordeling persoonlijke motieven laat meespelen. Als hij de auteur goedgezind is, zal hij wellicht wat milder zijn, maar als hij de auteur geen warm hart toedraagt, zal hij wellicht juist extra streng zijn. Het kan natuurlijk sowieso gebeuren dat een reviewer manuscripten die in strijd zijn

met zijn persoonlijke theorieën strenger beoordeelt dan nodig, of dat de reviewer juist onvoldoende kritisch is (Sokal, 1996a, b).

Een laatste cluster aan fouten die kunnen worden gemaakt in de wetenschap draait om de interpretatie van bevindingen en breder om verwachtingen van en meningen over wetenschap. Allereerst zij opgemerkt dat wetenschap zoals hiervoor beschreven overzichtelijk is. De crux is dat bij het bestuderen van de empirie regels moeten worden gevolgd waardoor de kans op een vals positieve conclusie wordt geminimaliseerd. Wetenschappelijk onderwijs is in deze visie eveneens overzichtelijk: dat draait erom studenten te leren hoe ze op een wetenschappelijke manier data kunnen verzamelen. Dat doet de docent door studenten actief te laten meekijken tijdens zijn eigen onderzoeksactiviteiten. Tegen deze eenvoudige achtergrond zijn er voortdurend maatschappelijke ontwikkelingen die de wetenschap beïnvloeden. Waar in het verdere verleden aan wetenschappers werd overgelaten om te bedenken welke onderwerpen aandacht verdienen, wordt dat momenteel gestuurd via subsidieverstrekingen, beroepsverenigingen, maatschappelijke wetenschaps-agenda's en zogenoemde *co-creation labs*. Ook is er meer dan voorheen behoefte aan het genereren van maatschappelijke impact ("valorisatie"). Tijd die daaraan wordt besteed kan uiteraard niet worden besteed aan het doen van onderzoek. Paradoxaal zijn wetenschappers die vaak in de media figureren lang niet altijd dezelfde als zij die de wetenschappelijke inzichten daadwerkelijk hebben bewerkstelligd. Tot slot zij opgemerkt dat protocollen die bedoeld zijn om ethisch handelen te bevorderen, soms als ongewenst neveneffect lijken te hebben dat er minder onderzoek wordt gedaan. Zo zijn sommige onderzoekers zo blij wanneer ze van de ethische toetsingscommissie toestemming krijgen om het voorgestelde onderzoek te doen, dat ze aan het daadwerkelijk uitvoeren van dat onderzoek niet meer toekomen. Datzelfde geldt voor

preregistraties, dat wil zeggen dat de onderzoeker met een tijdschrift afsprekt een bepaald onderzoek te doen en dat het tijdschrift dat onderzoek zal publiceren, ook als er sprake blijkt te zijn van een negatieve uitkomst. Nog voordat het onderzoek wordt uitgevoerd, schrijft de onderzoeker al een tekst met de rationale en de methode. Die tekst wordt voor de dataverzameling gepubliceerd. In de wandelgangen wordt er regelmatig over preregistraties gesproken als waren het volwaardige artikelen. Preregistraties voegen ook toe aan de toch al onoverzichtelijke stortvloed aan publicaties. De neiging om vooral te verwijzen naar de recentste output (inclusief preregistraties) leidt ertoe dat oudere bevindingen worden genegeerd en dat jonge onderzoekers daardoor het wiel opnieuw uitvinden.

Fouten in de interpretatie van bevindingen hangen soms ermee samen dat het in het algemeen moeilijk is om causale verbanden valide te detecteren. Mensen lijken meerdere criteria te gebruiken om een causaal verband vast te stellen (Einhorn & Hogarth, 1986). Allereerst dienen oorzaak en gevolg samen op te treden. Dit noemt men het criterium van de covariatie. Als het gevolg optreedt zonder dat de oorzaak aanwezig is, zwakt dat de perceptie van het veronderstelde causale verband af. Andersom geldt hetzelfde: als de oorzaak optreedt, maar het gevolg blijft uit, doet dat afbreuk aan de vermeende causaliteit. Deze covariatie is noodzakelijk maar niet voldoende voor een causaal verband. Soms wordt covariatie echter wel degelijk gezien als voldoende bewijs van causatie. Niet alleen moet er sprake zijn van covariatie, maar ook van een vaste temporele orde: de oorzaak gaat per definitie vooraf aan het gevolg. Ook hier geldt dat de volgorde op zichzelf onvoldoende is voor een causaal verband (*post hoc ergo propter hoc*). Op de derde plaats nemen we gemakkelijker een causaal verband aan als de tijd tussen oorzaak en gevolg kort is. Als iemand een vergif inneemt en een minuut later doodgaat, zal de observator er weinig moeite mee

hebben om te accepteren dat de dood werd veroorzaakt door het gif. Naarmate er meer tijd tussen oorzaak en gevolg verstrijkt, vertroebelt het causale verband. Dit verklaart deels waarom het maar moeilijk tot mensen lijkt door te dringen dat roken kan bijdragen aan het ontstaan van longziekten. In vakjargon wordt gezegd dat de contiguiteit hoog is als het interstimulusinterval klein is. Het waarnemen van een causaal verband wordt dus vergemakkelijkt door een hoge contiguiteit. Het vierde impliciete criterium voor de detectie van een causaal verband is de similariteit. Dat wil zeggen dat oorzaak en gevolg vergelijkbaar moeten zijn. Dat vleugelslagen van een vlinder hele orkanen tot gevolg kunnen hebben (het *butterfly-effect*), is om die reden ongelofelijk. Tot slot is het bij het concluderen van verbanden van belang alternatieve verklaringen uit te sluiten. Interpretatiefouten leiden tot onjuiste inzichten, die soms grappig, maar soms ook schadelijk kunnen zijn (De Bruijn, 2010).

De bovenstaande kritiekpunten indachtig, zal het niet verbazen dat een deel van het niet-wetenschappelijke publiek het vertrouwen in de wetenschap bij tijd en wijlen verliest, vooral als wetenschappelijke inzichten strijdig zijn met persoonlijke overtuigingen (Bloom & Weisberg, 2007; Lewandowsky et al., 2013).

Experimenteren bij juridische besluitvorming

Volgens artikel 338 Wetboek van Strafvordering kan de rechter tot een bewezenverklaring komen, "indien hij daarvan uit het onderzoek op de terechtzitting door den inhoud van wettige bewijsmiddelen de overtuiging heeft bekomen". Hoe de rechter tot die overtuiging komt, laat de wet vrij. Het is dus aan de rechter om te beslissen of hij bepaalde informatie zwaarder vindt wegen dan andere, en of hij ontlastende informatie als het ware mentaal in mindering brengt op de belastende informatie. Het is ook aan de rechter om te beslissen welke soort fout hij wil voorkomen en welke prijs hij daarvoor wil betalen

(Friedrich, 1993). Dat laatste neemt niet weg dat er een ongeschreven veronderstelling heerst dat het (nog) erger is om een onschuldige verdachte te veroordelen dan een schuldige vrij te spreken. Onder verwijzing naar deze *Blackstone's ratio*, zegt Kassin (1997): "in the context of a legal system founded on the ideal that it is better to acquit 10 guilty people than to convict 1 person who is innocent ... the false positive error – although less frequent than the false negative – presents the more serious problem" (p. 224). Daardoor zal de overtuiging uit artikel 338 Sv. behoorlijk sterk moeten zijn. In het Engels zal de *reasonable doubt* behoorlijk klein moeten zijn.

Hoe de rechter dan ook tot een bewezenverklaring komt, dit proces van waarheidsvinding heeft op het eerste gezicht niets te maken met de besproken wetenschappelijke concepten zoals hypothesen, experimenten, controlegroepen en publicaties. Psychologie en rechtsgeleerdheid zijn dan ook fundamenteel verschillende disciplines. Als academische discipline is psychologie een empirische wetenschap (in het Engels aangeduid met de term *science*). Rechtsgeleerdheid is dat niet. Dat is eerder een hermeneutische, constructieve, of normatieve wetenschap (zie Van den Bersselaar, 2007). Haack stipt het verschil aan als zij concludeert dat het woord wetenschap meeromvattender is dan het woord science: "... that not-quite-translatable German word Wissenschaft is broader than science" (2003, p. 344). Een principieel verschil is dat psychologie ontdekkingen pleegt te doen inzake van nature bestaande wetmatigheden in gedrag, terwijl het bij rechtsgeleerdheid meer gaat om het doen van uitvindingen, in het bijzonder, het verzinnen van regels. Ontdekken en uitvinden zijn overigens niet altijd evengoed te onderscheiden. Zo is het moeilijk om te zeggen of atoomenergie is ontdekt of uitgevonden.

Hoe verschillend rechtsgeleerdheid en psychologie ook mogen zijn, rode draad in dit boek is dat het proces van strafrechtelijke bewezenverklaring op minstens drie

manieren baat kan hebben bij wetenschap. Ten eerste moet worden bedacht dat er bij strafrechtelijke waarheidsvinding moet worden vertrouwd op empirische inzichten. De jurisprudentie en juridische argumenten in het algemeen zijn regelmatig doorspekt van empirische inzichten, soms zonder dat daar een empirisch-wetenschappelijke basis voor is. Crombag et al. (1994) destilleerden een aantal soms dubieuze rechterlijke empirische aannamen uit een aantal vonnissen. Een deel daarvan wordt weergegeven in Tabel 3. Afgaand op deze tabel, kan worden geconcludeerd dat het de rechtspraak ten goede zou komen wanneer empirische aannamen nog meer dan toch al een wetenschappelijke basis hebben. Inmiddels zijn er bij meerdere universiteiten initiatieven op het gebied van *Empirical legal science*.

Tabel 3. Door de rechter gebruikte ankers volgens Crombag et al. (1994).

-
1. Opsporingsambtenaren zijn betrouwbaardere getuigen dan gewone mensen.
 2. Op hun sterfbed liegen mensen zelden.
 3. Eens een dief, altijd een dief.
 4. Mensen met aanzien maken geen seksueel misbruik van hun kinderen.
 5. De meeste rechters slaan hun vrouw niet.
 6. Vrouwen die beweren verkracht te zijn, liegen zelden.
 7. Het Nederlands Forensisch Instituut maakt zelden fouten.
 8. De meeste Marokkaanse jongemannen maken zich bij herhaling schuldig aan delicten.
 9. Het Openbaar Ministerie vervolgt zelden zonder dat daarvoor een goede reden is.
 10. Orthopedagogen zijn meestal heel goed in staat om uit te maken wanneer kinderen de waarheid spreken of maar wat verzinnen.
-

Behalve dat waarheidsvinding stoelt op empirische inzichten, wordt er soms een direct beroep gedaan op de wetenschap. Rechters vragen advies aan velerlei deskundigen. Ook hier geldt dat niet elke forensisch wetenschapper optimaal wetenschappelijke informatie verschaft. Lilienfeld en Landfield (2008) beweren zelfs dat een deel van de forensische disciplines (of hun

beoefenaren) kenmerken vertoont van pseudowetenschap. Tabel 4 biedt een overzicht.

Tabel 4. Kenmerken van pseudowetenschap in forensische disciplines (Lilienfeld & Landfield, 2008).

Forensische discipline	Pseudowetenschappelijk kenmerk
1. Dactyloscopie	Onweerlegbaarheid en ad hoc manoeuvres
2. Grafologie	Afwezigheid van peer review
3. Waarheidsserum	Afwezigheid van zelfcorrectie
4. Verdachtenverhoor	Gebrekkige bescherming tegen confirmatiebias
5. Ooggetuigenverklaring	Nadruk op anekdotes
6. Polygrafie	Extravagante claims
7. Hypnose	Ad antequitem bias
8. Anatomisch correcte poppen	Omgekeerde bewijslast
9. Profiling	Afwezigheid van connectie
10. EMDR	Vakjargon

Tot slot wordt hier betoogd dat rechterlijke waarheidsvinding zou kunnen profiteren van procedures die gangbaar zijn in de wetenschap. Er zijn niet veel theorieën over hoe rechters tot hun overtuiging komen. Het heeft er in ieder geval alle schijn van dat wordt geteld hoeveel bewijsmiddelen er in het dossier zitten, dat een gewicht aan die bewijsmiddelen wordt gehecht, en dat de bewijsmiddelen niet op zichzelf staand worden gezien, maar in onderling verband. Zo betoogde Wigmore (1931) dat bewijsmiddelen in een (denkbeeldig) schema kunnen worden gezet om hun gewicht en onderlinge samenhang te ontwaren. Pennington en Hastie (1992) stellen in hun *story model* dat bewijsmiddelen pas enige betekenis krijgen als ze passen in een coherent verhaal (zie ook Greenspan & Scurich, 2016; Simon, 2004). Ook Crombag et al. (1994) benadrukken het belang van een coherent verhaal. In hun theorie van verhaal en verankering, dient het verhaal (dat wil zeggen de tenlastelegging) op aangewezen punten te worden verankerd met bewijsmiddelen. Die bewijsmiddelen

kunnen vervolgens weer worden verankerd met andere bewijsmiddelen. Op die manier wordt steeds verder afgedaald, totdat het niveau van feiten van algemene bekendheid wordt bereikt.

Vanwege de bescheidenheid van het aantal theorieën, maar ook vanwege het raadkamergeheim, blijft het proces van rechterlijke besluitvorming ietwat obscuur. Daardoor is het moeilijk doceerbaar en controleerbaar. Dat er volgens rechters zelf, bij de overtuiging “intuïtie, Fingerspitzengefühl, ervaring, of praktische wijsheid” komt kijken (Cleiren, 2010, p. 262), maakt het niet duidelijker. Reeds in 1966 schreven Kalven en Zeisel een grote groep Amerikaanse rechters aan en vroegen hun naar een zaak die zij onlangs hadden voorgezeten. De betreffende zaken waren beslist door een jury. De rechters werd gevraagd om aan te geven hoe hun oordeel over de schuldvraag zou hebben geluid als niet de jury maar zichzelf dat oordeel hadden moeten vellen. Op die manier konden de onderzoekers verschillen en overeenkomsten in het oordeel van jury en rechter in kaart brengen. Het ging om niet minder dan 3576 strafzaken. De resultaten worden weergegeven in Tabel 5.

Tabel 5. Percentage overeenstemming in strafrechtelijke oordelen van rechters en jury (Kalven & Zeisel, 1966).

	Jury veroordeelt	Jury spreekt vrij
Rechter veroordeelt	64%	19%
Rechter spreekt vrij	3%	14%

Uit deze tabel blijkt dat de overeenstemming tussen het oordeel van de rechter en de jury aanzienlijk is, namelijk 78%. Rechters komen dus niet zo vaak tot een ander oordeel dan dat leken doen. Van de 22% meningsverschil wordt 19% verklaard doordat de jury een verdachte vrijspreekt die volgens de rechter veroordeeld had moeten worden. De jury is dus verdachtenvriendelijker dan de rechter, hetgeen goed strookt met de presumptie

van onschuld. Er zijn verschillende voor- en nadelen te noemen van juryrechtspraak, denk aan democratische ideologie, temporele en financiële consequenties. Op grond van deze data kan echter worden geconcludeerd dat het tegenargument inhoudende dat de jury incapabel is om dergelijke belangrijke beslissingen te nemen, weinig hout snijdt. Kalven en Zeisel (1966) repliceerden hun onderzoek in "a comparable sample of some four thousand civil jury cases" (p. 63). De resultaten van die replicatie zijn vergelijkbaar met die uit de strafrechtelijke studie. Bij dit onderzoek kan overigens worden opgemerkt dat deelnemende rechters zich in sommige gevallen mogelijk hebben gecommitteerd aan het oordeel van de jury. Op het moment dat de rechters hun oordeel velden, waren ze immers op de hoogte van het oordeel van de jury. Kalven en Zeisel (1966) zeggen daarover: "Presumably the judge, in general, would find it more comfortable to report agreement. If, therefore, he nevertheless reports disagreement, it seems only reasonable to accept this statement" (p. 51). Uit dit citaat valt af te leiden dat de onderzoekers zich geen zorgen maken over de mogelijke conformatie van de rechter aan het juryoordeel. Het lijkt er eerder op dat ze nog meer overeenstemming hadden verwacht dan gevonden werd. Om deze kritiek te pareren hadden de onderzoekers de deelnemende rechters naar hun oordeel moeten vragen op het moment dat de jury beraad sloeg. Daarover zeggen Kalven en Zeisel (1966) echter: "It would have been extravagant to arrange for a staff of interviewers throughout the United States to personally interview a judge each time he finished a jury trial" (p. 46).

Kortom, de manier waarop rechters tot de overtuiging komen dat de verdachte het tenlaste gelegde feit heeft gepleegd, is niet veelvuldig onderzocht. Voor zover dat wel is gebeurd, lijkt de rechter geen gebruik te maken van specifieke juridische of wetenschappelijke methoden. Het oordeel van de rechter wijkt dan ook niet opvallend vaak af van dat van leken. De narratieve aanpak

waarin bewijsmiddelen worden samengesmeed tot een verhaal, zoals losse bakstenen een muur worden dankzij cement, heeft als nadeel dat hij niet beschermt tegen tunnelvisie (Rassin, 2007). Het kan immers gebeuren dat een incorrecte verdenking bewijsmiddelen aanzuigt en eindigt als overtuigend belastend verhaal. Ter illustratie, als een ons onbekend persoon naar ons glimlacht, is dat op zich vriendelijk, maar dezelfde glimlach wordt eng als we vermoeden dat deze persoon een psychopathische seriemoordenaar is die op het punt staat weer te gaan toeslaan.

Om het niveau van waarheidsvinding verder op te krikken, betogen sommigen dat er niet (alleen) narratief maar (ook) probablistisch moet worden gekeken naar strafrechtelijk bewijs. Dat komt veelal neer op een Bayesiaanse analyse van het bewijs (Fenton et al., 2013, 2016; Verhey et al., 2016). Deze aanpak zou de transparantie van het besluitvormingsproces ten goede komen. Het zou ook enige bescherming tegen tunnelvisie bieden, omdat elk bewijsmiddel niet alleen wordt beoordeeld op de steun voor het primaire scenario, maar ook voor een alternatief (onschuld) scenario. In die zin helpt de methode om de specificiteit van het bewijs in kaart te brengen. Er is ook kritiek op deze aanpak. Een praktisch obstakel is dat, waar het kwantificeren van niet-technisch bewijs al een uitdaging is, de prior odds soms moeilijk te schatten is. Het taxiprobleem is in dezen geen goed voorbeeld, omdat het gevoelsmatig gemakkelijker is om te achterhalen hoeveel taxi's van welke kleur er in de stad rondrijden, dan te achterhalen wat de accuratesse van een getuige in het onderscheiden van kleuren is. Stel echter dat we het taxiprobleem aanpassen naar een open stad waar ook talloze gekleurde taxi's uit andere steden rondrijden. In dat geval wordt het al moeilijker om een prior odds met enige nauwkeurigheid te schatten. Een ander, niet praktisch maar juridisch pijnpunt is dat een grote prior odds tenminste gevoelsmatig in strijd is met de vigerende

presumptie van onschuld. Sterker nog, omdat de rechterlijke overtuiging moet steunen op bewijs, is het de vraag of een prior odds überhaupt mag meetellen indien we deze niet per se als een bewijsmiddel beschouwen (daarop plakken we immers een likelihood ratio), maar als een soort voorafgaande overtuiging.

In zoverre de Bayesiaanse benadering van bewijs inzicht geeft in de specificiteit van bewijsmiddelen, is het een manier om het niveau van de bewijsoverweging op te tillen. Er zijn nog enkele wetenschappelijke trucs om dat te doen. Daarbij ga ik uit van de meest eenvoudige manier van bewijsevaluatie, die niet meer omvat dan het zoeken naar en in onderlinge samenhang bezien van voldoende bewijsmiddelen (minstens twee) om de vraag naar de bewezenverklaring positief te beantwoorden. De volgende aanpakken kunnen worden genoemd.

- Vertaal de vraag naar de bewezenverklaring in een hypothese, en voorspel welke bewijsmiddelen kunnen worden verwacht indien de hypothese klopt. Hierdoor verandert de besluitvorming van "zit er voldoende bewijs in het dossier dat aannemelijk maakt dat de verdachte de dader is?" naar een actieve en prospectieve inventarisatie van mogelijk bewijs ("als de verdachte dit feit heeft gepleegd, welke bewijsmiddelen kan ik dan redelijkerwijs verwachten?").
- Bekijk welk bewijs er daadwerkelijk is, maar vooral welk verwacht bewijs ontbreekt. Het ontbreken van verwacht bewijs wordt in het Engels *negative evidence* genoemd. Hoewel negatief bewijs nogal eens wordt wegverklaard ("er zijn geen vingerafdrukken omdat de dader handschoenen droeg", "de getuige herkent de verdachte niet, maar dat komt omdat de getuige de dader toch niet zo goed had gezien"), zeker in combinatie met ander wel gevonden bewijs, is negatief bewijs soms sterk

diagnostisch voor onschuld en relativeert het dus het bewijs (Liebman et al., 2012).

- Bekijk hoe goed elk bewijsmiddel ook past in een alternatief scenario waarin de verdachte onschuldig is. Dit scenariodenken is een equivalent van de genoemde likelihood ratio benadering. Een bewijsmiddel is alleen dan betrouwbaar als het wel past bij daderschap, maar niet in een onschuldscenario.
- Inventariseer welke informatie tegenspreekt dat de verdachte de dader is.

Deze vier benaderingen beogen het bewijs te relativeren en dienen daarmee als waarborg tegen vals positieve conclusies. Opgemerkt zij dat deze aanpakken gangbaar zijn, bijvoorbeeld in de medische diagnostiek (Mamede & Schmidt, 2023), maar ook in civiele procedures. De Bock schrijft bijvoorbeeld: "Een feit is voor de civiele rechter bewezen wanneer uit de beschikbare bewijsmiddelen redelijkerwijs kan worden afgeleid dat het feit zich heeft voorgedaan, en uit die bewijsmiddelen níet even goed kan worden afgeleid dat hetgeen de wederpartij met betrekking tot dat feit stelt, zich heeft voorgedaan, terwijl zich evenmin de situatie voordoet dat bewijsmateriaal dat redelijkerwijs verwacht mocht worden, ontbreekt" (2011, p. 69). In dit citaat komt het scenariodenken en het betrekken van negatief bewijs duidelijk naar voren.

THE PROPER SEAT

In 1988 schreef Willem Albert Wagenaar het artikel *The proper seat*, dat zijn oorsprong vindt in een rechtszaak waarin hij als deskundige een bijdrage had geleverd. De zaak draaide om het echtpaar Groscamp dat in de nacht werd staande gehouden door een politiepatrouille op de snelweg, omdat ze te hard reden. Dat ging niet heel gemakkelijk omdat er ook rekening moest worden gehouden met ander verkeer. Toen de politieauto de Citroen van de Groscamps eindelijk had weten in te halen, zat er ongeveer 150 meter tussen de auto's. De politie moest dus even achteruitrijden. Toen de politiemannen na ongeveer 30 seconden bij de Citroen aankwamen en het bestuurdersraam zagen openen, zagen ze een vrouw achter het stuur en een man op de bijrijdersstoel. Dat was vreemd, omdat een van de politiemannen zeker wist dat hij tijdens het inhalen een man achter het stuur had zien zitten. Vermoed werd dat het echtpaar van plaats had gewisseld in de tijd dat de politie naar hun auto manoeuvreerde. De man kreeg een boete voor hardrijden en de vrouw voor meeneed. Tijdens het proces werd Wagenaar gevraagd om iets te zeggen over de betrouwbaarheid van de verklaring van de politiemann, inhoudende dat hij zeker wist een man achter het stuur te hebben waargenomen.

Het artikel is om minstens twee redenen interessant. Enerzijds vanwege de manier waarop Wagenaar de onderzoeksvraag trachtte te beantwoorden. Anderzijds vanwege zijn manier van rapporteren aan de rechter. Wat dat eerste betreft, had Wagenaar een literatuuroverzicht kunnen schrijven over het effect van korte blootstelling, donkerheid, waarnemingshoek, maar ook de zekerheid van de politiemann. Hij koos er echter voor om een toegepast "experiment" te doen. Zijn gedachtegang was als volgt. Stel dat het wel degelijk mevrouw Groscamp was die reed, hoe groot is dan de kans dat de politiemann toch zou denken een man te hebben gezien? Uiteraard was de politiemann niet

beschikbaar als proefpersoon in dit experiment. Wagenaar besloot om 210 personen in de leeftijd van de politiemannin bloot te stellen aan mevrouw Groscamp onder vergelijkbare omstandigheden, dat wil zeggen in een donkere ruimte, kortdurend en zijwaarts. Vervolgens vroeg hij wat de proefpersonen hadden gezien. Niet minder dan 116 mensen (55%) dachten een man te hebben gezien. Wagenaar concludeerde dat de verklaring van de politiemannin niet zo betrouwbaar was, omdat zelfs als mevrouw Groscamp reed er ongeveer 55% kans is dat hij een man dacht te hebben gezien. Wagenaar had oog voor de beperkingen van zijn toegepaste onderzoek. Zo waren de proefpersonen geen politiemannen, met corresponderende vermeende waarnemings-superioriteit. Ook is onduidelijk of de experimentele omstandigheden die van de daadwerkelijke waarneming goed nabootsten. Hij zegt daarover: "Uncertainty cannot be avoided when basic scientific principles are generalized to real-life situations" (p. 507). Wagenaar laat zien hoe dataverzameling een bijdrage kan leveren aan de discussie over de kracht van bewijsmiddelen. Dat voegt nuance toe die er niet is als we het dogma hanteren dat politiemensen betere waarnemers zijn dan gemiddeld.

Over de discussie met de rechter schrijft Wagenaar: "Professor, am I correct in saying that 55% of your subjects mistook Mrs. Groscamp for a man?" "Yes your honor", "In that case I must conclude that there is a 55% chance that she was indeed driving the car, and that she therefore did not commit perjury" (p. 503). Wagenaar vond deze gedachtegang niet helemaal correct. Zijn conclusie dat er 55% kans is dat de politiemannin een man dacht te hebben gezien, als in werkelijkheid mevrouw Groscamp achter het stuur zat, is niet hetzelfde als de conclusie van de rechter die luidt dat er 55% kans is dat mevrouw Groscamp reed als de politiemannin een man zegt te hebben gezien. Ter vergelijking, de conclusie dat de meeste verkrachters mannen zijn betekent niet dat de meeste mannen

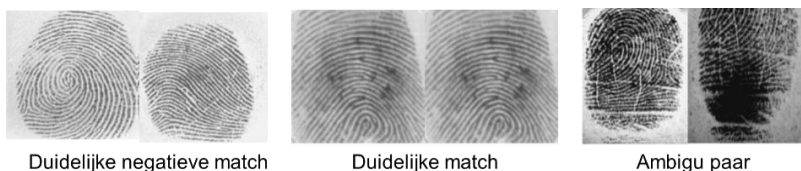
verkrachters zijn. Wagenaar had nog wel zijn best gedaan om deze *transposed conditional* redeneerfout te voorkomen. Dat brengt ons bij de tweede reden waarom dit artikel interessant is. Wagenaar noemt als een grote beperking van zijn onderzoek dat hij alleen heeft onderzocht wat de kans is op de verklaring dat er een man achter het stuur zat, terwijl dat in werkelijkheid een vrouw was. Hij onderzocht daarmee de kans op een vals positieve uitkomst. Hij heeft niet onderzocht hoe groot de kans was dat de politiemann een man dacht te hebben gezien in geval het ook daadwerkelijk meneer Groscamp was die reed (de terecht positieve kans). Door het niet draaien van deze controleconditie, mag het onderzoek geen experiment heten. Wagenaar liet dit deel van het onderzoek na, alleen al om kosten te besparen. In plaats daarvan veronderstelde hij deze terecht positieve kans op de maximale 100%: als het wel meneer Groscamp was die reed, is de kans dat de politiemann een man dacht te hebben gezien 100%. Bedenk dat het echtpaar opmerkelijk veel op elkaar leek. Met deze twee likelihoods (100% en 55%) schatte Wagenaar de kracht van de verklaring van de politiemann op een likelihood ratio van maximaal 1,82. Wagenaar haalde dus het tweehonderd jaar oude Bayesiaanse model van stal, en rapporteerde in likelihood ratio format. Hij betoogde dat alle psychologen dat zouden moeten doen, mochten zij ooit door de rechter worden gevraagd om advies te geven over de kracht van bewijsmiddelen. Sterker nog, ook forensisch psychologen, die informatie geven over de toerekeningsvatbaarheid van de verdachte zouden volgens hem Bayesiaans moeten rapporteren. Niet "Examination of Mr. Y has revealed that he committed the crime in an attack of madness", maar: "Examination of Mr. Y has revealed that he commits crimes in attacks of madness" (p. 505).

Volgens Wagenaar beschermt het rapporteren in likelihoods de deskundige en de rechter tegen allerlei soorten bias. Zo impliceert een likelihood ratio de beperking dat er uitsluitend een uitspraak wordt gedaan

over de bewijskracht van het ter beoordeling voorgelegde bewijsmiddel. Dat is de taak van de deskundige. Hij doet een uitspraak over de relatieve likelihood van het bewijs, uitgaande van het primaire scenario versus het alternatieve scenario. Hij doet geen uitspraak over de probability dat het primaire scenario correct is. Dat is de taak van de rechter, die een overzicht heeft van het gehele dossier. Het fundamentele onderscheid tussen likelihood en probability voorkomt rolverwarring. Immers, de deskundige gaat niet op de stoel van de rechter zitten. De deskundige beperkt zich tot een oordeel over dat ene bewijsmiddel.

Het gebruik van likelihood ratio's kan de deskundige ook beschermen tegen tunnelvisie, gedefinieerd als "the class of effects through which an individual's preexisting beliefs, expectations, motives, and situational context influence the collection, perception, and interpretation of evidence during the course of a criminal case" (Kassin et al., 2013, p. 45). Deze bescherming vloeit voort uit de verplichting om na te denken over de waarschijnlijkheid van het bewijsmiddel, gegeven het alternatieve scenario (ofwel de onschuldhypothese). Bescherming tegen contexteffecten (dat wil zeggen onbedoelde invloeden) zou ook kunnen worden gevonden in het gebruik van likelihood ratio's. Dat er in het domein van deskundigenanalyses sprake kan zijn van contexteffecten, werd bijvoorbeeld aannemelijk gemaakt in een onderzoek van Dror et al. (2005). Zij vroegen 27 vrijwilligers om te beoordelen of 96 paren vingerafdrukken *matchten* (zie Figuur 5 voor enkele voorbeelden). Sommige paren matchten duidelijk niet, andere wel, en tot slot waren er ambigue paren. Verder werden sommige paren vergezeld van contextinformatie, zoals een foto van een neutraal deel van de plaats delict versus een foto van een bebloed lijk. Terwijl de beoordelingen van de duidelijke paren niet werd beïnvloed door de contextinformatie, vonden de onderzoekers dat er vaker een match werd gezien in de ambigue paren indien deze werden vergezeld door emotionele contextinformatie

(66%) dan wanneer er geen contextinformatie was (46%). Behalve dat likelihood ratio's bescherming kunnen bieden tegen contexteffecten, kan die bescherming ook worden gezocht in blinding of *linear sequential unmasking*, dat wil zeggen dat de deskundige alleen die informatie krijgt die nodig is voor het beoordelen van het betreffende bewijs, terwijl hij wordt afgeschermd van andere dossierkennis.



Figuur 5. Vingerafdrukken uit Dror et al. (2005).

Tot slot kan het gebruik van likelihood ratio's bescherming bieden tegen *allegiance* bias (dat wil zeggen, de verleiding om de conclusie van het rapport zodanig te formuleren dat die gunstig uitpakt voor de opdrachtgever; Murrie et al., 2013). Likelihood ratio's zijn nu eenmaal numeriek en daar valt weinig aan te marchanderen. Ondertussen lijken deskundigen zich niet al te druk te maken over het risico op bias. Kukucka et al. (2017) vroegen aan 403 deskundigen of bias in het algemeen een probleem is voor forensische wetenschappen. Deze vraag werd bevestigend beantwoord door 71% van de respondenten. Dat bias ook een rol speelt in het vakgebied van de deskundigen werd door 52% onderkend. Dat zijzelf last zouden kunnen hebben van bias, bevestigde nog slechts 26%. Tegen deze achtergrond wordt gesproken van een *bias blind spot*.

Terwijl het al opmerkelijk is dat deskundigen geplaagd worden door bias, is het ook nog eens zo dat psychologen, die vooroplopen bij het waarschuwen tegen bias, zelf ook even vatbaar zijn voor bias als ieder ander (Bogaard et al., 2014). Met name klinisch psychologen zijn gewend aan een aantal professionele principes die haaks

staan waarheidsvinding. Greenberg en Shuman (1997) komen zelfs tot de conclusie dat het onmogelijk is voor een psycholoog om de rol van therapeut te combineren met die van forensisch deskundige. De auteurs noemen een aantal fundamentele verschillen. Om te beginnen is voor de therapeut de cliënt de opdrachtgever, zij het indirect, terwijl de opdrachtgever van de rechtspsycholoog de rechter is. Het doel van de therapeut is om het welbevinden van de cliënt te bevorderen, terwijl dat van de rechtspsycholoog is om bij te dragen aan de waarheidsvinding. Verder is er een verschil in vereiste attitude en vaardigheden. In een therapeutische context staan empathie, diagnostiek en behandeling centraal, in de forensisch context gaat het om gepaste afstandelijkheid en wetenschappelijke toetsing. Verder is een therapeutisch context cliënt-gecentreerd: het gaat om het welbevinden van de cliënt. De deskundigencontext is juist waarheidsgecentreerd. Het klinkt wellicht bevreemdend, maar in een therapeutische context is de waarheid van secundair belang. Stel bijvoorbeeld dat een cliënt zich bij de psycholoog meldt met klachten die zijn terug te voeren naar een traumatische gebeurtenis. Dan zal de behandelaar het narratief over die gebeurtenis als uitgangspunt (moeten) nemen. Het is niet de bedoeling om dan een feitenonderzoek te starten naar de vraag of het betreffende incident daadwerkelijk heeft plaatsgehad. Dat zou in strijd zijn met de vertrouwensrelatie. Crombag et al. (1994) verwoordden het als volgt: "In de gewone psychiatrische praktijk is de subjectieve beleving van de cliënt de maat van alle dingen: of iets waar is of niet, is niet van belang; van belang is slechts hoe de cliënt het ziet" (p. 328). Dit klinische uitgangspunt is geschikt voor de behandelcontext, maar het kan tot verwarring leiden wanneer de behandelaar als deskundige een uiteenzetting gaat geven over de posttraumatische klachten van de cliënt (die inmiddels getuige is in een strafzaak tegen de vermeende veroorzaker van de traumatische gebeurtenis). Het is dan mogelijk dat

de rechter in een redenering vervalt inhoudend dat als de deskundige overtuigend uiteenzet wat de posttraumatische klachten zijn, dat trauma echt gebeurd moet zijn (en de verdachte het daarom veroorzaakt heeft). Boesch en al. (1998) benadrukken dan ook dat therapeuten niet als deskundigen moeten optreden. Shuman et al. (1998) concluderen dat als een behandelaar een deskundigenrapport schrijft over zijn eigen cliënt, het zo goed als onmogelijk is om enerzijds een empathische relatie te onderhouden met die cliënt en anderzijds een zakelijk rapport te produceren: "Either way, therapy loses" (p. 521).

Dat klinische psychologie en rechtspsychologie niet altijd goed samengaan, wordt mede veroorzaakt doordat de basisheuristieken van klinisch psychologen en strafrechtjuristen tegengesteld zijn. Vanwege de presumptie van onschuld is een vals positieve fout (het veroordelen van een onschuldige) ernstiger dan een vals negatieve fout (het vrijspreken van een schuldige). In de klinische setting is juist een vals negatieve fout ernstiger dan een vals positieve. Het is voor een hulpverlener erger om een aandoening over het hoofd te zien (vals negatieve diagnose) dan om een niet-bestaande kwaal te behandelen (vals positieve diagnose), hoewel deze vuistregel deels afhangt van de ingrijpendheid van de kwaal en de remedie (Rassin & Merckelbach, 1999). Ook Richards et al. (2015) waarschuwen tegen het te gemakkelijk combineren van klinische en forensische rollen en spreken van een potentiële *role bias*. Zajac et al. (2013) concluderen, waar het gaat om het optreden als deskundige in seksueel misbruikzaken: "At the minimum these experts will have earned a PhD in psychology or a related discipline, hold a position at a research institution, and have published a significant number of empirical papers in internationally recognised peer-reviewed journals (for the most part, conference presentations or posters are no substitute for peer-reviewed journal publications). It is also our view that people whose primary occupation is clinical practice are not

automatically qualified to give evidence on human memory, even if they have published the occasional paper" (p. 615). Daar kan dan bij worden opgemerkt dat ook een PhD niet beschermt tegen bias.

Wagenaar was met zijn artikel *The proper seat* een pionier, waar hij betoogde dat forensisch psychologen moeten rapporteren in termen van likelihoods. Zijn betoog is echter niet opgepikt. Paradoxaal genoeg is het de afgelopen jaren gebruikelijk geworden om technisch-forensische rapporten te formuleren in likelihood ratio format, tenminste in Europa en Australië (Thompson et al., 2018). De *Association of Forensic Science Providers* (AFSP) schreef in 2009 dat "across all scientific disciplines" (p. 161), "the opinion or conclusion should express the degree of support for one proposition or the other depending upon the value of the likelihood ratio" (p. 163) en dat het deskundigenoordeel dus rust op "the estimation of a likelihood ratio" (p. 161). Als drijfveren voor dit advies noemt de AFSP balans (een likelihood ratio vereist immers kennisname van niet alleen het primaire scenario, maar ook van een alternatief scenario), logica (een likelihood ratio vereist een prospectieve redenering, van scenario naar waarschijnlijkheid van het bewijs), transparantie (hoewel moeilijk, is een likelihood ratio inzichtelijk) en robuustheid (een likelihood ratio die stoelt op onderbouwde schattingen kan goed worden gecontroleerd door anderen). De AFSP publiceerde ook een overzicht van likelihood ratio marges met bijbehorende verbale omschrijvingen (zie Tabel 6).

Tabel 6. Likelihood ratio marges (AFSP, 2009).

Verbale omschrijving	Likelihood ratio
<i>Weak support</i>	1 - 10
<i>Moderate support</i>	10 - 100
<i>Moderately strong support</i>	100 - 1.000
<i>Strong support</i>	1.000 - 10.000
<i>Very strong support</i>	10.000 - 1.000.000
<i>Extremely strong support</i>	1.000.000 <

Het Nederlands Forensisch Instituut (NFI, 2017) heeft een eigen variant van de likelihood ratio marges gepubliceerd (zie Tabel 7). Deze tabel wijkt op twee manieren af van de internationale conventie. Op de eerste plaats zijn de marges iets anders, en op de tweede plaats, drukt het NFI de bewijskracht uit in waarschijnlijkheid in het primaire scenario vergeleken bij het alternatieve scenario, terwijl de AFSP spreekt van steun voor het primaire scenario.

Tabel 7. Likelihood ratio marges (NFI, 2017).

Verbale omschrijving	Likelihood ratio
Ongeveer even waarschijnlijk	1 - 2
Iets waarschijnlijker	2 - 10
Waarschijnlijker	10 - 100
Veel waarschijnlijker	100 - 10.000
Zeer veel waarschijnlijker	10.000 - 1.000.000
Extreem veel waarschijnlijker	1.000.000 <

Met dat laatste ondervangt de Nederlandse versie een onvoorzien probleem van de internationale conventie. Mensen blijken soms namelijk te denken dat *weak support* betekent dat het bewijs geen of onvoldoende steun voor het primaire scenario inhoudt. Martire et al. (2013) gaven een casus over een inbraak aan 75 studenten. In de casus zat een deskundigenrapport over een voetafdruk. De deskundige concludeerde in verschillende versies van de casus dat de likelihood ratio van die afdruk 4,5, 450, of 495.000 was. De helft van de deelnemers kreeg niet deze getallen, maar de corresponderend verbale omschrijving (*weak support*, *moderately strong support*, of *very strong support*, respectievelijk). Deelnemers beantwoordden twee keer de vraag "Based on the available evidence I believe that it is ... times more likely that the accused is guilty than not guilty", een keer voor en een keer na het lezen van het deskundigenrapport. Zoals verwacht, gold voor de meeste condities dat de deelnemers overtuigder waren geraakt na lezing van het rapport. Uitzondering was echter de versie waarin de deskundige concludeerde dat de voetafdruk weak

support opleverde voor de schuldhypothese. In die groep zakte de overtuiging juist. Martire et al. (2013) spreken van een paradoxaal *weak evidence effect*.

Zoals te zien, kunnen likelihood ratio's in forensisch technische analyses behoorlijk groot zijn. Dan gaat het veelal om analyses op identificatieniveau ofwel bronniveau. De sterke bewijskracht is eraan te danken dat vingerafdrukken en DNA zeer uniek zijn. Als er een likelihood ratio wordt geschat, zal daarom wel een hoog getal in de teller staan, maar een zeer laag getal in de noemer. Het wordt anders als de verdediging een verweer voert op activiteitsniveau. Stel dat de politie een vingerafdruk heeft veiliggesteld in een winkel die nachtelijk bezoek heeft gehad van een inbreker. Het NFI concludeert dat de kans op het vinden van deze vingerafdruk zeer veel groter is als hij van de verdachte is dan wanneer hij van een willekeurig ander persoon is. Stel nu dat de verdediging niet betwist dat de verdachte in de winkel is geweest, maar dat wordt betoogd dat hij daar simpelweg een dag voor de inbraak was als reguliere klant en dat zijn vingerafdruk toen is achtergebleven. Nu bewijst de vingerafdruk niet meer dat de verdachte de dader is. In de woorden van Kokshoorn et al. (2017): "More often than not, the source of DNA traces found at the crime scene is not disputed, but the activity or timing of events that resulted in their transfer is" (p. 115). Er moet nu een analyse op activiteitsniveau plaatsvinden, waarbij allerlei ruisveroorzakende vragen een rol gaan spelen, zoals hoe lang blijft een vingerafdruk detecteerbaar, had de verdachte vettige handen tijdens zijn bezoek overdag, is er toevallig een schoonmaak geweest tussen dat bezoek en de inbraak? Al deze verstoringen variabelen drukken de likelihood ratio. Zo geven Kokshoorn et al. voorbeelden waarin een volledige DNA-analyse met een extreem grote likelihood ratio op identificatieniveau, op activiteitsniveau nog slechts een likelihood ratio oplevert van tussen de 9 en 91. Als het gaat om een gemengd DNA-

profiel, is de maximaal haalbare likelihood ratio nog slechts 4,74 tot 9.

Zoals gezegd, is Wagenaar's betoog niet opgepikt door forensisch psychologen. Technisch-forensische disciplines hebben de likelihood ratio inmiddels wel omarmd, overigens zonder enige verwijzing naar Wagenaar (1988), of Bayes (1763). In 2022 hebben Rassin et al. geprobeerd om Wagenaar's boodschap opnieuw onder de aandacht van psychologen te brengen, overigens zonder enig merkbaar gevolg. In dat artikel is betoogd dat de marges voor likelihood ratio's in psychologische analyses lijken op die zoals gangbaar in de wetenschap, en niet zoals die die worden gehanteerd in de technische bronanalyses (zie Tabel 8).

Tabel 8. Likelihood ratio marges (Rassin et al.; 2022).

Verbale omschrijving	Likelihood ratio
Weak/anecdotal support	2 - 3
Positive/substantial support	3 - 10
Positive/strong support	10 - 30
Strong/very strong support	30 - 100
Very strong/decisive support	100 <

Wagenaar eindigde zijn artikel somber. Volgens hem snappen juristen niet goed wat likelihood ratio's zijn en verwarren ze die met de posterior odds, en: "A worse problem is that the experts themselves are not always fully aware of the distinction" (p. 509).

HET HERKENNEN VAN IWAN

In 1989 schreef Wagenaar het boek *Het herkennen van Iwan*. De aanleiding voor dat boek was een casus die draaide rond John Demjanjuk. Deze man werd in 1976 toevallig herkend door overlevenden van concentratiekamp Treblinka. Deze getuigen zien in Demjanjuk ene Iwan de verschrikkelijke, een wrede bewaker (zelf krijgsgevangene) die verantwoordelijk wordt gehouden voor de vergassing van 850.000 gevangenen in de periode 1942-1943. Hij zou in een stilstaande tank hebben gezeten die met ronkende motor uitlaatgassen de gaskamers inblies. Meer dan dertig jaar later denken enkele overlevenden Iwan te herkennen in de Oekraïense immigrant John Demjanjuk. De verdachte ontkent. Willem Albert Wagenaar diende in de strafzaak tegen Demjanjuk (die inmiddels door de Verenigde Staten van Amerika was uitgeleverd aan Israël) als deskundige en gaf commentaar op de verschillende herkenningen van Demjanjuk door de getuigen. Later schreef Wagenaar een boek over zijn ervaringen in deze strafzaak. In dat boek geeft hij niet minder dan 50 regels voor het correct uitvoeren van een Osloconfrontatie. Sommige van die regels zijn belangrijker dan andere. Sommige regels lijken inmiddels achterhaald door moderne technische mogelijkheden. Maar in ieder geval illustreren de regels wat de kern is van een Osloconfrontatie. Ze maken duidelijk hoe en waarom een Osloconfrontatie (mits goed uitgevoerd) een sterkere herkenning oplevert dan een eenpersoonsconfrontatie. Hieronder worden Wagenaar's regels opgesomd en soms van commentaar voorzien.

1. *De betrokkenheid bij een misdaad kan niet worden bewezen door middel van een eenpersoons- of meerpersoonsidentificatietest, wanneer de getuige de verdachte al voor het misdrijf kende.* Het centrale idee onderliggend aan de Osloconfrontatie is dat als de getuige de verdachte aanwijst, die identificatie dan niet of nauwelijks aan een andere oorzaak kan worden

geattribueerd dan dat de verdachte de dader is. Als een getuige verklaart dat hij een hem bekend persoon heeft herkend tijdens het delict, vond de herkenning reeds op dat moment plaats. Als deze getuige vervolgens wordt onderworpen aan een Osloconfrontatie zal hij allicht de verdachte herkennen, maar die herkenning is nu niet meer uitsluitend toe te schrijven daaraan dat de getuige de verdachte gezien moet hebben toen die het strafbare feit pleegde. Het kan namelijk ook zo zijn dat de getuige de hem bekende persoon aanwijst, waarvan hij dacht dat hij hem het strafbare feit heeft zien plegen. De Jong et al. (2005) zeggen het als volgt: "identification problems of a different kind may occur with persons already known to a witness ... (e.g., nephew) ... a subsequent identity test in which you are asked to identify your nephew is of course useless and even nonsensical" (p. 88).

Deze eerste regel heeft bredere relevantie als wordt bedacht dat de verhouding tussen de getuige en de dader voor de politie een gegeven is. Daar heeft de politie geen invloed op, maar het is dus wel een omstandigheid die de politie moet achterhalen om te beslissen of een herkenningsprocedure überhaupt zinvol is. Er zijn meer van dergelijke externe factoren, die worden aangeduid met de term *estimator* variabelen. Dit ter onderscheiding van systeemvariabelen die wel onder invloed van de politie staan (denk aan de manier waarop de politie de confrontatie uitvoert). Wagenaar en Van der Schrier (1996) onderzochten het effect van twee andere estimator variabelen, te weten afstand en lichtsterkte tijdens het waarnemen van het incident, op de accuratesse van een latere uitkomst van een goed uitgevoerde Osloconfrontatie. Hun bevindingen worden samengevat in Tabel 9. De gevonden likelihoods volgen een logisch patroon: hoe dichterbij en beter verlicht, des te sterker de latere herkenning. Wagenaar en Van der Schrier introduceerden dan hun zogenoemde vuistregel van 15: voor een stevige identificatie (met een likelihood ratio van minstens 15) mag

de getuige niet verder dan 15 meter en bij niet minder dan 15 lux hebben waargenomen.

Tabel 9. Likelihoods van terechte (vet) en valse herkenningen (Wagenaar & Van der Schrier, 1996).

Afstand (meters)	Lichtsterkte (lux)								
	0,3	2	3	5	10	30	150	300	3000
3	14 9	46 10	61 8	79 8	82 6	82 6	82 3	86 2	86 2
5	11 5	25 11	54 5	50 1	68 1	79 1	82 0	86 0	86 1
7		7 5	36 9	43 6	71 5	68 5	82 3	79 2	86 1
12			4 4	25 6	43 4	43 4	57 1	57 1	61 1
20				11 5	18 9	18 9	43 8	39 8	57 5
30					14 4	25 7	21 6	32 6	36 6
40						7 3	14 8	25 8	29 9

De Jong et al. (2005) rapporteerden vergelijkbare likelihoods, maar dan voor de situatie waarin een getuige een hem/haar bekend persoon een delict ziet plegen. In dat geval vindt de herkenning dus tijdens het delict plaats, en heeft een confrontatieprocedure op het politiebureau (althans een Osloconfrontatie) geen zin meer.

2. *Identificatie van een verdachte die overigens totaal onbekend is aan de getuige, moet geschieden door middel van een meerpersoonsidentificatietest, niet door middel van een eenpersoonstest.* Met deze richtlijn toont Wagenaar zich voorstander van de Osloconfrontatie. Inmiddels is er onderzoek gepubliceerd waaruit kan worden afgeleid dat een herkenning in een Osloconfrontatie inderdaad meer waarde heeft dan een herkenning in een eenpersoonsconfrontatie (*show-up*). Steblay et al. (2003) verzamelden acht studies naar de kracht van *line-ups* en *show-ups*. De vergelijking leverde verschillende resultaten op. Als wordt gekeken naar de uitkomst van *target present*

displays, dat wil zeggen dat de verdachte inderdaad de dader is (het primaire scenario), werd hij in 47% van de show-ups herkend en in 45% van de line-ups. In geval van *target absent displays* (de verdachte is niet de dader; het alternatieve scenario), werd de verdachte (ten onrechte) in 23% van de show-ups herkend en in 17% van de line-ups. Deze data lijken niet veel te verschillen, maar in termen van likelihood ratio's, is de kracht van een herkenning in een show-up ($47\% : 23\% =$) 2 en in geval van een Osloconfrontatie ($45\% : 17\% =$) 2,6, afgerond 3. Verder bleek dat in geval van show-ups, 69% van alle resultaten correct waren, en slechts 51% van alle resultaten in line-ups. Dat komt doordat er bij line-ups, anders dan bij show-ups, de mogelijkheid bestaat dat de getuige een aantoonbaar onschuldige figurant herkent, en zich aldus blijkt te vergissen. Die uitkomst is niet mogelijk bij show-ups, omdat figuranten daar afwezig zijn.

Tabel 10. Likelihoods van terechte (vet) en valse herkenningen (De Jong et al., 2005).

Afstand (meters)	Lichtsterkte (lux)								
	0,3	2	3	5	10	30	150	300	3000
3	0 0	79 0	100 5	100 0	93 8	96 3	94 0	97 0	97 0
5	0 0	68 14	89 8	94 14	94 3	97 6	92 0	92 7	100 3
7		63 3	63 10	89 3	93 6	93 3	97 6	96 3	90 0
12			60 3	70 6	76 6	81 5	80 3	92 5	100 2
20				32 6	52 14	65 9	77 13	79 6	100 14
30					20 10	29 8	32 11	45 8	40 5
40						7 3	17 5	17 10	30 2

3. De aanwijzing van een verdachte door middel van confrontaties met politiefoto's, kan niet worden gebruikt als bewijs van identiteit. Een cruciaal verschil tussen een

Osloconfrontatie en een politieserie is dat in dat laatste geval de figuranten niet aantoonbaar onschuldig zijn. Alle getoonde foto's zijn immers van verdachten uit eerdere zaken. Welbeschouwd is er sprake van een aaneenschakeling van show-ups.

4. *Getuigen die deel hebben genomen aan een confrontatie met behulp van politiefoto's moeten in principe worden uitgesloten van latere identificatietests.* Politiefoto's kunnen volgens Wagenaar (1989) worden gebruikt als opsporingsinstrument. Als een getuige voor dit middel wordt ingezet en daarna nog een bewijsconfrontatie ondergaat, bestaat de kans dat hij de verdachte in tweede instantie herkent, maar dan niet per se als de dader, maar omdat hij hem eerder zag in de politieserie. Een positieve herkenning kan dan het gevolg zijn van een bronverwarring (*source monitoring error*). Er wordt ook wel gesproken van een *false fame effect* (Jacoby et al., 1989) of een *bystander effect*. Die laatste term werd oorspronkelijk gereserveerd voor de situatie waarin de ene getuige een andere getuige per abuis aanzag voor de dader. Men spreekt overigens ook van een bystander effect als mensen anderen niet te hulp schieten, omdat ze denken dat iemand anders dat wel zal doen (Darley & Latané, 1968).

5. *Getuigen moeten niet worden gebruikt voor confrontaties met behulp van politiefoto's wanneer een werkelijke identificatietest kan worden uitgevoerd.* Met deze regel geeft Wagenaar (1989) aan dat hij een bewijsconfrontatie (Osloconfrontatie) prefereert boven een opsporingsconfrontatie.

6. *Er moet minder gewicht worden toegekend aan identificatietests wanneer die vooraf zijn gegaan door een confrontatie met foto's uit een politiearchief, zelfs wanneer bij de twee procedures verschillende getuigen zijn gebruikt.* Wagenaar (1989) formuleerde verschillende regels vanuit het onderliggende idee dat ooggetuigenbewijs feilbaar is. Deze regel is daar een van. De vermeende feilbaarheid speelt twee keer indien de verdachte is opgespoord via een

herkenning en het bewijs vervolgens ook uit herkenningen bestaat.

7. *Geen enkele getuige moet worden gevraagd om dezelfde verdachte meer dan eenmaal te identificeren.* Door getuigen herhaaldelijk te confronteren met de verdachte, ontstaat het risico op een valse herkenning als gevolg van de eerder genoemde bronverwarring.

8. *Identificatie door ooggetuigen moet niet aanvaard worden als bewijs, als er niet ook een andersoortig bewijs voorhanden is.* Regels 8 en 9 zijn geschreven vanuit de gedachte dat het risicovol is om een bewijsconstructie uitsluitend te baseren op (feilbaar) ooggetuigenbewijs.

9. *Bewijs van identiteit door een identificatietest moet worden ondersteund door aanvullend bewijs van een andere soort, niet door andere identificatietests, zelfs niet wanneer deze gebruik maken van andere getuigen.*

10. *Wanneer er meerdere getuigen beschikbaar zijn, moet worden getracht een aanvankelijke identificatie te ontzenuwen met behulp van deze andere getuigen.* Met deze ogenschijnlijk wellicht ietwat vreemde regel poogt Wagenaar (1989) het wetenschappelijke principe van falsificatie te benadrukken. Concreet zou men bij een positieve herkenning van de verdachte in een initiële procedure, een tweede getuige kunnen onderwerpen aan een Osloconfrontatie waarin de figuranten nog strenger zijn geselecteerd op hun gelijkenis met het signalement.

11. *Alle identificatiepogingen, ongeacht of zij succes hadden of niet, moeten worden gerapporteerd.* Deze regel beoogt te voorkomen dat een geverbaliseerde identificatie van de verdachte te flatteus is. Stel bijvoorbeeld dat er in het dossier een proces-verbaal zit van een herkenning in een Osloconfrontatie. Stel verder dat er vier getuigen waren, waarvan er één de verdachte herkende, maar de drie andere niet. Die drie nonherkenningen (negatief bewijs) zijn wellicht diagnostisch en daarom is het van belang om ook van deze uitkomsten processen-verbaal op te stellen. Er zijn aanwijzingen dat nonherkenningen

inderdaad relevant zijn. Wells en Lindsay (1980) heranalyseerden bijvoorbeeld data uit een studie van Loftus (1976) en ontdekten dat in geval van target present displays, 84% van de getuigen de dader aanwezen, terwijl in target absent compilaties, 60% van de getuigen een valse herkenning hadden. Onder de aanname van een gelijke verhouding tussen target present en absent displays, bleek dus 58% van de herkenningen correct (84 van de 144). Maar van de (16 + 40 =) 56 nonherkenningen bleek 71% correct. De onderzoeker concludeert: "there is no justifiable logic for approaching a lineup procedure with a set for considering an identification of the suspect to be informative while considering a nonidentification to be uninformative" (Wells & Lindsay, 1980, p. 777).

In hun overzichtsartikel includeerden Clark et al. (2008) 94 experimenten en vonden dat in geval van target present displays, de dader in 46% van de gevallen werd herkend, terwijl in 21% een figurant werd aangewezen en in 33% niemand werd aangewezen. In geval van target absent displays, werd in 13% van de gevallen de verdachte onterecht herkend, in 35% een figurant en in 52% niemand. Na correctie voor steekproefgrootte en andere verschillen tussen de bestudeerde experimenten, kwamen de onderzoekers uit op de volgende likelihood ratio's. Een herkenning is 5,4 keer zo waarschijnlijk als de verdachte de dader is dan wanneer hij dat niet is, het herkennen van een figurant is twee keer zo waarschijnlijk als de verdachte onschuldig is dan wanneer hij schuldig is (de likelihood ratio voor het primaire scenario is dan 0,5), en een volledige verwerping van de line-up is 2,1 keer zo waarschijnlijk als de verdachte onschuldig is dan wanneer hij de dader is (likelihood ratio: 0,5). Dit duidt erop dat een herkenning bewijs van daderschap oplevert, maar een nonherkenning bewijs van onschuld. Door mogelijke prior odds te variëren, kwamen Wells en Olson (2002) tot de conclusie dat "the exonerating value of filler identification and "not there"

responses can actually exceed the incriminating value of identifications of the suspect" (p. 155).

Nonherkenningen zijn dus relevant en de rechter moet er kennis van kunnen nemen. Dat is de reden om deze regel te formuleren. Desondanks concludeerden Wells en Lindsay (1980): "neither jurors and judges nor police investigators place much faith in the eyewitness who says 'this is not the man'" (p. 776). Clark en Wells (2008) voegen toe: "... particularly easy for the casual observer to dismiss the diagnostic value of foil identifications because they are known immediately to be mistaken identifications; the eyewitness has made a mistake so the witness must have a bad memory. At one level this is true – the witness has made a mistake. But the nature of the mistake suggests that the eyewitness is saying "Your suspect in this lineup looks less like the perpetrator than does this foil," which should logically reduce the fact-finder's confidence that the suspect is the perpetrator" (p. 419).

12. *Processen-verbaal van identificatietests die een negatief resultaat hadden, moeten alle informatie verschaffen die nodig is om dat resultaat te interpreteren.*

13. *De bewijskracht van een eerste identificatie kan met behulp van identificaties door andere ooggetuigen slechts marginaal worden vergroot.* Wagenaar (1989) is van mening dat het risicovol is om een gehele bewijsconstructie te baseren op eenzelfde soort bewijs, in dit geval ooggetuigenherkenningen.

14. *Onderzoekers moeten al het mogelijke doen om onderlinge contacten tussen getuigen te vermijden.* Getuigen kunnen elkaar (onbedoeld) beïnvloeden met als gevolg dat hun herinnering wordt aangetast. Daarom is het raadzaam om onderling contact te voorkomen.

15. *Onderzoekers moeten, voorafgaand aan de identificatietest, aan de getuigen vragen of zij de zaak met andere mensen hebben besproken. De antwoorden op zulke vragen moeten worden genoteerd.* Als het niet lukt om onderling contact te voorkomen, is het op zijn minst

relevant om te weten of getuigen contact hebben gehad, zodat de eventuele positieve uitkomst van de confrontatie mede in het licht van het gedocumenteerde contact kan worden geïnterpreteerd.

16. Getuigen moeten worden uitgesloten van deelname aan identificatietests als is vastgesteld dat zij hun identificatie geheel of gedeeltelijk zouden kunnen baseren op informatie die zij van anderen hebben verkregen.

17. Getuigen moeten, voordat zij deelnemen aan een identificatietest, een verbale beschrijving van de dader geven zoals zij die zich herinneren. Deze beschrijving moet deel uitmaken van het proces-verbaal van de daaropvolgende identificatietest. Aangezien er altijd een kans is op een valse herkenning, ook in een goed uitgevoerde Osloconfrontatie, moet worden voorkomen dat nongetuigen deelnemen aan confrontaties. Om dat te bereiken, wordt getuigen gevraagd een zo specifiek mogelijk signalement te geven van de dader. Hierdoor kan niet alleen worden uitgesloten dat mensen die de dader eigenlijk niet goed genoeg hebben gezien toch een gokje mogen wagen in een met-pijn-en-moeite opgetuigde Osloconfrontatie, maar kan ook worden gezien of het signalement voldoende past op de verdachte. Als dat namelijk niet zo zou zijn, is er iets vreemds aan de hand, ofwel met de getuige, of met de verdachte. Dat getuigen eerst een signalement moeten geven van de dader voordat ze worden onderworpen aan de line-up, heeft niet alleen voordelen, maar ook een nadeel. Bedacht moet worden dat het geven van een signalement een vertaalslag vergt van een visuele herinnering naar een verbale omschrijving. Het zou vervolgens kunnen gebeuren dat die verbale omschrijving de visuele herinnering gaat beïnvloeden. Dit *verbal labelling* of *verbal overshadowing* effect werd voor het eerst beschreven door Carmichael et al. in 1932. Zij gaven hun proefpersonen plaatjes zoals die in de linker kolom van Figuur 6. Elk plaatje werd vergezeld van een beschrijving, maar die verschilde tussen de deelnemers.

Enige tijd daarna probeerden de proefpersonen de plaatjes uit hun hoofd na te tekenen, en toen bleek dat deze reproducties spontaan vervuild bleken te zijn geraakt door de verbale omschrijvingen die er eerder hadden bij gestaan. Meissner en Brigham (2001) schreven een overzichtsartikel over vijftien onderzoeken naar dit effect toegepast op het herkennen van gezichten. Zij concludeerden dat het geven van een signalement de accuratesse van latere identificatiepogingen inderdaad licht drukt (Fisher's $Z_r = -.12$). Ook vonden ze dat het verbal labelling effect sterker was als de confrontatie snel volgt op het geven van het signalement (binnen tien minuten), maar bijna geheel verdwijnt na enige retentie (30 minuten). De regel dat getuigen een signalement moeten geven van de dader beoogt dus het aantal valse herkenningen te drukken, maar kan er tevens toe leiden dat het aantal terechte herkenningen daalt.

Originele figuur	Verbale omschrijving en uiteindelijke reproductie	Verbale omschrijving en uiteindelijke reproductie
	 Four	 Seven
	 Two	 Eight
	 Broom	 Gun
	 Curtains in a window	 Diamond in a rectangle

Figuur 6. Voorbeelden uit Carmichael et al.'s (1932) verbal labeling studie.

18. *Als een vroegere beschrijving niet klopt met het uiterlijk van de verdachte, en de discrepantie kan niet op een eenvoudige wijze worden verklaard, dan bestaat er een voorkeur om de getuige die de beschrijving produceerde uit te sluiten van verdere identificatietests.* Deze regel lijkt duidelijk. Als de herinnering van de getuige (zoals blijkend uit het door hem gegeven signalement) niet strookt met de verdachte, is er iets mis. Het toch uitvoeren van een confrontatie kan dan uitmonden in een toevalstreffer (valse herkenning), waarna de opmerkelijke mismatch tussen signalement en verdachte zomaar eens zo kunnen worden vergeten.

19. *De getuige moet geïnstrueerd worden dat hij slechts één persoon mag aanwijzen, namelijk de persoon die beantwoordt aan de verbale beschrijving die eerst door dezelfde getuige voorafgaande aan de test is geproduceerd.* Stel dat een getuige wordt geconfronteerd met zes portretfoto's van mensen die voldoen aan het signalement. Een foto is van de verdachte. Strikt statistisch is de kans op een valse herkenning nu 1 op de 6 (17%) of zelfs 1 op de 7 (14%), aangezien de getuige ook mag concluderen dat hij niemand herkent. Deze verhouding, waarbij door toevoeging van zinnige figuranten de statistische kans op een valse herkenning kan worden gedrukt, gaat alleen op als de getuige slechts één persoon mag aanwijzen.

20. *De instructie die aan de getuige wordt gegeven moet letterlijk worden gerapporteerd in het proces-verbaal.* Deze regel beoogt het mogelijk te maken om de instructie aan de getuige achteraf te controleren.

21. *De instructie aan de getuige moet benadrukken dat de gezochte persoon mogelijk niet in de testgroep voorkomt, en dat een positieve identificatie daarom alleen moet plaatsvinden wanneer de getuige er zeker van is dat hij die persoon heeft herkend.* Met deze instructie wordt beoogd om de getuige conservatief te maken. Hij moet zich realiseren dat hij niet verplicht is om iemand aan te wijzen,

en dus alleen iemand moet aanwijzen als hij die persoon positief herkent.

22. *Politiefunctionarissen die de leiding hebben bij identificatietests moeten de noodzakelijke training hebben ontvangen, en moeten in staat zijn om hun professionele kwalificaties te demonstreren in geschreven en mondelinge verklaringen.* De politiebeambte zou de getuige onbedoeld kunnen beïnvloeden, bijvoorbeeld door te vragen: "Herkent u nummer 2 niet? Kijk nog eens goed". Om dit te voorkomen, dient de begeleidende politiefunctionaris goed op de hoogte te zijn van de procedure rondom bewijsconfrontaties.

23. *Een testgroep in een identificatietest mag nooit meer dan één verdachte bevatten.* Net als regel 19, beschermt deze regel een van de cruciale kenmerken van een bewijsconfrontatie, te weten dat de aanwezigheid van figuranten de kans op een valse herkenning drukt. Als er meerdere verdachten in een line-up staan (of uitsluitend verdachten), komt dit kenmerk in het gedrang.

24. *Aan getuigen moet worden verteld dat er slechts één verdachte in de testgroep voorkomt. Wanneer zij meer verdachten aan de politie hebben beschreven, moet worden verteld welke van deze verdachten zij geacht worden te identificeren.* Deze regel beoogt de getuige, net als regel 21, conservatief te maken in zijn beslissing om iemand te herkennen.

25. *Als er meer verdachten zijn die door dezelfde getuige geïdentificeerd moeten worden, is het noodzakelijk om voor iedere verdachte een aparte identificatietest te houden. Zulke achtereenvolgende tests mogen nooit figuranten bevatten die aan dezelfde getuige bij een vorige test al zijn aangeboden.* Mocht deze regel worden overtreden, dan zal de getuige tijdens de tweede confrontatie de figuranten die daar aanwezig zijn en dat ook waren bij de eerste confrontatie, kunnen herkennen. De getuige kan dan reconstrueren dat die figuranten inderdaad figurant zijn en

geen verdachte. Daarmee neemt het aantal zinnige figuranten af met alle gevolgen van dien.

26. *Aan getuigen moet worden verteld dat zij slechts één persoon bij een identificatietest mogen aanwijzen.* Deze regel vloeit voort uit dezelfde rationale als bijvoorbeeld regels 19 en 23.

27. *Identificatietests moeten niet minder dan zes personen bevatten en liefst tien tot twaalf personen.* Een van de personen in de line-up is uiteraard de verdachte. Wagenaar (1989) stelt voor dat er tussen de vijf en elf figuranten worden geïncorporeerd. De kans op een valse herkenning neemt af met de toename van het aantal figuranten volgens de formule $1 : (n + 1)$. Niettemin is er sprake van een afnemende meerwaarde en bij elf figuranten is de kans op een valse herkenning nog slechts 8%. Ondertussen wordt het praktisch moeilijk om een line-up met zoveel mensen te organiseren die allen aan het signalement voldoen. Daarom adviseert Wagenaar een plafond van elf figuranten. Bovendien zou de getuige bij al te veel figuranten door de bomen het bos niet meer kunnen zien. De ondergrens van vijf figuranten wordt niet sterk onderbouwd: "In de literatuur heb ik niet gevonden dat testgroepen van minder dan zes personen op een positieve manier worden besproken, hetgeen wellicht betekent dat er een stilzwijgende overeenstemming is over zes als minimale groepsomvang" (Wagenaar, 1989, p. 91). Eenendertig jaar later was deze onderbouwing nog niet veel verbeterd, afgaand op de opmerking van Wells et al. (2020), betreffende hetzelfde advies (overigens zonder enige verwijzing naar Wagenaar): "The recommendation that there should be at least five known-innocent fillers (thereby creating a six-person line-up) for a single suspect is somewhat arbitrary" (p. 19).

28. *Alle figuranten die deelnemen aan een identificatietest moeten beantwoorden aan de beschrijving die eerder door getuigen is gegeven.* Wederom geldt hier dat figuranten pas iets toevoegen aan het besluitvormingsproces als zij zinnige

figuranten zijn. Daartoe moeten ze tenminste aan het door de getuige gegeven signalement voldoen en dientengevolge dus enigszins lijken op de verdachte.

29. *Vroegere beschrijvingen van de dader waaraan het uiterlijk van een verdachte niet voldoet, kunnen niet worden gebruikt voor de selectie van figuranten.* Omdat ons uiterlijk onderhevig is aan veranderingen, kan een verouderd signalement problemen veroorzaken. Iets dergelijks kan bijvoorbeeld gebeuren in een casus zoals die van John Demjanjuk, waarin er sprake is van jarenlange retentie.

30. *Een identificatietest moet met voorzichtigheid worden beoordeeld wanneer er niet minstens één eerdere beschrijving is waaraan de verdachte beantwoordt.* Beargumenteerd kan worden dat Wagenaar (1989) met deze regel wellicht iets te voorzichtig is geweest. Immers, als er geen signalement van de dader is waaraan de verdachte voldoet (en dit kan niet worden verklaard door lange retentie), dan is er een probleem, en moet een Osloconfrontatie worden afgeraden.

31. *Onder normale omstandigheden moeten leden van een homogene groep, zoals politieagenten, niet worden gebruikt als deelnemers in een identificatietest.* Zelfs als figuranten voldoen aan het signalement, zouden zij door hun gedrag signalen kunnen afgeven waaruit de getuige zou kunnen afleiden dat zij de verdachte niet zijn. Neem als stereotype voorbeeld dat de organisatoren van een Osloconfrontatie in persoon een handvol collega's van de Mobiele Eenheid vragen om figurant te zijn. Dan is het denkbaar dat de verdachte op de een of andere manier toch afsteekt bij deze figuranten, omdat hij niet tot dezelfde groep behoort. Wellicht hebben de ME-ers allemaal dezelfde kisten aan, of kijken ze allemaal iets stoerder uit hun ogen.

32. *Figuranten moeten zodanig worden geselecteerd dat anderen, die geen getuige van het misdrijf zijn geweest, geen aanwijzingen zullen vinden die helpen om te raden wie de verdachte is.* Een goede manier om te toetsen of een

line-up eerlijk, niet gebiased is, is om een aantal mensen, die geen getuigen waren van het incident, het door de getuige gegeven signalement van de dader te geven, en hun vervolgens te onderwerpen aan de confrontatie. Mocht dan blijken dat van deze nongetuigen een significant groter aantal dan statistisch verwacht (zie Tredoux, 1998), uitgerekend de verdachte zegt te zouden hebben aangewezen, dan steekt de verdachte blijkbaar af tegen de figuranten en is de line-up niet goed. Omdat Doob en Kirshenbaum (1973) een van de eerste artikelen schreven over line-up *fairness* en bias, noemen we zo'n oefenconfrontatie ook wel een Doob en Kirshenbaumtest.

33. *Figuranten moeten zodanig worden geselecteerd dat de getuigen geen enkele aanwijzing vinden die kunnen helpen te raden dat één van de figuranten de verdachte niet kan zijn.* Figuranten moeten aantoonbaar niet betrokken zijn geweest bij de onderzochte misdaad. Getuigen moeten ook worden geïnstrueerd dat figuranten onschuldig zijn, opdat zij zich terdege realiseren dat ze een fout kunnen maken. Maar getuigen moeten uiteraard niet kunnen zien of aanvoelen wie de figuranten zijn en wie de verdachte is.

34. *Het materiaal gebruikt bij foto-identificatietests moet worden bewaard voor later onderzoek, en moet automatisch ter hand worden gesteld aan de verdediging. Het proces-verbaal moet beschrijven in welke volgorde de foto's zijn aangeboden, of hoe de foto's waren gerangschikt als zij tegelijkertijd werden aangeboden.* Deze regel poogt controle van de line-up achteraf mogelijk te maken. Met deze regel stipt Wagenaar (1989) aan dat een confrontatie simultaan of sequentieel kan worden uitgevoerd. Volgens sommige onderzoekers geniet de sequentiële variant eigenlijk de voorkeur. Clark et al. (2008) vergeleken bijvoorbeeld de resultaten van 81 simultane en negen sequentiële procedures. Bij deze simultane confrontaties bleken er 47% terechte en 14% valse herkenningen op te treden. Na het wegen van deze percentages, kwamen de onderzoekers uit op een likelihood ratio van 5,5. Voor de

sequentiële confrontaties werden 40% terechte herkenningen, 8% valse herkenningen, en na weging, een likelihood ratio van 6,4 gevonden. Steblay et al. (2011) cumuleerden de resultaten van 49 onderzoeken naar de effecten van simultane en sequentiële confrontaties en vonden een likelihood ratio van 5,8 voor simultane, en 7,7 voor sequentiële procedures. Zij keken ook naar de likelihood ratio's (onder het primaire scenario) voor nonherkenningen. Voor simultane confrontaties bleek die 0,5 te zijn, en voor sequentiële 0,3. De reden voor de superioriteit van sequentiële confrontaties zou kunnen zijn dat het onmogelijk is om de meest lijkende persoon aan te wijzen. De getuige kan namelijk niet vergelijken, omdat alle personen (of foto's van hen) een voor een worden getoond. Daarom vergt een sequentiële procedure absolute beslissingen. Bij een simultane confrontatie zou de getuige, terwijl hij niemand echt positief herkent, alsnog de meest op zijn herinnering lijkende persoon kunnen identificeren. Dat zou kunnen leiden tot een toename van het aantal vals positieven. Om relatief kiezen onmogelijk te maken, mag de getuige, nadat hij een persoon heeft afgewezen als dader (en dus door is gegaan naar de volgende) niet meer terugkeren naar de vorige. Ook mag hij, nadat hij een persoon heeft geïdentificeerd, niet alsnog de rest van de line-up bekijken. Verder mag de getuige niet weten hoeveel personen hij te zien krijgt, om te voorkomen dat hij, eenmaal aangekomen bij de voorlaatste, denkt dat hij dan maar de laatste moet identificeren. Overigens is het aan te bevelen om in een sequentiële confrontatie, de verdachte niet op de laatste en ook niet op de eerste positie te zetten. Sommige onderzoekers betwisten overigens het verschil in bewijskracht tussen simultane en sequentiële confrontaties (Gronlund et al., 2014; Mickes et al., 2012). Verder zij op deze plek opgemerkt dat de Nederlandse politie veel van de door Wagenaar (1989) beschreven richtlijnen heeft overgenomen. In het *Handboek Confrontatie* (Van Amelsvoort, 2023) wordt de term Osloconfrontatie

overigens gereserveerd voor de simultane bewijsconfrontatie.

35. *Identificatietests met levende personen moeten op foto's worden vastgelegd. Zulke foto's moeten automatisch aan de verdediging ter hand worden gesteld.* Deze en de volgende twee regels dienen ter controle achteraf.

36. *Namen en adressen van deelnemende figuranten moeten worden vastgelegd, maar niet automatisch aan de verdediging ter hand gesteld. Wanneer de verdediging daarom vraagt, kunnen zij wel worden overhandigd.*

37. *Het is aan te bevelen dat van identificatietests met levende personen video-opnamen worden gemaakt, wanneer dat maar enigszins mogelijk is.*

38. *Aan deelnemers in een identificatietest moet nauwkeurig worden verteld hoe zij zich dienen te gedragen.* Ook deze regel beoogt de kans te reduceren dat de verdachte op de een of andere manier, zoals door zijn timide houding, afsteekt tegen de figuranten en aldus daardoor zou kunnen worden geïdentificeerd.

39. *Veranderingen die plaatsvinden in de periode tussen de misdaad en de identificatietest moeten zoveel mogelijk worden teniet gedaan.* Als er zodanig veel tijd is verstreken tussen het incident en de confrontatie dat de dader er mogelijk anders is gaan uitzien dan het signalement dat door de getuige is gegeven, kan het zo zijn dat de verdachte dus niet voldoet aan de beschrijving van de dader (maar toch de dader is). Dat is natuurlijk problematisch. Concreet: als de dader geen snor had, maar de verdachte wel, zou men aan de verdachte kunnen vragen om zijn snor te scheren ten behoeve van de confrontatie.

40. *Wanneer het maar mogelijk is moeten identificatietests worden afgenomen door politiebeambten die zelf niet weten wie in de testgroep de verdachte is.* Als het lukt om deze regel uit te voeren, ondervangt dat het risico dat de politieambtenaar de getuige (onbewust) beïnvloedt in de richting van een identificatie van de verdachte.

41. *Het aanbieden van foto's of tests met levende personen moet nauwkeurig worden beperkt tot een van te voren vastgelegde periode, en de getuigen moeten hiervan op de hoogte worden gesteld.* Uit onderzoek van bijvoorbeeld Dunning en Perretta (2002) blijkt dat accurate herkenningsbeslissingen in korte tijd worden genomen. Het is daarom niet nodig en zelfs niet zinnig om getuigen zeer lang te laten kijken naar de line-up. Dat leidt namelijk niet tot betere beslissingen.

42. *Alle relevante details van de voorbereiding en afname van identificatietests moeten worden vastgelegd.* Deze regel past in het rijtje maatregelen die het controleren van de confrontatie achteraf mogelijk maken.

43. *Er mag geen feedback aan getuigen worden gegeven over de vraag of zij de verdachte dan wel een figurant hebben aangewezen.* Het doel van deze regel is tweeledig. Op de eerste plaats voorkomt deze regel dat als er twee getuigen zijn die beiden de confrontatie doorlopen, de eerste getuige de tweede zou kunnen souffleren. Op de tweede plaats is het aan te raden om getuigen te vragen hoe zeker ze zijn van hun keuze (zie regel 47). Dit zelfvertrouwen is echter gevoelig voor feedback. In een onderzoek van Luus en Wells (1994) werd het verband tussen feedback en zelfvertrouwen sprekend geïllustreerd. Aan het onderzoek namen in totaal 140 psychologiestudenten deel. Zij ondergingen het experiment in tweetallen. Bij aankomst in het lab werd hun verteld dat ze zouden gaan proberen om auditieve stimuli te koppelen aan visuele stimuli. Het ging met name erom stemmen aan gezichten te koppelen. Hun werd ook meegedeeld dat ze zouden worden gefotografeerd in het kader van het onderzoek. Toen ze naar een kamer werden gewezen waar het onderzoek zou plaatsvinden, troffen ze daar een vrouw aan die verschrikt opkeek en vervolgens via een andere deur de kamer verliet met een apparaat onder haar arm. Vlak daarna kwam de experimentleidster binnen en werd na enig overleg geconcludeerd dat de betreffende vrouw

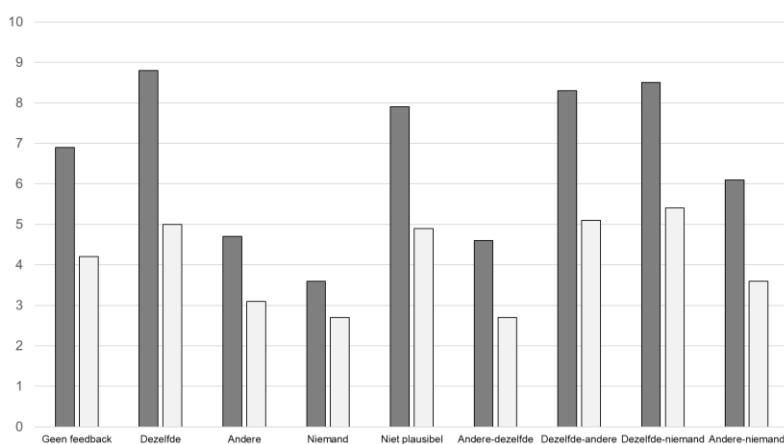
een essentieel apparaat had ontvreemd. Hoogstwaarschijnlijk was de dievegge een studente die eerder had deelgenomen aan dit onderzoek. Niemand anders was namelijk op de hoogte van de aanwezigheid van de waardevolle apparatuur in deze kamer. Gelukkig waren foto's voorhanden van alle zes vrouwelijke studenten die eerder hadden deelgenomen aan dit onderzoek. De proefpersonen werden elk naar een kamer gebracht alwaar ze moesten beoordelen of de door hen geziene dievegge inderdaad in de fotoserie aanwezig was. De experimentleidster vertelde de deelnemer dat de andere proefpersoon zojuist de foto's had gezien en vergeleek de herkenning van de proefpersoon met die van de andere proefpersoon. Uiteindelijk waren er negen verschillende soorten feedback. Sommige proefpersonen kregen bijvoorbeeld te horen dat hun collega-proefpersoon niemand had aangewezen. Anderen werd verteld dat hun collega een andere foto aanwees. Sommigen ervoeren dat hun collega dezelfde foto had aangewezen. Dan was er nog een groep deelnemers die te horen kreeg dat hun collega een andere foto had aangewezen, terwijl de op die foto afgebeelde vrouw totaal niet op de dievegge leek (een niet plausibele optie). Uiteraard was er een controleconditie waarin proefpersonen helemaal geen feedback ontvingen. Tot slot waren er vier condities waarin de feedback veranderde. In een van die condities meldde de experimentleidster aanvankelijk dat de collega dezelfde foto had aangewezen, maar corrigeerde ze zichzelf binnen twee minuten en concludeerde ze dat de collega toch een andere had aangewezen. Zo ook was er een conditie waarin geswitcht werd van dezelfde foto naar een nonherkenning, een waarin de switch van een andere foto naar een nonherkenning ging en een conditie waarin de feedback wisselde van een andere foto naar dezelfde. Deze variabele feedback werd gegeven om te achterhalen of het zelfvertrouwen – net als de feedback – veranderbaar is.

Na afloop van de fotoconfrontatie werden de proefpersonen geïnterviewd door een beveiligingsmedewerker. Het tien minuten durende interview werd gefilmd. Tijdens dit interview vroeg de beveiligingsmedewerker onder andere hoe zeker de proefpersoon zich voelde over de juistheid van de herkenning. De proefpersoon diende dit aan te geven door een getal te noemen tussen 1 (*helemaal onzeker*) en 10 (*heel zeker*).

Wat de deelnemers niet wisten, is dat de diefstal geënceneerd was. Verder beoogden de onderzoekers om zo veel mogelijk deelnemers te verleiden tot een valse herkenning. Daartoe zetten ze vier middelen in. Zo was de foto van de dievegge niet aanwezig in de aan getuigen getoonde reeks van zes. Elke herkenning was dus per definitie vals. Daarnaast was de diefstal zodanig geënceneerd dat de blootstellingstijd kort was. De proefpersonen zagen de dievegge dus niet goed. Verder leken vijf van de zes foto's absoluut niet op de dievegge, terwijl één foto enigszins leek. De instructie luidde: "Which of these women did it?" (Luus & Wells, 1994, p. 717). Daarmee werd een cruciale clausule, inhoudende dat de dader er mogelijk niet bij zat, aan de deelnemers onthouden. Tot slot moet worden opgemerkt dat het in werkelijkheid ging om *bogus* feedback; de feedback was gefingeerd en maakte deel uit van de experimentele manipulatie.

De onderzoekers vonden dat 97% van de deelnemers inderdaad werd verleid tot een valse herkenning. Slechts 4 van de 140 proefpersonen weigerden iemand aan te wijzen. De zelfgerapporteerde zekerheid in de juistheid van de herkenning wordt weergegeven in Figuur 7 (donkergekleurde balk). Het valt op dat het vertrouwen in afwezigheid van feedback relatief groot was (6,9), ondanks dat alle deelnemers zich vergisten. Ook bleek dat het effect van de initiële feedback niet werd gecorrigeerd door aanpassing van die feedback.

In het tweede deel van het onderzoek werden de video-opnamen van de getuigenverklaringen getoond aan 378 studenten. Elke student in dit deel van het onderzoek zag één filmfragment. Daarna moesten de deelnemers aangeven “to which extent they believed that the witness had made an accurate identification” (Luus & Wells, 1994, p. 719). Dit deden zij op een schaal van 1 (*helemaal niet accuraat*) tot 7 (*zeer accuraat*). De resultaten van dit deel van het onderzoek worden ook weergegeven in Figuur 7 (lichtgekleurde balk). De data volgen het patroon van het zelfgerapporteerde vertrouwen. Dit betekent dat de zekerheid van proefpersonen uit het eerste deel van het onderzoek door de deelnemers aan het tweede deel werd geïnterpreteerd als indicator van de herkenningssuccesrate, hetgeen in casu onterecht was.



Figuur 7. Zelfvertrouwen (donker; 1-10) en waargenomen accuraatheid (licht; 1-7; Luus & Wells, 1994).

In hun overzichtsartikel vonden Sporer et al. (1995) een correlatie tussen zekerheid en accuraatheid ($r = 0,37$), maar alleen voor getuigen die iemand aanwezen. De correlatie voor de nonherkenkers ($r = 0,12$) was niet significant. Met een alternatieve analyse berekenden de onderzoekers dat getuigen die de correcte persoon aanwezen gemiddeld 76% scoorden op een schaal voor

zekerheid, terwijl de vergissers 54% scoorden. Voor de nonherkenners was het gemiddelde 71% onder de terechte nonherkenners en 64% onder de vals negatieve herkenners. Het verband tussen zelfvertrouwen en accuratesse is er dus wel, maar het is conditioneel en ietwat kwetsbaar. Niettemin wordt in de recente literatuur gepropageerd dat zelfvertrouwen als een kenmerk van accuratesse moet worden gezien (Wixted & Wells, 2017). Nu is het een ding om de herkenning van een getuige die onzeker is met voorzichtigheid te betrachten, maar een ander ding om de herkenning van een zekere getuige, vanwege die zekerheid, sterker te achten. Niettemin concluderen Wells et al. (2020) dat "eyewitness confidence is a useful cue to the accuracy of a witness's decision" (p. 22). Zij adviseren verder om de zekerheid na te vragen op bijvoorbeeld een schaal van 0% tot 100%, en wel onmiddellijk na de (non)herkenning: "Note that forcing the witnesses to withhold their confidence reports for as little as 5 min has been found to undermine the predictive value of confidence" (p. 22).

44. *De advocaat van de verdachte moet in staat worden gesteld bij identificatietests aanwezig te zijn.* Deze regel dient ter bevordering van transparantie en controle.

45. *Identificaties moeten niet worden aanvaard wanneer getuigen verklaren dat zij alleen maar een gelijkenis zien.* In beginsel dienen alle personen in de line-up op elkaar te lijken en ook op de dader, omdat het signalement zoals gegeven door de getuige in de meeste gevallen leidend is bij de samenstelling. Er is pas sprake van een identificatie als de getuige positief uitspreekt dat hij een van de getoonde personen herkent als de dader.

46. *Volledige herkenningen die slechts na een zeer lange bedenktijd tot stand komen moeten niet aanvaard worden als positieve identificaties.* Correcte herkenningsbeslissingen worden veelal vrij snel genomen, zelfs binnen 10 tot 12 seconden (Dunning & Perretta, 2002). Van de beslissingen die in deze korte tijd worden genomen bij een

uit zes foto's bestaande line-up bleek 90% correct. Van de langzamere beslissingen was nog slechts 50% correct.

47. *Andere uitdrukkingen van onzekerheid moeten worden geregistreerd, en kunnen leiden tot een verwerping van de identificatie.* Wagenaar (1989) betoogde niet dat herkenningen door zelfverzekerde getuigen als accuraat moeten worden gezien (anders dan Wells et al., 2020), maar wel dat herkenningen van twijfelende getuigen minder serieus moeten worden genomen.

48. *Identificaties in de rechtszaal kunnen nooit worden aanvaard als bewijs van identiteit.* Mocht de getuige ter zitting worden gevraagd of hij de verdachte herkent, dan worden daarmee tal van regels voor het correct uitvoeren van een bewijsconfrontatie geschonden. Bovendien is er dan sprake van een show-up. Daarbij komt dat de context sturend is. De getuige weet immers dat de hem "getoonde" persoon de verdachte is, en het zou heel wat vragen van de getuige om ter zitting te verklaren dat hij de verdachte niet herkent.

49. *Een overtreding van de regels die niet in het proces-verbaal wordt gerechtvaardigd moet leiden tot een verwerping van de identificatietest.* Welbeschouwd is niet elke door Wagenaar (1989) genoemde regel even cruciaal voor een valide uitslag. Niettemin toonde hij zich met deze regel vrij streng.

50. *Wanneer wordt aangetoond dat overtreding van regels de validiteit van de test aanzienlijk heeft verkleind, moet worden besloten om de gehele test te verwerpen, zelfs wanneer de overtreding is gerechtvaardigd in het proces-verbaal.* Sommige van de regels zijn cruciaal voor de waarborging van de validiteit van de (positieve) uitkomst. Als dergelijke regels worden geschonden, zal de resulterende herkenning slechts schijnbewijs tegen de verdachte opleveren. Met deze laatste regel probeert Wagenaar (1989) dat te voorkomen.

De vijftig regels van Wagenaar bouwen waarborgen uit een wetenschappelijk experiment in in een confrontatie-

procedure: er is bijvoorbeeld sprake van een “blinde” begeleider om de getuige niet te beïnvloeden, en er zijn “controle”-condities in de vorm van figuranten. Daarmee laat Wagenaar zien hoe een bewijsprocedure kan profiteren van wetenschappelijke principes. Wells en Loftus zeiden het in 1990 mooi: “Police lineups as experiments: Social Methodology as a framework for properly conducted lineups”, helaas zonder enige verwijzing naar het werk van Wagenaar (1989). Ook in 2020 deden Wells et al. nog alsof zij de eersten waren die ooit hadden nagedacht over hoe je een confrontatie kunt spekken met wetenschappelijke waarborgen: “The 1998 scientific review paper was the first set of science-based recommendations regarding how to conduct lineups that was endorsed by a scientific society” (p. 4). Hun aanbevelingen zijn een kopie van die van Wagenaar met uitzondering van de nummering en met een groot verschil. Waar Wagenaar ermee begon uit te leggen dat een Osloconfrontatie alleen zinnig is in het geval dat de getuige een hem/haar onbekend persoon een delict heeft zien plegen, zeggen Wells et al.: “prior familiarity with the culprit ... generally increases the accuracy of a witness’s description and identification” (p. 10). Dat klopt, maar als de politie een confrontatie overweegt, is nog niet duidelijk of de verdachte de dader is.

Tot slot zij opgemerkt dat de regels van Wagenaar (1989) wetenschappelijke regels zijn. Het is interessant om te zien dat die regels *grosso modo* zijn overgenomen in de opleiding van de Nederlandse politie (Van Amelsvoort, 2023). De rechter is er bij de waardering van herkenningen echter op geen enkele manier aan gehouden. Lang niet alle rechters zullen bekend zijn met deze kwaliteitsregels (zie ook Wise & Safer, 2004). Sterker nog, onder wetenschappers bestaat ook nog geen consensus over alle regels (Benton et al., 2006; Kassin et al., 2001).

THE MEMORY OF CONCENTRATION CAMP SURVIVORS

Tijdens de Tweede Wereldoorlog richtte de Duitse bezetter een strafkamp in in Nederland. In dat kamp, genaamd Erika, heeft een tot kapo bevorderde gevangene, Marinus de Rijke, zich bijzonder wreed gedragen tegenover de gevangenen. Toen het kamp in 1943 werd gesloten, zijn meerdere ex-gevangenen als getuige gehoord over de misdaden van De Rijke. Deze verhoren zijn gedocumenteerd in de periode 1943-1948. Veertig jaar later, tussen 1984 en 1988, werd er opnieuw onderzoek gedaan naar De Rijke, en toen zijn weer getuigen gehoord. Sommige mensen zijn zowel in de jaren veertig als in de jaren tachtig gehoord. Dat biedt een unieke mogelijkheid om te bestuderen hoe consistent ons geheugen is. Wagenaar en Groenweg (1990) kregen toegang tot de dossierstukken en schreven er een artikel over.

Allereerst verdient het onderzoek enkele kritische kanttekeningen. De gegevens zijn om meerdere redenen uniek. Zo hebben sommige getuigen traumatische ervaringen gehad, variërend van het ondergaan van mishandelingen tot het waarnemen van executies. Daarom kunnen herinneringen aan dergelijke trauma's worden vergeleken met meer alledaagse herinneringen, zoals die aan namen en uiterlijk van andere aanwezigen in het kamp. Die vergelijking is interessant omdat er meningen zijn over verschillen tussen traumatische en niet-traumatische herinneringen. Zo zouden de eerstgenoemde "op het netvlies staan gebrand" en niet onderhevig zijn aan vergeten en andere geheugenperikelen. Op de tweede plaats laten de gegevens toe om het zeer langetermijngeheugen te onderzoeken. Een retentieperiode van 40 jaar is voor reguliere experimenten onmogelijk. Omdat sommige getuigen zowel in het eerste als in het tweede onderzoek zijn gehoord, is tot slot de temporele

consistentie te onderzoeken. Afgezet tegen een laboratoriumexperiment heeft dit onderzoek ook nadelen. Die worden veroorzaakt doordat de data niet op een wetenschappelijke manier zijn verzameld. Zo kregen verschillende getuigen soms verschillende vragen. Getuigen die twee keer zijn gehoord, kregen de eerste keer andere vragen dan de tweede keer. Deze ongestructureerde manier van dataverzameling maakt vergelijkingen complex. Omdat het een veldstudie is, weten de onderzoekers zelf niet wat er daadwerkelijk is gebeurd. Ze kennen de *ground truth* niet. Dit maakt de interpretatie van bevindingen bij voorbaat moeilijk. Stel dat een getuige in de jaren 40 details noemt, die hij in de jaren 80 niet noemt. Is hij die dan vergeten, komt dat vergeten door tijdsverloop of doordat hij oud is, is er echt sprake van vergeten, of is er in de jaren 80 niet goed doorgevraagd, had hij toen geen zin om alles te vertellen? Stel dat iemand juist in de jaren 80 details noemt die hij niet in de jaren 40 noemde. Zijn dat dan details die vertraagd zijn bovengekomen, of zijn dat verzinsels, pseudoherinneringen? Wat hebben de getuigen in de voorbijgaande 40 jaar gedaan, hebben ze vaak gedacht en gesproken over kamp Erika, of juist niet?

De belangrijkste bevindingen van Wagenaar en Groeneweg (1990) zijn als volgt: "Camp experiences were generally well-remembered, although specific but essential details were forgotten. Among these were being maltreated, forgetting names and appearances of the torturers, and forgetting being a witness to murder. Apparently, intensity is not a sufficient safeguard against forgetting" (p. 77). Meer in detail vonden zij dat bijna alle getuigen, ook na 40 jaar, nog wisten hoe de dagelijkse gang van zaken was en hoe de barakken eruit zagen. Van de 78 getuigen vergisten elf zich in de jaren 40 in de datum waarop ze het kamp in gingen. In de jaren 80 vergisten zich 19 getuigen in de datum. Dit is een voorbeeld van de achteruitgang van herinneringen door tijdsverloop en/of ouderdom. Zes getuigen sloegen niet aan op de naam De

Rijke. Veertien getuigen herkenden De Rijke's foto niet tijdens het tweede onderzoek. Opmerkelijk is dat dit patroon van vergeten hetzelfde was onder getuigen die mishandeld waren als onder hen die dat niet waren geweest. Verder waren er getuigen die in eerste instantie verklaarden te hebben gezien hoe een medegevangene werd verdrinken, maar dit later niet meer meldden en desgevraagd zelfs zeiden dat nooit te hebben verklaard. Ook waren er fouten in de zin dat een getuige aanvankelijk iemand als slachtoffer noemde maar later juist als dader van een mishandeling.

Op basis van de bevindingen concluderen Wagenaar en Groeneweg (1990) dat traumatische herinneringen niet fundamenteel afwijken van reguliere herinneringen. Ze zijn in de kern soms opvallend duurzaam, maar er kunnen ook opmerkelijke spontane inconsistenties optreden (Schmidt et al., 2000). Daarin staan de auteurs niet alleen.

Recenter onderzochten Engelhard et al. (2008) het geheugen van 133 infanteristen aan hun vier maanden durende uitzending naar Iraq in 2004. De militairen werden twee keer geïnterviewd, na vijf en na vijftien maanden na hun terugkomst van de missie. Op beide meetmomenten kregen de soldaten 22 vragen over traumatische gebeurtenissen die ze al dan niet hadden meegemaakt. Tabel 11 geeft voorbeelden van de items, inclusief de (in)consistentie van de antwoorden tussen de eerste en de tweede meting. Zoals te zien, was er sprake van een behoorlijke inconsistentie. Gemiddeld werd elk van de 22 items door 17,5% van de deelnemers inconsistent beantwoord. Ook valt op dat vergeten (ja-nee verandering) voorkomt, maar toevoegen (nee-ja veranderingen) evenzeer. In totaal veranderden de 133 militairen 259 keer van ja naar nee, en 249 keer van nee naar ja. In afwezigheid van een gouden standaard inzake de waarheid, is het onmogelijk om te beoordelen of een verandering een verbetering of een verslechtering is.

Tabel 11. Aantallen consistente en inconsistente antwoorden ($N = 133$) op enkele vragen over het al dan niet meemaken van traumatische ervaringen na vijf en vijftien maanden (Engelhard et al., 2008).

	Consistent		Inconsistent		Percentage inconsistente antwoorden
Na vijf maanden	Nee	Ja	Ja	Nee	
Na vijftien maanden	Nee	Ja	Nee	Ja	
<i>Seeing dead or injured non-Dutch soldiers.</i>	79	16	22	16	29%
<i>Seeing dead or injured Dutch soldiers.</i>	55	52	13	13	20%
<i>Witnessing an explosion.</i>	20	90	14	9	17%
<i>Witnessing violence.</i>	15	10	13	5	14%
		0			
<i>Being shot at.</i>	33	75	6	19	19%
<i>Being injured due to an assault.</i>	109	5	14	5	14%
<i>Being sexually harassed.</i>	116	3	11	3	11%
<i>Fearing to be taken hostage.</i>	16	88	18	18	27%

Een ander voorbeeld van de kwetsbaarheid van traumatische herinneringen is te vinden in een onderzoek van McNally et al. (2004). Deze onderzoekers stelden tien mensen met pseudoherinneringen aan een ontvoering door buitenaardse wezens bloot aan transcripten van hun ontvoering en aan transcripten van daadwerkelijk meegemaakte vervelende gebeurtenissen los van de vermeende ontvoering (bijvoorbeeld een pijnlijke tandartsbehandeling). Zij ontdekten dat de confrontatie met de pseudoherinnering even stresserend was als die met de echte herinnering, zowel indien gemeten met een zelfrapportage als met fysiologische reacties. Hieruit valt af te leiden dat traumatische herinneringen in werkelijkheid hardnekkige pseudoherinneringen kunnen zijn.

Net zoals alle andere herinneringen kunnen traumatische herinneringen dus fouten bevatten. Samengevat in de woorden van Wagenaar en Groeneweg

(1990): "There is no reason to distrust such testimony more than in other violent crimes. The degree of conflict between the testimonies in the De Rijke case is probably quite normal (p. 87).

Volgens de auteurs kunnen de gevonden inconsistenties niet worden geattribueerd aan perikelen tijdens waarneming en *encoding*. Daarbij zij opgemerkt dat het geheugen welbeschouwd een verzameling processen is. Allereerst moet de informatie worden waargenomen en opgeslagen (*encoding*). Dan moet de informatie een tijd worden vastgehouden in het langetermijngeheugen (*retentie* of *storage*). Tot slot moet de herinnering onder bewuste controle weer worden opgehaald (*retrieval*) en worden herinnerd in het werkgeheugen. Bij al deze processen kan iets misgaan. Om te beginnen is er nogal wat onderzoek dat erop duidt dat wij relatief oppervlakkig waarnemen en daarom menig detail missen. Klassiek is het onderzoek waarbij wandelaars op straat worden aangesproken door een vreemde die om richting vraagt. Tijdens het gesprek lopen twee werkers met een houten plank tussen de proefpersoon en de vreemdeling door. Als zij zijn gepasseerd heeft een ander persoon stiekem, terwijl ze uit beeld van de proefpersoon waren omdat de houten plank het zicht blokkeerde, de plaats van de vreemdeling ingenomen. Ongeveer een derde van de proefpersonen bleek deze verandering niet op te merken, hetgeen illustreert hoe oppervlakkig we waarnemen (Simon & Levin, 1998). Doordat we relatief weinig aandacht kunnen schenken aan de overvloed aan informatie in onze omgeving, vallen dergelijk veranderingen, vooral als ze buiten ons zicht, of gradueel plaatsvinden niet op. In de literatuur wordt gesproken van *change blindness* (Rensink, 2002).

In een klassiek onderzoek van Simons en Chabris (1999) bleek het bijna de helft (46%) van de observatoren niet op te vallen als er midden tussen een groep basketballers een verkleed persoon in een gorillapak

opdook. De observatoren waren te geconcentreerd op de wedstrijd aan het letten om deze misplaatste figuur waar te nemen. De auteurs spreken van *inattentional blindness*. Dit fenomeen heeft een voor getuigen funeste manifestatie te weten het *weapon focus* effect: de meest opvallende stimulus slokt al onze aandacht op, waardoor we andere stimuli niet waarnemen (Stebly, 1992). Dit is frappant, want blijktbaar is te oppervlakkig waarnemen niet goed voor onze gedetailleerde herinnering achteraf, maar gaat een te geconcentreerde waarneming eveneens ten koste van de kwaliteit, omdat we het overzicht missen. Deffenbacher (1985) introduceerde dan ook de *Yerkes-Dodsoncurve*: een parabool die het verband tussen de aandacht (emotionele betrokkenheid) bij de waarneming en de kwaliteit van de herinnering naderhand weergeeft. In deze curve, ligt een optimale waarneming ergens in het midden.

Niet alleen is onze aandacht beperkt en wordt deze soms automatisch gestuurd door de omgeving, onze waarneming wordt ook beïnvloed door voorkennis en verwachtingseffecten (Balcetis & Dunning, 2006). We completeren onze waarneming op basis van onze voorkennis. Verwachtingen kunnen ertoe leiden dat we dingen waarnemen die er in het geheel niet zijn (hallucinaties). Merckelbach en Van de Ven (2001) lieten 44 proefpersonen luisteren naar drie minuten witte ruis en vertelden dat er in die ruis fragmenten van Bing Cosby's *White Christmas* zouden kunnen voorkomen. Hoewel er in werkelijkheid geen muziek voorkwam in de opname, gaf 32% aan de muziek toch te hebben gehoord. In een vervolgstudie met 108 deelnemers vonden Van de Ven en Merckelbach (2003) een vergelijkbaar percentage mensen (35%) dat een auditieve hallucinatie ervoer. Rassin en Van der Heiden (2006) onderwierpen 50 deelnemers aan drie minuten bruine ruis gecombineerd met muziekfragmenten van Roxy Music en Cabaret Voltaire, en vertelden vooraf dat er in het geluidsfragment vloeken hoorbaar zouden kunnen zijn. Ook deze instructie was vals, maar niettemin gaf 74%

van de deelnemers aan een of meerdere vloeken te hebben gehoord.

Zelfs als een getuige het incident onder optimale omstandigheden heeft kunnen waarnemen, is daarmee nog niet zeker dat de af te leggen verklaring over dit incident valide zal zijn. Daarvoor is ook opslag in het geheugen nodig. Dat gaat niet altijd goed en de vergeetcurve is steil. Om te beginnen, komt alle informatie in het korte termijn geheugen (*short term memory*; STM). Het STM duurt maar enkele seconden of maximaal minuten en is beperkt in omvang: er passen slechts ongeveer zeven stukjes informatie in (Miller, 1956). Denk aan een getal, bijvoorbeeld 83549863. Het kost moeite om dat te onthouden, en de beste manier om dat te doen, oftewel om het te verplaatsen naar het langetermijngeheugen (*long term memory*, LTM) is herhaling en of *chunking* tot minder informatie (bijvoorbeeld 83-54-98-63; Sternberg, 1996). Veel informatie gaat verloren omdat het niet wordt opgeslagen in het LTM. Het opslaan van het 8-cijferige getal is al een uitdaging, maar met een woordenreeks wordt het wellicht nog moeilijker. Neem als voorbeeld de reeks "noot, geluid, piano, zang, radio, band, melodie, hoorn, concert, instrument, symfonie, jazz, orkest, kunst en ritme". Als we een beperkte hoeveelheid tijd hebben, zal het moeilijk zijn om alle vijftien woorden te onthouden. In theorie zullen de eerste woorden goed blijven hangen. Dat komt doordat daar niets aan voorafging. Er is dus geen proactieve of anterograde interferentie. Naast het *primacy*-effect is er ook een (iets kleiner) *recency*-effect voor de laatste woorden. De laatste woorden hebben namelijk geen concurrentie van opvolgers, er is dus geen retrograde interferentie. De middelste woorden worden in het algemeen het slechts onthouden.

In de literatuur wordt onderscheid gemaakt tussen episodische en semantische herinneringen. De eerste categorie betreft herinneringen aan concrete gebeurtenissen. De tweede categorie is abstracter. Het gaat

niet om concrete gebeurtenissen maar om algemene kennis of betekenis. Wij weten bijvoorbeeld wat het woord "stoel" betekent. Die betekenis is opgeslagen in ons geheugen. Maar we weten waarschijnlijk niet meer wanneer we het woord stoel hebben geleerd. Voor getuigen is vooral het episodisch geheugen belangrijk. Zij moeten weten wat wanneer gebeurde. Uit onderzoek blijkt dat blijven herhalen van herinneringen in eerste instantie leidt tot opslag als episodische herinnering in het LTM, maar vervolgens ook kan leiden tot een overgang naar een meer semantische herinneringen die concrete episodische details verloren heeft. Zo ontdekten Van den Hout en Kindt (2003) dat hoe vaker we eenzelfde handeling uitvoeren (bijvoorbeeld het gasfornuis uitdraaien, het licht uitdoen, of de deur afsluiten), des te minder zeker we worden dat we die handeling daadwerkelijk hebben uitgevoerd. Dat komt waarschijnlijk doordat de herhaling leidt tot een semantisch karakter van de herinneringen, hetgeen ten koste gaat van concrete details.

Eenmaal opgeslagen in het LTM, dient de informatie daar een tijd te blijven. Helaas is ook dat niet vanzelfsprekend. Jenkins en Dallenbach (1924) bespraken een onderzoek van Ebbinghaus uit 1885 waarin proefpersonen een reeks onzinwoorden dienden te onthouden. Na 20 minuten waren ze echter al 41,8% vergeten. De deelnemers werden vervolgens nog herhaaldelijk gevraagd alle woorden te reproduceren. Hun prestatie nam af met het verstrijken van de tijd. Na een uur was 55,8% vergeten, na 8,8 uur was 64,2% vergeten, na 24 uur was 66,3% vergeten, na twee dagen was 72,2% vergeten, na zes dagen was 74,6% vergeten en na 31 dagen was maar liefst 78,9% weg. Dit onderzoek laat zien dat de vergeetcurve inderdaad steil is, vooral aanvankelijk. Recenter onderzoek laat zien dat naarmate we informatie oppervlakkiger opslaan (een reeks betekenisloze woorden slaan we minder "diep" op dan een verhaal, of een traumatische ervaring), we hem sneller vergeten. Het

nadeel van een “diepere” opslag is dan wel weer dat die meer semantisch is en dus relatief weinig gedetailleerd (Fisher & Radvansky, 2018).

Tijdsverloop zorgt niet alleen voor een kwantitatief verlies van details, maar een complexe herinnering aan een gebeurtenis kan ook kwalitatief achteruitgaan. Johnson et al. (1988) gaven 72 studenten vier geheugenopdrachten. Zij dienden zich een recente gebeurtenis, zoals een ontmoeting, bibliotheek-, of tandartsbezoek voor de geest te halen, eenzelfde gebeurtenis maar dan uit hun jeugd, en ze dienden een droom, fantasie, of niet uitgevoerde intentie (een recente en een uit hun jeugd) voor de geest te halen. Vervolgens scoorden ze elk van deze vier mentale voorstellingen op 39 kenmerken, op een schaal van 1 tot 7. Met deze *Memory Characteristics Questionnaire* (MCQ) probeerden de onderzoekers te achterhalen of er verschillen zijn tussen oude en recente herinneringen en tussen herinneringen aan echte gebeurtenissen en “intern gegenereerde” gebeurtenissen. De vier mentale voorstellingen verschilden niet op alle 39 kenmerken, maar die waarop er wel verschillen bestonden, worden weergegeven in Tabel 12.

De 39 items van de MCQ worden soms geclusterd tot hoofdcategorieën zoals *clarity*, *sensory information*, *context*, *thoughts and feelings*, en *intensity* (Suengas & Johnson, 1988). In het algemeen lijken herinneringen aan echte gebeurtenissen iets hoger te scoren dan intern gegenereerde voorstellingen. Ook lijken recentere voorstellingen hoger te scoren dan oudere. Dat betekent ook dat recente fantasieën kwalitatief lijken op oude herinneringen aan echte gebeurtenissen. Als we de levendigheid (de MCQ-score) van onze mentale voorstellingen gebruiken als maat voor de echtheid van de inhoud, zullen we ons dus soms vergissen. Er treedt dan een *source monitoring error* (misattributie) op. Zo vroegen Rassin et al. (2001) aan twee groepen studenten of ze wel eens realistisch droomden. Die vraag werd door 257 van de

340 respondenten (76%) bevestigend beantwoord. 13,8% Van de proefpersonen gaf aan herinneringen te koesteren waarvan ze niet wisten of die teruggingen op een echte gebeurtenis. 8,8% Van hen gaf aan ooit iemand te hebben aangesproken op basis van een herinnering, die toen niet op een echte gebeurtenis bleek te zijn gebaseerd.

Tabel 12. MCQ-scores (1-7) van verschillende mentale voorstellingen (Johnson et al., 1988).

	Recente gebeurtenis	Recente fantasie	Jeugd gebeurtenis	Jeugd fantasie
Visual details	6.0	5.1	5.2	4.6
Sound	4.7	3.1	3.8	3.4
Smell	3.0	1.8	2.3	1.9
Taste	3.1	1.6	2.5	1.3
Complexity	2.4	3.4	2.8	3.3
Realism	6.4	5.0	5.8	3.9
Location	6.6	5.5	6.2	5.8
Setting	6.0	4.9	5.8	5.7
Spatial object arrangement	6.1	4.5	5.0	4.7
Time	5.9	4.0	4.0	4.0
Year	6.6	4.9	4.1	3.3
Season	6.1	4.8	3.8	3.7
Day	4.8	2.9	2.6	2.1
Hour	5.1	2.9	3.2	3.0
Emotional tone	5.5	4.5	4.8	4.5
Seeming implications	3.1	4.8	3.6	3.4
Intensity	4.4	5.4	4.8	5.0
Prior events	4.7	3.1	2.7	2.5
Posterior events	4.9	3.1	3.1	2.5
Covert rehearsal	3.3	4.5	3.0	3.5

Uit onderzoek blijkt kortom dat van alles wat we potentieel kunnen onthouden, een aanzienlijk deel niet ons LTM bereikt, en zelfs van de informatie die dat wel doet, wordt een groot deel alsnog vergeten. Naast deze normale beperking van ons geheugen, is er in sommige gevallen sprake van amnesie als een psychiatrisch symptoom. In de *Diagnostic and Statistical Manual of Mental Disorders* (DSM-

5; American Psychiatric Association; APA, 2013) worden verschillende syndromen gedefinieerd waarin amnesie een symptoom is. Zo is er bij *posttraumatic stress disorder* (PTSD) sprake van “inability to remember an important aspect of the traumatic event (typically due to dissociative amnesia)”. Dissociatieve amnesie is op zichzelf een stoornis volgens de DSM-5. Deze niet biologisch verklaarbare amnesie kan lokaal (betrekking hebbend op een afgebakende periode), systematisch (betrekking hebbend op een bepaald onderwerp of persoon) of gegeneraliseerd (volledig) zijn.

Voor de waarheidsvinding is het verloren gaan van waardevolle informatie ongelukkig. Er zijn dan ook pogingen gedaan om technieken te ontwikkelen die vergeten informatie op de een of andere manier toch weer naar boven kunnen halen (retrieval). Hierbij wordt er vanuit gegaan dat vergeten informatie nog steeds in het LTM aanwezig is, maar niet meer te vinden is. In de woorden van Schacter (1999) is er dan sprake van tijdelijk vergeten (*blocking*), dit in tegenstelling tot permanent vergeten ofwel *transience*. Hypnose is lange tijd als een methode gezien waarmee vergeten informatie alsnog kan worden opgehaald. De kracht van hypnose is erin gelegen dat de gehypnotiseerde persoon zich sterk concentreert op de stem van de hypnotiseur. Deze concentratie zou helpen bij het terughalen van de herinnering. De afgelopen decennia is forensische hypnose echter ietwat in ongebruik geraakt. Probleem is namelijk dat door de sterke concentratie op de hypnotiseur, diens instructies en vragen een zwaar aandeel krijgen in het herinneren door de gehypnotiseerde. Als de hypnotiseur (per ongeluk) een suggestieve vraag stelt, zal dat een grote invloed hebben op de herinnering van de getuige (Kebbell & Wagstaff, 1998; Pinizzotto, 1989). Daardoor kan de verklaring onbetrouwbaar worden en/of zelfs geacht worden in strijd met de verklaringsvrijheid te zijn afgelegd. De Hoge Raad heeft gedicteerd dat een onder hypnose afgelegde verklaring van een verdachte (*NJ* 1985,

135), noch die van een getuige (*NJ* 1998, 798) als bewijsmiddel kan worden gebruikt.

Een eenvoudige geheugenverbeterende verhoormethode is het sluiten van de ogen tijdens het ophalen van herinneringen. Perfect et al. (2008) lieten hun proefpersonen een film zien en stelden er vervolgens 15 vragen over. Sommige deelnemers kregen een algemene vrije reproductie-instructie (*free recall*), anderen dienden hun ogen te sluiten. Over vijf van deze onderzoeken, antwoordden de deelnemers in de controleconditie gemiddeld 4,8 vragen correct, vergeleken bij 7 in de gesloten ogen conditie. In de controleconditie werden gemiddeld 2 vragen fout beantwoord (de rest kon worden beantwoord met een weet-niet-optie), vergeleken bij 1,4 in de gesloten ogen conditie.

Een methode waarnaar inmiddels een behoorlijke hoeveelheid onderzoek is gedaan, is het cognitieve interview (CI; Geiselman et al., 1986). Dit interviewprotocol kent vier opmerkelijke aspecten. Om te beginnen is er de context *reinstatement* instructie: de getuige moet in gedachte teruggaan naar vlak voor het incident. Door het terughalen van de context, wordt het geheugen aangewakkerd. Het geheugen werkt namelijk associatief en het oproepen van de context, roept dus (hopelijk) ook allerlei relevante herinneringen op. In hun klassieke onderzoek lieten Godden en Baddeley (1975) achttien duikers een woordenlijst bestuderen ("gewoon" aan land, of via een geluidsverbinding, onderwater). Vervolgens dienden de duikers de woorden te herinneren, wederom aan land, of onder water. Duikers die de woorden in dezelfde context opsloegen en ophaalden, konden zich gemiddeld meer woorden herinneren (12,5) dan duikers voor wie de encoding en retrieval context verschilden (8,5). Dit verschil bleek in een tweede studie niet te kunnen worden verklaard doordat de laatstgenoemde duikers hun duikpak moesten aan- of uittrekken tussen encoding en retrieval. Duikers die in dat tweede onderzoek namelijk nog meer afleiding

moesten ondergaan, bijvoorbeeld door extra rondjes te zwemmen, bleken namelijk niet slechter te gaan presteren. Daarom kon de *disruption* hypothese worden verworpen en bleek het verschil dus inderdaad aan de context(verandering) toe te schrijven. Vandaar dat wordt gesproken van *context-dependent* of *state-dependent learning*.

Dat vergeten soms slechts een tijdelijk fenomeen is, laat zich ook mooi illustreren met het *tip-of-the-tongue* fenomeen. En dat associatie een grote rol speelt, valt af te leiden uit *flashbulb memories*: dat zijn herinneringen aan indrukwekkende publieke gebeurtenissen zoals rampen, die wij ons vaak herinneren in combinatie met de context waarin we het nieuws voor het eerst hoorden (Brown & Kulik, 1977).

Naast context reinstatement, kent het CI de *report everything* instructie. Ook hier geldt dat details die op zichzelf wellicht minder relevant zijn niettemin kunnen dienen als retrieval cue voor andere, wel relevante details. Bovendien moet deze instructie de getuige ervan doordringen dat hij "alles" moet vertellen. Als getuige weet je immers niet altijd welke informatie voor de politie interessant is. In plaats van het zelf aanleggen van interne criteria voor de vermeldenswaardigheid, kan de getuige beter alles vertellen, waarna de politie de verkregen informatie filtert. Dit geldt temeer in het opsporingsonderzoek.

De derde en de vierde instructies dienen niet zozeer om meer informatie te herinneren, maar juist om te voorkomen dat de getuige details in de verklaring inbouwt die niet daadwerkelijk stoelen op herinneringen, maar die passen in het script dat hij heeft van het waargenomen incident. Zo kan een getuige van een bankoverval al snel denken donkere bivakmutsen en machinegeweren te hebben gezien, omdat dat wellicht past in het plaatje van een overval. Om scriptactivatie te drukken kan de *changed perspective* en de *changed chronology* instructie worden

gegeven. De getuige moet dan zijn waarneming vertellen vanuit een ander perspectief en/of in een niet-chronologische volgorde. Dat is moeilijk, waardoor de standaard voorstelling van het incident minder kans moet krijgen om de daadwerkelijke herinnering te verdrukken.

In een meta-analyse van het CI (Memon et al., 2010) waarin 46 onderzoeken en 2.887 proefpersonen werden betrokken, bleek het CI inderdaad tot meer correct herinnerde details te leiden dan een standaard herhaalde ondervraging ($d = 1,2$), maar ook tot een lichte toename van het aantal fouten ($d = 0,24$). Verder bleek het CI meer toe te voegen naarmate de getuige ouder is. Ook bleek het CI minder effectief naarmate het waargenomen incident meer emoties oproept. Tot slot bleek dat het CI niet leidt tot meer correcte herinneringen, maar wel tot meer verzinsels (confabulaties) naarmate er meer tijd verstrijkt tussen waarneming en interview. Dat is, gezien vanuit het oogpunt van de validiteit van de verklaring dus wel een probleem. In afwezigheid van ander bewijs waaraan de verklaring kan worden getoetst, is het onmogelijk om de validiteit van nieuwe details in een tweede interview op hun merites te beoordelen. Zijn dat daadwerkelijk correcte details, die eerder niet werden herinnerd (*reminiscence*, dat wil zeggen dat bij subsequente interviews eerder genoemde details worden ingeruild voor nieuwe details, of zelfs hypermnésie, waarbij er in een subsequent interview in absolute aantallen meer details dan initieel worden gerapporteerd; Kelley & Nairne, 2003; La Rooy et al., 2005), of zijn het incorrecte details?

Een laatste initiatief ter verdediging tegen vergeten als gevolg van retentie is het *Self Administered Interview* (SAI, Hope et al., 2011; in Nederland bekend als het Zelfformulier Getuigen; ZEG). Idee achter dit instrument is dat getuigen zo snel mogelijk hun verhaal moeten kunnen doen, zodat ze een optimaal aantal details kunnen rapporteren en hun herinnering onmiddellijk geconsolideerd wordt. Bovendien kan de politie dan optimaal gebruikmaken

van alle informatie die de getuigen aanleveren. Dat snel horen van getuigen is vooral moeilijk bij grote incidenten met meerdere getuigen. In dergelijke gevallen kan de politie een SAI uitdelen. Dat heeft de vorm van een formulier dat de getuige invult en daarmee zijn verklaring aflegt. Het SAI telt zeven instructies. A) Om te beginnen is er net als in het CI een context reinstatement ("Where were you, what were you doing, who were you with, how were you feeling, what was happening, who was involved, what could you see, and what could you hear?") en een report everything instructie. B) Daarnaast dient de getuige de dader zo gedetailleerd mogelijk te beschrijven. C) De derde instructie luidt dat de getuige een tekening maakt (plattegrond) van het incident. D) De getuige dient verder een beschrijving te geven van de andere getuigen. E) De volgende sectie in het SAI vraagt om een beschrijving van eventuele voertuigen die betrokken waren bij het incident. F) De getuige dient ook de waarnemingsomstandigheden te beschrijven (het weer, tijdstip, licht, blootstellingsduur, eventuele obstakels). G) Tot slot is er de mogelijkheid om aanvullende informatie te geven, die nog niet aan bod is gekomen in de voorgaande instructies. Er is nog weinig onderzoek naar het effect van het SAI, maar er is een toenemende interesse in het instrument vanuit de praktijk.

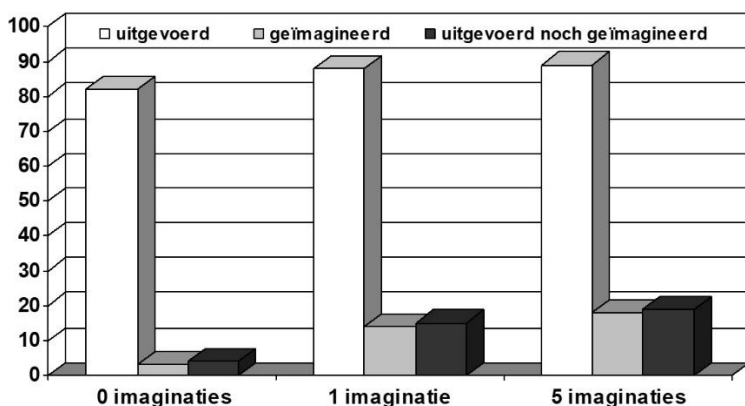
Zoals ook blijkt uit het onderzoek van Wagenaar en Groeneweg (1990) werkt tijdsverloop niet alleen vergeten in de hand, het maakt ook de weg vrij voor onbewuste, spontane veranderingen in het geheugen. Het eerder genoemde verbal labelling effect (Carmichael et al., 1932) waarbij een separate visuele en verbale herinnering spontaan samensmelten, is daar een voorbeeld van. Een ander voorbeeld van spontane reconstructies stoelt op de associatieve werking van het geheugen. Neem de woordenreeks "noot, geluid, piano, zang, radio, band, melodie, hoorn, concert, instrument, symfonie, jazz, orkest, kunst en ritme". Bij het reproduceren van deze reeks, zouden we ons zomaar het woord "muziek" kunnen

herinneren, terwijl dat er niet tussen staat. Deze pseudoherinnering kan ontstaan omdat het woord muziek thematisch goed past in de woordenreeks. Roediger en McDermott (1995) onderwierpen proefpersonen aan verschillende van dergelijke geassocieerde woordenlijsten en ontdekten dat de deelnemers gemiddeld 65% van de aangeboden woorden reproduceerden, maar ook 40% van de niet genoemde, maar wel geassocieerde woorden (de zogenoemde *critical lures*). Na deze onmiddellijke herinneringstaak, kregen de deelnemers een herkenningstaak waarin de critical lures waren opgenomen. Maar liefst 81% van de critical lures werd "herkend". Dit effect staat te boek als het *Deese-Roediger-McDermott (DRM) effect*. Monds et al. (2013) onderwierpen 109 studenten aan twaalf DRM lijsten van elk tien woorden. De helft van de lijsten was neutraal (met critical lures als school, muziek en berg) en de andere helft was meer traumatisch (met thema's als bloed, verkrachting en pijn). Deelnemers konden zich net iets meer traumatische woorden herinneren (42,72) dan neutrale (40,07). Zij rapporteerden ook meer neutrale critical lures (0,66) dan traumatische (0,39). De verschillen waren significant maar klein. Bij de herkenningstest waren er geen verschillen tussen neutrale en traumatische woordenlijsten.

Het eerder genoemde false fame effect, waarbij mensen een niet bestaande naam bij een tweede blootstelling gaan attribueren aan een bekend persoon, alleen maar omdat die naam hen nu bekend voorkomt, is ook een voorbeeld van een spontane reconstructie. Omdat dit effect onbewust en spontaan optreedt, simpelweg met het verstrijken van de tijd, noemen Jacoby et al. (1989) het ook wel een *sleeper-effect*.

Thomas en Loftus (2002) includeerden 210 studenten in hun drie sessies durende onderzoek naar imaginatie-inflatie. Tijdens sessie 1 moesten proefpersonen 18 handelingen uitvoeren en 18 handelingen slechts in gedachten uitvoeren (bijvoorbeeld een potlood slijpen).

Vierentwintig uur later vond een imaginatiesessie plaats: sommige handelingen werden niet geïmagineerd, terwijl andere één of zelfs vijf keer werden voorgesteld. Tijdens sessie 3, twee weken later, werden de proefpersonen 54 handelingen voorgelezen. Zij moesten daarbij aangeven of ze die handelingen tijdens de eerste sessie hadden uitgevoerd of niet. Het percentage bevestigende antwoorden wordt weergegeven in Figuur 8.



Figuur 8. Percentage handelingen dat geacht werd te zijn uitgevoerd (Thomas & Loftus, 2002).

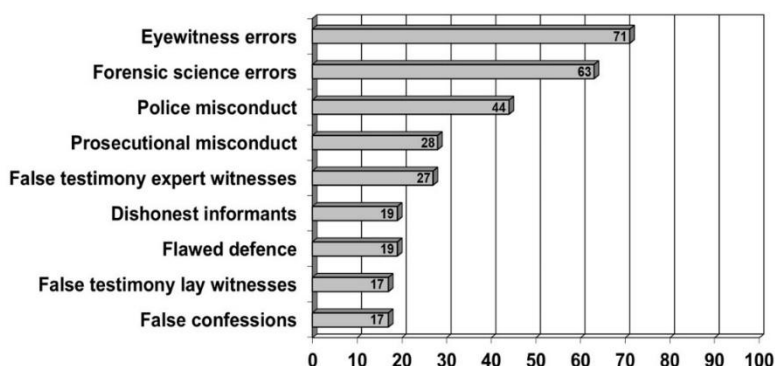
Daarbij wordt onderscheid gemaakt naar de daadwerkelijke uitvoering van de handeling tijdens sessie 1, en het aantal imaginaties tijdens sessie 2. In de figuur valt af te lezen dat daadwerkelijk uitgevoerde handelingen voor een groot deel als zodanig werden herinnerd. Een klein percentage handelingen dat slechts was voorgesteld in sessie 1 werd echter ten onrechte ook als echt uitgevoerde handeling herinnerd. Dit percentage stijgt met de toename van het aantal imaginaties. Datzelfde geldt voor handelingen die tijdens sessie 1 helemaal niet werden gepresenteerd, niet als uit te voeren handeling, noch als te imagineren handeling. Hieruit volgt dat herhaalde

imaginatie inderdaad kan leiden tot pseudo-herinneringen. Opgemerkt moet worden dat het daarbij geen verschil uitmaakte of het om alledaagse handelingen ging (zoals een handtekening zetten en een knoop in een veter leggen) of om meer bizarre handelingen (bijvoorbeeld op een dobbelsteen gaan zitten en een woord schrijven met nagellak).

Wegner et al. (1981) ontdekten dat met het verloop der tijd nuances in onze herinnering verloren gaan. Zij lieten hun proefpersonen bijvoorbeeld een verzonnen krantenkop lezen ("Bob Talbert linked with Maffia") en vroegen enige tijd later om deze Bob Talbert te beoordelen op enkele dimensies (bijvoorbeeld goed-slecht; aardig-gemeen; eerlijk-oneerlijk), op een schaal van 1 tot 7, waarbij hogere scores een negatievere evaluatie weergaven. Talbert scoorde gemiddeld een 4,25. Sommige krantenkoppen waren in vraagvorm geformuleerd ("Is Bob Talbert linked with Maffia?"), of zelfs ontkennend ("Bob Talbert not linked with Maffia"). Het viel op dat de evaluatie na het lezen van de vraagvorm net zo negatief was (4,33) als die na de stelvorm. En ook de ontkennende krantenkop leverde toch een vergelijkbare score van 3,73 op. Blijkbaar ging de nuance die dient uit te gaan van de vraagvorm of ontkenning snel verloren.

De conclusies uit het onderzoek van Wagenaar en Groeneweg (1990) en alle andere genoemde onderzoeken zijn als volgt. Om te beginnen lijken traumatische herinneringen sterk op niet-traumatische "normale" herinneringen. Ze zijn soms even duurzaam, maar ook even gevoelig voor vergeten en reconstructies. Vergeten en spontane reconstructie hebben tot gevolg dat oprechte getuigen en verdachten die meerdere malen worden verhoord, een grote kans lopen om inconsistenties in hun verklaringen te laten zien. Tegelijkertijd zouden liegende getuigen en verdachten wel eens kunnen proberen om zo consistent mogelijk uit de bus te komen, door hun verhaal goed te oefenen (men spreekt van de *repeat versus*

reconstruct hypothesis). Als dat zo is, dan zou de consistentie van verklaringen geen sterk criterium voor oprechtheid zijn, ondanks dat het wel met enige regelmaat als zodanig wordt gebruikt (Verschuere et al., 2025). Desondanks blijken vergissingen van getuigen een grote bron van rechterlijke dwalingen te zijn, bijvoorbeeld in een onderzoek naar de gebezigde bewijsmiddelen in 86 zogenoemde DNA *exoneration cases* (Saks & Koehler, 2005; zie Figuur 9).



Figuur 9. Foutenbronnen in rechterlijke dwalingen (Saks & Koehler, 2005).

CRASHING MEMORIES

Op 4 oktober 1992 stort een Boeing 747 van El Al op een flat in de Amsterdamse Bijlmer, met als gevolg meer dan 40 slachtoffers. Van deze ramp zijn er geen filmopnamen. Enige jaren later vroegen Crombag et al. (1996) aan 193 studenten en universiteits-medewerkers of ze de filmopnamen hadden gezien. Niet minder dan 107 respondenten (55%) beantwoordde deze vraag bevestigend. Op een vervolgvraag antwoordde 59% van die 107 mensen dat ze hadden gezien dat de brand onmiddellijk na impact ontstond, 23% dat de brand na enige vertraging ontstond, en 18% dat ze zich niet konden herinneren wanneer de brand precies ontstond. Uit een tweede onderzoek kwamen vergelijkbare resultaten.

Dit onderzoek dat de cynische naam *crashing memories* meekreeg, laat zien dat mensen zeer gevoelig zijn voor suggestieve vragen, dat wil zeggen vragen die informatie veronderstellen die (nog) niet door de ondervraagde zelf is geopperd. Zo bevat de vraag of je een film hebt gezien de implicatie dat er een film is. Informatie die op de een of andere manier door de vraagsteller wordt gesuggereerd, wordt ook wel posthoc informatie genoemd. Suggestibiliteit is funest in contexten waarin er objectieve informatie moet worden verkregen door middel van communicatie, denk aan getuigenverhoren en therapeutische gesprekken. Probleem is dat als posthoc informatie deel gaat uitmaken van de herinnering van de ondervraagde, de herinnering niet meer authentiek is (zelfs als de informatie correct is). Het is daarom van belang te inventariseren aan welke bronnen van posthoc informatie de getuige is blootgesteld. Posthoc informatie kan namelijk op verschillende manieren in beeld komen, bijvoorbeeld via media, of gesprekken met derden, maar ook tijdens het verhoor door de politie. Neem ter illustratie het onderstaande fragment, afkomstig uit een strafdossier (marginaal aangepast). In deze casus had de politie, na een

melding, een gewond en bewusteloos slachtoffer in een brandend huis gevonden. De politie had redenen om de aangetroffen situatie te interpreteren als het gevolg van een overval met mishandeling en brandstichting. De politie had ook al een concrete verdachte op het oog. Toen het slachtoffer enkele weken later werd verhoord, bleek er echter sprake te zijn van ernstig geheugenverlies. Het slachtoffer had geen idee wat er was gebeurd, laat staan wie de eventuele dader was. Tijdens de verschillende verhoren werd onder meer het volgende gezegd (V = vraag van verhoorder; A = antwoord van getuige).

V: "Maar dan is het de vraag, wat weet je nog."

A: "Helemaal niks en weet ook niets van de politie."

V: "Dus eigenlijk wat jou is overgekomen dat je daar lag, daar heb je geen verklaring voor uit je herinneringen."

A: "Nee, helemaal niets. Mijn eerste herinnering is van vorige week, vanaf 4 weken weet ik al niks meer."

V: "Wij hebben jou aangetroffen met letsel en het blijkt ernstig hersenletsel te zijn en dat is jou aangedaan, dat heeft iemand gedaan. Dat is voor ons duidelijk, voor de politie is het duidelijk dat iemand jou zwaar heeft mishandeld."

A: "Ja oké, dus dit heb ik daar van overgehouden?"

V: "Dit heb je daar van overgehouden. Alles. Ja je hebt allemaal plekken. En als ik jou nou zou vragen, is er iemand die dat heeft gedaan bij jou, wie zou dat dan zijn?"

A: "De eerste waar ik dan aan denk dat is R."

V: "Wat zei je?"

A: "Die sukkel."

V: "Was zei je daarvoor?"

A: "R. Ik weet zijn voornaam even niet meer, R ik weet zijn voornaam niet maar dat is de eerste waar ik op kom."

V: "Eric R."

A: "Ja"

...

A: "Ik bedenk mij nu wel iets, dat ik thuis was en dat hij er was ... , maar dat heb ik ook wel heel vaak gedroomd omdat

ik er heel vaak bang voor was. Ik weet niet of het nu echt zo is geweest of een droom is geweest."

V: "En hij is dus duidelijk Eric R?"

A: "Ja."

V: "En denk je dat zo of zie je iets?"

A: "Nou ik zie mijzelf daar nog. Ik weet het niet. Hij heeft er misschien ook wel voor gezorgd dat ik het niet goed meer weet, dat weet ik niet."

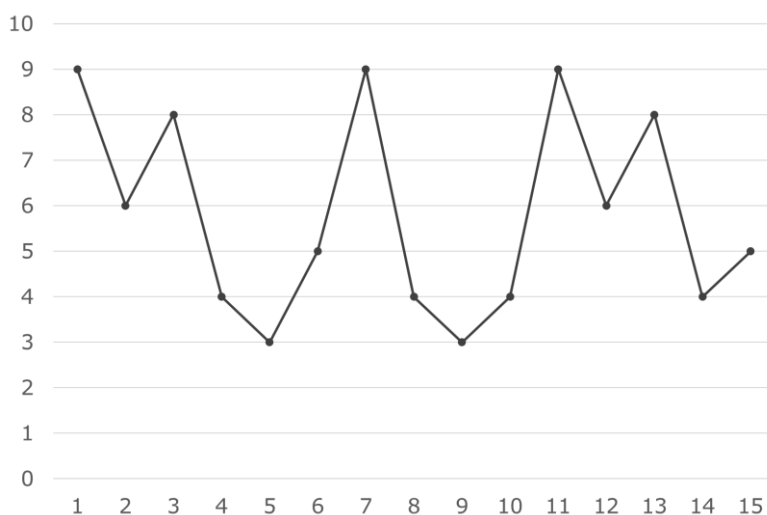
Dit fragment begint met een getuige die geen herinneringen heeft en dus ook niets kan verklaren. De politie informeert de getuige dat er sprake is van een misdrijf en vraagt vervolgens wie dat gepleegd zou kunnen hebben. De getuige noemt dan iemand, wiens volledige naam hij niet weet, maar met wie hij zich wel een conflict herinnert te hebben. De politie vertelt hem dan de volledige naam van die persoon, want dat is de verdachte. Vervolgens gaat de politie samen met de getuige werken aan het oplossen van de amnesie (*memory work*). Dit doen ze door veel te praten met de getuige over de conflicten die hij in het verleden had met de verdachte. Ook laten ze de getuige stukken lezen over de verdachte. Zo gauw de getuige in herinnering de verdachte voor zich ziet, neemt de politie aan dat het gaat om herinneringen aan echte gebeurtenissen. Hoewel dit een behoorlijk sturende aanpak is, lijkt de politie zich daar niet van bewust. In het proces-verbaal maken de verbalisanten de volgende aantekening. "Verbalisanten leggen uit er een afweging gemaakt moet worden wanneer wij getuige dingen vertellen in verband met het onderzoek en ook in verband met de betrouwbaarheid van de verklaring met betrekking tot het geheugen. Dat we dit heel zorgvuldig moeten doen en dat we getuige daarom tot nu toe nog heel weinig verteld hebben. Getuige geeft aan hier begrip voor te hebben maar dat het voor de verwerking nodig is om meer informatie te krijgen."

V: "Dan wordt het onbetrouwbaarder of tenminste, dat wordt het echt voer voor advocaten. En nou heb je helemaal

jou vrije verhaal, wat jij wist, wat jij je herinneren kon, wat jij te vertellen hebt over dingen zonder dat wij informatie ingebracht hebben.”

A: “Ja, klopt. Dat hebben jullie goed gedaan.”

Dodier en Denault (2018) beschrijven een grafische weergave van een verhoorfragment, waarmee in een oogopslag de kwaliteit van dat fragment (in termen van de afwezigheid van suggestie) kan worden gezien. In deze *Griffiths questionograph* worden negen typen vragen onderscheiden. Drie daarvan zijn kwalitatief goed: open vragen, doorvragen (wie, wat, waar, wanneer, hoe) en gesloten vragen ter afronding van een onderwerp. In de grafiek krijgen deze een score van 9, 8 en 7. Lagere scores (6-1) worden toegekend aan suggestieve vragen zoals open vragen die de getuige onderbreken, gesloten vragen, meerkeuzevragen, meningen en suggestieve vragen. In Figuur 10 staat een questionograph die een gefingeerd verhoor weergeeft dat begint met een open vraag, maar vervolgens wordt gekenmerkt door enkele “slechte” vragen.



Figuur 10. Voorbeeld van een Griffiths questionograph.

Een questionograph kan worden ingezet in de opleiding aan verhoorders, maar ook ter beoordeling van een verhoor. Overigens zijn suggestieve vragen vooral dan funest voor de validiteit van een verklaring als de getuige meegaat in de suggestie. Als blijkt dat de getuige weerstand biedt tegen de suggestie, heeft de suggestieve vraag blijkbaar geen effect gehad. Daarbij moet niet alleen worden gekeken naar de reactie die onmiddellijk op de suggestieve vraag volgt, maar ook naar alle latere reacties. Vanwege het sleeper-effect kan een met een suggestie geplant zaadje immers ook met vertraging tot een pseudoherinnering leiden, die in een later verhoor als een ogenschijnlijk authentieke verklaring wordt afgelegd.

Zoals gezegd, kan posthoc informatie niet alleen tot een getuige komen tijdens het verhoor, maar ook al daaraan voorafgaand. Tversky en Marsh (2000) toonden aan dat als we een waargenomen gebeurtenis doorvertellen in een niet-zakelijke context, bijvoorbeeld met als doel om iemand ergens van te overtuigen, of als entertainend verhaal, we accenten leggen, selectief vertellen en soms ietwat overdrijven. Als we vervolgens echter dezelfde gebeurtenis zakelijk willen navertellen (als getuige), dan zullen de accenten en selecties in het vorige gesprek ook in onze reconstructie tijdens het latere gesprek doorwerken.

Crombag (1999) beschrijft het verhaal van Edward Green. Deze kunstenaar werkte met verschillende naaktmodellen. Met sommigen van hen had hij seks. Het verhaal eindigt ermee dat vier van die modellen na enige tijd aangifte deden van seksueel misbruik door Green. Dat is opmerkelijk omdat de seksuele relaties destijds vrijwillig en probleemloos leken. De contacten waren serieel. Crombag beschrijft een proces langs welk de normale seksuele contacten achteraf alsnog tot misbruik zijn geworden in de beleving van de aangeefsters. De modellen kenden elkaar van hun studie. Zij hebben op enig moment onderling gepraat over hun bijbaantje als naaktmodel. Tijdens die gesprekken ontdekten ze dat meerdere van hen

seks hadden gehad met de kunstenaar. Door deze ontdekking voelden ze zich minder speciaal. Door vervolgens te reconstrueren dat de kunstenaar eigenlijk misbruik van hen had gemaakt, konden ze hun eigen gedrag beter verklaren en voelden ze zich minder slecht over zichzelf. Deze interpretatie zorgde voor een reductie van cognitieve dissonantie. Crombag noemt dit proces *collaborative storytelling* en benadrukt dat de aangeefsters niet logen. Ze waren echt overtuigd geraakt dat ze waren misbruikt.

Verder kan worden opgemerkt dat sommige therapeutische interventies, onbedoeld, enkele componenten bevatten die doen denken aan het uitlokken van geheugenfouten. Zo wordt er bij *Eye Movement Desensitisation and Reprocessing* (EMDR), een populaire behandeling van PTSD, van de patiënt gevraagd om een mentale voorstelling te maken van het traumatische incident om vervolgens veelvuldige saccadische oogbewegingen te maken. De bedoeling is dat het maken van die oogbewegingen tijdens het terugdenken aan het incident, de emotionele beleving van het incident drukt. Sommigen zijn echter bezorgd dat deze aanpassing van de (beleving van de) herinnering de validiteit van de herinnering aantast (Houben et al., 2020). Dat is in de therapeutische context niet zo erg, maar als deze patiënt later als getuige zou worden gehoord, is dat wel een probleem. En dat speelt dan alleen in het primaire scenario, waarin de patiënt bij aanvang van de behandeling een redelijk adequate herinnering aan het incident had. Als het alternatieve scenario in werkelijkheid correct zou zijn, en de patiënt dus helemaal geen herinnering aan het incident kan hebben, althans niet op de manier waarop de patiënt zelf denkt, dan is het maken van een voorstelling van dat incident, wetenschappelijk gezien, niets minder dan een methode om een pseudoherinnering uit te lokken (Rassin & Verschuere, 2025; Thomas & Loftus, 2002).

Opgemerkt zij dat de gevoeligheid voor suggestieve interventies niet er se samenhangt met het optreden van spontane reconstructies. Patihis et al. (2018) onderwierpen hun proefpersonen bijvoorbeeld aan twintig DRM-lijsten maar stelden hen ook suggestieve vragen over een (niet bestaand) filmpje van de *United airlines* 93 vliegtuigcrash. Hoewel zowel spontane (DRM) als gesuggereerde pseudoherinneringen optraden, bleken beide soorten niet gecorreleerd. Ook bleken persoonlijkheidskenmerken nauwelijks te voorspellen wie van de deelnemers pseudoherinneringen zou oplopen. Alleen *fantasy proneness* en dissociatie bleken enige voorspellende waarde te bezitten. Dat neemt niet weg dat in de literatuur meerdere persoonlijkheidskenmerken worden genoemd die enigszins voorspellen of mensen vatbaar zijn voor spontane of gesuggereerde geheugenfouten. Zo zouden schizofrenie en alexithymie samenhangen met spontane geheugenfouten, terwijl intelligentie, extraversie, psychopathie en openheid juist negatief daarmee samenhangen. Gevaarvermijding, nieuwsgierigheid, negatieve coping, compliance en uiteraard suggestibiliteit zouden samenhangen met gesuggereerde fouten, terwijl intelligentie en autisme daar juist negatief mee samenhangen (zie Battista et al., 2024).

Het crashing memories onderzoek was niet het enige en ook niet het eerste in het domein van suggestie en pseudoherinneringen. In een klassieke studie toonden Loftus en Palmer (1974) 45 proefpersonen een film waarin twee auto's tegen elkaar botsten. Enige tijd later vroegen ze om een schatting van de snelheid op het moment van collusie. Ze gebruikten verschillende formuleringen voor de botsing, namelijk: "How fast were the cars going when they contacted each other?", "(...) hit (...)", "(...) bumped (...)", "(...) collided (...)" en "(...) smashed (...)". De geschatte snelheid bleek te worden beïnvloed door de in de vraag gebruikte term. De respectievelijke schattingen bedroegen 31,8 mijl per uur, 34,0, 38,1, 39,3 en 40,5. De

onderzoekers gebruikten verschillende filmpjes en mede daardoor is het onbekend hoe hard de auto's daadwerkelijk reden. Niettemin bleek de schatting te zijn beïnvloed door dat ene woordje. In een tweede onderzoek kregen 150 proefpersonen een vergelijkbaar filmpje te zien. Sommigen kregen de vraag naar de snelheid met het woord hit, anderen met het woord smashed. Deze keer werd ook gevraagd of ze gebroken glas hadden gezien in het filmfragment – in werkelijkheid was daarvan geen sprake. Van de deelnemers in de hit-conditie gaf 14% een bevestigend antwoord. Dat percentage bedroeg 32% in de conditie waarin het woord smashed was gebezigd. Blijkbaar beïnvloedde de bewoording niet alleen de geschatte snelheid maar ook de visuele herinneringen aan het fragment.

Kaasa et al. (2013) keken naar het effect van dit actief zoeken naar getuigen of slachtoffers. Zij lieten 129 studenten (gemiddelde leeftijd 19 jaar) meedoen aan een onderzoek waarbij ze werden begeleid door een vrouwelijke labassistente. Een maand later werden de deelnemers gebeld met het verzoek om hun deelname aan het experiment te evalueren. De telefonist probeerde de deelnemers negatieve uitspraken over de labassistente te ontlokken. Dit gebeurde door vragen te stellen van het type "Are you sure she didn't say anything inappropriate? I know this happened a few weeks ago, so why don't you think about it for a minute?". Toen zij werden teruggebeld met het verzoek om een anoniem klachtenformulier in te vullen, was 39% daartoe bereid. Als het om een niet-anoniem klachtenformulier ging, was 15% bereid het in te vullen. De assistente had zich in werkelijkheid in alle gevallen protocollair correct gedragen. Toen de proefpersonen tijdens de *debriefing* werd gevraagd waarom ze belastend hadden verklaard over de onderzoekster, zei 39% dat te hebben gedaan om te voorkomen dat ze andere proefpersonen ongepast zou behandelen. Sommigen (12%)

hadden het gedaan omdat ze de assistente onaardig vonden.

Hoewel het crashing memories onderzoek dus niet het eerste was op het gebied van suggestie, heeft het achteraf gezien zelf een traditie in gang gezet. Er is inmiddels sprake van een heus crashing memories paradigma waarin onderzoekers een publieke gebeurtenis die niet is vastgelegd op beeld gebruiken als uitgangspunt om proefpersonen suggestief te bevragen over het niet-bestaande filmfragment. Zo is in Nederland de aanslag op Pim Fortuyn en de aanslag van tramschutter Gokman T gebruikt als publiek incident waarvan er geen filmopnamen bekend zijn. De resultaten van de verschillende onderzoeken worden samengevat in Tabel 13. Op basis van deze onderzoeken kan worden geconcludeerd dat grofweg 43% van de mensen meegaat in de suggestieve vraag en de helft daarvan (23%) denkt zich vervolgens zelfs details te herinneren.

Tabel 13. Overzicht van crashing memories onderzoeken.

Auteurs	N	Film gezien	Herinnert details
Crombag et al. (1996)	286	168	87
Ost et al. (2002)	45	20	-
Granhag et al. (2003)	182	82	-
Jelicic et al. (2006)	83	52	19
Ost et al. (2006)	48	19	-
Sjoden et al. (2009)	160	103	30
Smeets et al. (2006)	120	47	25
Wilson & French (2006)	100	36	34
Ost et al. (2008)	100	28	-
Smeets et al. (2009)	88	58	6
Patihis & Loftus (2016)	202	74	68
Patihis et al. (2018)	365	122	108
Rassin (2022)	319	98	60
Totaal	2089	907	484

Inmiddels is gebleken dat pseudoherinneringen, ook die welke in experimenten zijn uitgelokt, behoorlijk florissant en hardnekkig kunnen zijn (Arce et al., 2023;

Calado et al., 2021; McNally et al., 2004; Scoboria et al., 2017). De belangrijkste implicaties van het crashing memories en vergelijkbare onderzoeken zijn als volgt. Het is cruciaal om te achterhalen aan welke posthoc invloeden een getuige is blootgesteld. Hoe meer blootstellingen, des te minder authentiek de getuigenverklaring zal zijn. Tijdens het politieonderzoek, dient de politie zich verder te onthouden van suggestieve interventies. Tot slot zou de politie expres enkele suggestieve controlevragen aan de getuige kunnen stellen, over niet-zaakgerelateerde onderwerpen welteverstaan, om te achterhalen of de getuige de neiging heeft om zich te laten leiden door suggesties. Als dat zo is, moet de politie nog voorzichtiger tewerk gaan. Nog beter is het om een deskundige een test te laten afnemen bij de getuige, die op een gevalideerde manier meet hoe suggestibel iemand is (de Gudjonsson Suggestibility Scale; GSS; Gudjonsson, 1984, 1997).

FALSE CONFESSIONS

In 1999 publiceerden Horselenberg et al. een experimenteel onderzoek naar valse bekentenissen. In dat experiment deden 34 studenten een reactietijdentaak. Zij moesten al naar gelang de symbolen die zij op een computerscherm zagen zo snel mogelijk bepaalde toetsen op het toetsenbord indrukken. Vanwege een *bug* mochten ze echter niet de shifttoets indrukken, omdat daardoor de computer zou crashen. Als ze toch die toets indrukten, zouden ze 80% van hun beloning mislopen. Als de proefpersoon enige tijd bezig was, crashte de computer. Vervolgens werd de proefpersoon beschuldigd van het indrukken van de shifttoets. De proefleider vroeg om een schriftelijke schuldbekentenis te ondertekenen. Op enig moment kwam ook een persoon binnen die niets met het onderzoek te maken had. Die vroeg aan de proefpersoon wat er allemaal was gebeurd. In werkelijkheid was dit een *stooge*, die probeerde te achterhalen of de proefpersoon zelf geloofde de shifttoets te hebben ingedrukt. Wat de proefpersonen niet wisten is dat de software zo geprogrammeerd was dat na 96 *trials* een crash zou optreden en dat geen enkele proefpersoon daadwerkelijk de shifttoets had ingedrukt. Wat bleek? Zevenentwintig deelnemers (82%) ondertekenden de schuldbekentenis, 14 (42%) gaven achteraf aan de stooge aan zelf te geloven de shifttoets te hebben ingedrukt en 19 (58%) confabuleerden allerlei redenen waarom ze de toets hadden ingedrukt.

Dit onderzoek werpt licht op de bereidwilligheid van mensen om iets toe te geven wat ze in werkelijkheid niet hebben gedaan, zelfs als daar negatieve consequenties aanhangen. Dit soort onderzoek is tegelijkertijd moeilijk, omdat vanwege ethische restricties nooit de druk en strafdreiging kunnen worden nagebootst op een niveau dat in de buurt komt van wat verdachten tijdens een verhoor te verduren krijgen. Daarom moet er bij de extrapolatie van

deze experimentele bevindingen naar echte verhoren een grote stap worden gezet.

Dat bekentenissen vals kunnen zijn, is nauwelijks nog te ontkennen. In Amerika hadden ze de *Central Park Five*, in Groot-Brittannië de *Guildford four* en de *Birmingham Six* (zie Gudjonsson, 2003) en in Nederland hadden we onder meer de Schiedammer parkmoordzaak, de Puttense moordzaak en de Arnhemse villamoordzaak. Vraag is op welke schaal valse bekentenissen voorkomen. Dat is een moeilijk te onderzoeken vraag, omdat lang niet altijd aan het licht komt dat een bekentenis vals is. Saks en Koehler (2005) bestudeerden 86 zaken waarin aantoonbaar sprake was van een rechterlijke dwaling en ontdekten dat in 17% van de gevallen een valse bekentenis debet was aan de dwaling. Maar dat betekent natuurlijk niet dat 17% van alle bekentenissen vals is. Gudjonsson et al. (2009) pakten het anders aan. Zij ondervroegen 24.627 jongvolwassenen en ontdekten dat 2.726 van hen wel eens door de politie waren verdacht van een strafbaar feit en in verband daarmee waren verhoord. Van deze verdachten beweerden er 375 dat ze tijdens dat verhoor een (gedeeltelijke) valse bekentenis hadden afgelegd. Dit zou neerkomen op een percentage valse bekentenissen van 13,8%.

Onderzoek zoals dat van Horselenberg et al. (1999) is experimenteel en daarom wetenschappelijk sterk. Dat onderzoek was overigens niet het eerste in die traditie. Het was een replicatie van een eerdere studie van Kassin en Kiechel (1996). Zij ontwikkelden een paradigma waarmee de context van een politieverhoor vrij dicht kan worden benaderd. Vijfenzeventig studenten namen deel aan een reactietijdentaak op de computer. Ze doorliepen het onderzoek in tweetallen: een van hen las een lijst met verschillende woorden voor en de andere moest in reactie op de voorgelezen woorden zo snel mogelijk op bepaalde toetsen van een toetsenbord drukken. Voor aanvang van het onderzoek werd benadrukt dat het verboden was om de alt-toets in te drukken, omdat dat zou leiden tot het

vastlopen van de computer. Enkele minuten na aanvang van het onderzoek begaf de computer het. De proefleider kwam de kamer binnen en beschuldigde de proefpersoon van het indrukken van de alt-toets. Uiteindelijk werd de proefpersoon gevraagd om een bekentenis te ondertekenen die inhield dat hij de alt-toets had ingedrukt, tegen de instructie van de proefleider. Van de 75 deelnemers ondertekende 69% een schuldbekentenis. De reden voor het ondertekenen van de bekentenis is daarmee nog niet bekend; wellicht bekende de deelnemer omdat hij daarmee van het gezeur dacht af te zijn. Het zou echter ook zo kunnen zijn dat de deelnemer zelf was gaan geloven dat hij de alt-toets had ingedrukt. Om dat te achterhalen werden de proefpersonen naderhand opgewacht door een medewerker van de onderzoeker. Deze handlanger vroeg, zogenaamd uit interesse, hoe het onderzoek was verlopen. De reacties van de proefpersonen werden opgenomen. Uit de analyse van deze gegevens bleek dat 28% inderdaad te kennen gaf te denken de alt-toets te hebben ingedrukt; 9% ging zelfs over tot het confabuleren van allerlei details. In werkelijkheid had geen enkele proefpersoon de alt-toets ingedrukt.

Vermeldenswaardig is dat de onderzoekers de druk op een deel van de proefpersonen vergrootten door de tweede deelnemer, in werkelijkheid ook een handlanger, de valse beschuldiging van de proefleider te laten bevestigen. Daarnaast werd de cognitieve belasting gemanipuleerd door het tempo van de reactietijdentaak te variëren (43 versus 67 voorgelezen woorden per minuut). Zowel toename van sociale druk als die van cognitieve belasting deed het aantal bekentenissen, internalisering en confabulaties stijgen. In de zwaarste conditie (extra sociale druk plus extra cognitieve inspanning) was het bekenpercentage zelfs 100%. De resultaten worden samengevat in Tabel 14.

De alt-toetsstudie van Kassin en Kiechel (1996) is van belang omdat, zoals gezegd, ze de context van een

politieverhoor redelijk goed benadert. Bovendien werd in het paradigma onderscheid gemaakt tussen zogenaamde bekentenissen die *pressured compliant* en bekentenissen die *pressured internalized* zijn (zie Gudjonsson, 2003; Kassin, 1997, 2005). De eerste soort wordt afgelegd in de hoop dat daarmee een eind komt aan het verhoor. De verdachte blijft zelf overtuigd van zijn onschuld. De tweede soort is het resultaat van de gegroeide, maar onterechte overtuiging van de verdachte dat hij zich inderdaad heeft schuldig gemaakt aan iets wat verboden is. Daarnaast kan ook nog de vrijwillige valse bekentenis worden genoemd, die wordt afgelegd om de dader te beschermen, of wegens de behoefte aan aandacht of roem (zie Kassin, 1997).

Tabel 14. Percentage bekentenissen (Kassin & Kiechel, 1996).

	Geen getuige		Belastende getuige		Totaal
	Langzaam	Snel	Langzaam	Snel	
Bekentenis	35%	65%	89%	100%	69%
Internalisering	0%	12%	44%	65%	28%
Confabulatie	0%	0%	6%	35%	9%

Ondanks de kracht van de alt-toetsstudie kunnen er kanttekeningen worden geplaatst bij dit onderzoek. Zo is het indrukken van een alt-toets niet heel dramatisch en het is mogelijk ook niet een heel bewuste actie. Verder was iedereen in dit onderzoek onschuldig, want niemand had de alt-toets aangeraakt. Om dat te ondervangen – en daarmee een nog realistischere situatie te creëren – ontwikkelden Russano et al. (2005) het volgende paradigma. Een grote groep studenten ($N = 330$) nam in tweetallen deel aan een onderzoek waarin allerlei problemen moesten worden opgelost. Sommige van die problemen mochten de proefpersonen samen oplossen, terwijl andere opdrachten individueel dienden te worden voltooid. Voor de individuele opdrachten gold een samenwerkingsverbod. Terwijl het tweetal aan de slag was, werd een aantal proefpersonen

door hun collega om hulp gevraagd bij een individuele opdracht. De tweede proefpersoon was evenwel een handlanger van de onderzoeker. Na afloop van het onderzoek zei de proefleider dat beide deelnemers dezelfde fout hadden gemaakt bij een van de individuele opdrachten. Daarom was de verdenking van samenwerking gerezen en werden de twee deelnemers individueel aan de tand gevoeld daarover. Tijdens het gesprek paste de proefleider twee elementen uit de negenstappenmethode toe, te weten het minimaliseren ("I'm sure you didn't realize what a big deal it was") en het doen van een belofte in ruil voor het ondertekenen van een bekentenis ("things could probably be settled pretty quickly"; Russano et al., 2005, p. 483). Deze interventies werden niet in alle gevallen toegepast. Uiteindelijk waren er vier condities: een conditie met een verhoor zonder trucs, een waarin een belofte werd gedaan, een waarin geminimaliseerd werd, en een waarin zowel sprake was van een belofte als van minimalisatie. Vanwege de gebruikte onderzoeksmethode was er een groep schuldige verdachten ontstaan (namelijk diegenen die hadden samengewerkt) en een groep onschuldige verdachten (degenen die niet hadden samengewerkt). Het percentage bekenners wordt weergegeven in Tabel 15. Deze resultaten tonen aan dat de gebruikte verhoortrucs het aantal terechte bekentenissen inderdaad doen stijgen, maar dat het percentage valse bekentenissen eveneens en zelfs in sterkere mate stijgt. De likelihood ratio van een bekentenis neemt daarmee navenant af.

Tabel 15. Percentage deelnemers dat een terechte respectievelijk valse bekentenis aflegde, als functie van verhoorprotocol (Russano et al., 2005).

	Terechte bekentenis	Valse bekentenis	Likelihood ratio
Standaard verhoor	46%	6%	7,67
Belofte	72%	14%	5,14
Minimaliseren	81%	18%	4,50
Belofte + minimaliseren	87%	43%	2,02

Inmiddels zijn er meerdere studies verricht van dit kaliber, waarin soms net een andere vraag centraal stond. Zo is het onderzoek van Horselenberg et al. (1999) ook interessant omdat werd gezien of de persoonlijkheid van de "verdachte" voorspelt of hij zal overgaan tot het afleggen van een valse bekentenis. Het is voor psychologen altijd interessant als individuele verschillen bepaalde effecten voorspellen. Persoonlijkheidskenmerken die in dit kader interessant zijn, zijn onder meer suggestibiliteit, *compliance*, *acquiescence*, and *fantasy proneness*. Gudjonsson (2003) definieert *interrogative suggestibility* als "the extent to which, within a closed social interaction, people come to accept messages communicated during formal questioning, as the result of which their subsequent behavioural response is affected" (p. 345). De GSS is vrij complex om af te nemen. Probleem is dat suggestibiliteit ongevoelig is voor introspectie en zelfrapportage. Als we namelijk zouden weten dat we ons een pseudoherinnering hebben laten aanpraten, is dat per definitie geen pseudoherinnering meer. Bij de GSS leest de onderzoeker daarom een verhaal voor, waarna de respondent een vrije reproductie geeft. Daarna stelt de onderzoeker twintig vragen, waarvan er vijftien suggestief zijn, in de zin dat ze informatie impliceren die in werkelijkheid niet in het verhaal zit. Achteraf kan worden geteld hoe vaak de respondent in zijn antwoord meegaat met de suggestie. Daarna stelt de onderzoeker de twintig vragen nogmaals, onder vermelding dat de respondent nu beter zijn best moet doen. Als de respondent na deze aansporing zijn antwoorden aanpast, vergeleken bij de eerste ronde, wordt dat gezien als een gevoeligheid voor sociale druk. Het meegaan in de suggesties tijdens de eerste ronde en het aanbrengen van veranderingen in de tweede ronde, bepalen samen de suggestibiliteitsscore. Gudjonsson gaat ervanuit dat suggestibiliteit een belangrijke kwetsbaarheid is. Uiteraard bewijst een hoge score op de GSS niet dat de door deze getuige afgelegde verklaring onjuist is. Een lage score

bewijst evenmin dat de betreffende verklaring juist zou zijn. Naast suggestibiliteit van de getuige, spelen ook de overtuigingskracht van de communicator, de plausibiliteit van de suggestie, en de verhoorcontext een rol. In het onderzoek van Horselenberg et al. (1999) kon het afleggen van een valse bekentenis niet worden voorspeld met de persoonlijkheid van de proefpersoon. Niettemin bleek in de recentste meta-analyse dat scores op de GSS samenhangen met het afleggen van een valse bekentenis in laboratoriumonderzoeken (Otgaar et al., 2021).

Gudjonsson (1989) ontwikkelde eveneens de Gudjonsson Compliance Scale (GCS). Aangezien compliance geen geheugenprobleem is, maar een sociale neiging om anderen naar de mond te praten, is de GCS een zelfrapportage instrument.

Compliance lijkt op *acquiescence*, maar de laatste neiging betreft het bevestigend beantwoorden van vragen ("jaknikken"). Acquiescence kan worden gemeten met een lijst met conflicterende stellingen. Respondenten die conflicterende stellingen bevestigend beantwoorden worden geacht acquiescence te vertonen (Winkler et al., 1982).

Van fantasy proneness is bekend dat het samenhangt met pseudoherinneringen. Deze neiging kan worden gemeten met de Creatieve Ervaringen Vragenlijst (CEV; Merckelbach et al., 2022). Zoals eerder opgemerkt, is nog lang niet duidelijk of en hoe persoonlijkheidskenmerken samenhangen met spontane dan wel gesuggereerde pseudoherinneringen en/of het zeggen van dingen die men zelf niet gelooft (zie Battista et al. 2024).

Tabel 16 biedt een overzicht van de resultaten van onderzoeken in deze traditie. Op basis van deze experimenten is de likelihood ratio van een bekentenis (74% / 39%) dus lager dan 2. Een en ander betekent niet dat 39% van alle bekentenissen vals is. Dat aantal is namelijk tevens afhankelijk van het aantal onschuldigen dat als verdachte wordt verhoord. Als de door Kasson (2005, p.

216) veelvuldig gehoorde opmerking (“No, because I do not interrogate innocent people”) zou kloppen, is er niets aan de hand: 39% van nul is nul. Als, aan de andere kant van het continuüm, de politie consistent de verkeerde verdachten zou arresteren en verhoren, dan zouden alle afgelegde bekentenissen vals zijn. In werkelijkheid ligt de base rate van onschuldige verdachten ergens in het midden en zal dus een deel van de afgelegde bekentenissen vals zijn.

Tabel 16. Samenvatting van de resultaten van de experimenten naar (valse) bekentenissen.

Auteurs	<i>N</i>	Terechte bekentenissen	Valse bekentenissen
Kassin & Kiechel (1996)	75	-	69%
Forrest et al. (2002)	56	-	61%
Horselenberg et al. (2003)	34	-	82%
Redlich & Goodman (2003)	96	-	70%
Candel et al. (2005)	50	-	36%
Russano et al. (2005)	296	72%	20%
Horselenberg et al. (2006)	56	-	68%
	9	-	11%
	12	-	8%
Forrest et al. (2006)	98	-	82%
Billings et al. (2007)	99	-	12%
Blair (2007)	196	-	28%
Klaver et al. (2008)	219	-	40%
Nash & Wade (2009)	30	-	100%
	30	-	93%
Perillo & Kassin (2011)	71	-	61%
	38	-	60%
	59	-	58%
Narchet et al. (2011)	180	89%	29%
Horgan et al. (2012)	132	89%	32%
Jordan et al. (2012)	63	17%	4%
Pimentel et al. (2015)	173	-	48%
Frenda et al. (2016)	88	-	53%
Totaal	2169	74%	39%

Dat valse bekentenissen in wetenschappelijk onderzoek zo frequent optreden, is een probleem, omdat de politie in een spagaat raakt: Druk uitoefenen leidt

potentieel tot valse bekentenissen, maar zonder druk zal menig verdachte niet willen verklaren en wordt het moeilijk om zaken op te lossen. In veel zaken zal de politie erop zijn gebrand om de verdachte te doen bekennen. Een bekentenis is gevoelsmatig een sterk bewijsmiddel (zie Kassin, 1997, 2005), ondanks dat artikel 341, lid 4 Sv. luidt: "Het bewijs dat de verdachte het ten laste gelegde feit heeft begaan, kan door den rechter niet uitsluitend worden aangenomen op de opgaven van den verdachten". Hoewel verschillende onderzoeken wisselende statistieken hebben opgeleverd, is het veilig om te stellen dat in de meerderheid van de gevallen (ongeveer 70%) waarin uiteindelijk een veroordeling wordt uitgesproken, een bekentenis wordt afgelegd (zie Gudjonsson, 2003; Inbau et al., 2001; Russano et al., 2005). Dit onderzoek betreft dan met name de Britse en Amerikaanse situatie.

Er zijn enkele studies die aannemelijk maken dat een bekentenis veel indruk maakt. Kassin en Neumann (1997) lieten 62 psychologiestudenten verschillende twee pagina's tellende casusbeschrijvingen lezen, over een moord, zedendelict, mishandeling, of diefstal. In alle versies was er enig indirect bewijs, maar er waren ook verschillen. Zo was er soms sprake van een getuige die de dader had gezien, maar dan wel van enige afstand en bij weinig licht. Soms was er een bekende van de verdachte die hem beschreef als een agressief of misdadig type. En soms legde de verdachte een bekentenis af, zij het nadat de politie enige druk had uitgeoefend. In de controleconditie waarin alleen sprake was van indirect bewijs zou 11% van de deelnemers de verdachte veroordelen. In de varianten waarin een bekende een negatieve persoonlijkheidsschets gaf, was dat gemiddeld 21%. Als er een belastende getuigenidentificatie was, bedroeg het veroordelingspercentage 37%. In de variant met de bekentenis werd het meest veroordeeld, namelijk door 56% van de deelnemers.

In een onderzoek van Kassin en Sukel (1997) lazen vijftientig studenten een casus over ene Wilson, die werd

verdacht van het doodslaan van zijn vrouw en buurman. Wilson zou beiden hebben doodgestoken toen hij thuiskwam en hen in bed aantrof. Wilson en zijn vrouw waren al enige tijd van elkaar afgegroeid. Hij beweerde onschuldig te zijn; hij zou zijn vrouw en buurman dood hebben aangetroffen toen hij thuiskwam. Wat hem verdacht maakte, in de ogen van de officier van justitie, is dat hij eerst zijn advocaat had geraadpleegd voordat hij de politie belde. Bovendien had hij een tijdje geleden een privédetective ingehuurd om zijn vrouw te schaduwen. Het stimulusmateriaal was een 24 pagina's tellend dossier. Een deel van de proefpersonen wist alleen dat Wilson eerst zijn advocaat had gebeld en dat hij een privédetective had ingehuurd (de controlegroep). In vier andere condities werd toegevoegd dat Wilson de dubbele doodslag bekend had. In twee van die condities stond in het dossier dat Wilson onmiddellijk had bekend; er was dus geen sprake van druk tijdens het verhoor. In twee andere condities was er wel sprake van druk: Wilson was gedurende het gehele verhoor geboeid en de verhoorder had op een bepaald moment zelfs met zijn vuurwapen gezwaaid. Er was nog een complicerende variabele: sommige proefpersonen lazen dat ze bij hun taak als jurylid geen gebruik mochten maken van de bekentenis, omdat de rechter die als onrechtmatig had bestempeld. Andere deelnemers mochten de bekentenis wel gebruiken bij de beantwoording van de schuldvraag. Deelnemers dienden aan te geven of ze Wilson zouden veroordelen en in hoeverre zijn (eventuele) bekentenis een rol had gespeeld in hun overwegingen.

De onderzoekers herhaalden hun experiment met 80 andere studenten, maar dan met de aanpassing dat in alle versies iets meer bewijs zat. Zo had een buurman Wilson na het delict zien weglopen van de plaats delict met iets wat leek op een mes in zijn handen. De resultaten van beide studies staan in Tabel 17. In het kort duiden de resultaten erop dat de deelnemers die de bekentenis niet mochten meewegen weliswaar dachten dat ze dat inderdaad niet

deden, maar Wilson toch in gelijke mate veroordeelden als deelnemers die de bekentenis wel mochten gebruiken. Het lijkt er dus op dat het moeilijk is om informatie niet mee te wegen. De auteurs noemen in dezen het assimilatie-effect, waarmee ze bedoelen dat de belastende waarde van de bekentenis cognitief onbewust afstraalde op de andere bewijsmiddelen.

Tabel 17. Zelf-gerapporteerde invloed van de bekentenis (0-100%), en veroordelingspercentages als functie van verhoorprotocol (Kassin & Sukel, 1997).

	Geen bekentenis	Lage druk; (niet) toegestaan		Hoge druk	
		Wel	Niet	Wel	Niet
Studie 1					
Gewicht	-	59	24	41	12
Veroordelingspercentage	6	29	18	24	29
Studie 2					
Gewicht	-	68	50	56	19
Veroordelingspercentage	19	63	50	50	44

Afgaande op deze bevindingen heeft een bekentenis een aanzienlijke impact op de beantwoording van de schuldvraag en is het afleggen van een bekentenis, vanuit de positie van de (onschuldige) verdachte althans, sterk te ontraden. Daar staat dan echter tegenover dat het halsstarrig blijven ontkennen in aanwezigheid van een overdaad aan bewijs, als proceshouding nogal negatief kan overkomen op de rechter.

De subjectieve bewijswaarde van een bekentenis is mogelijk gelegen in de veronderstelling dat als de verdachte zelf toegeeft het delict waarvan hij wordt verdacht te hebben gepleegd, dat dan inderdaad wel zo zal zijn. Dat een verdachte iets gaat bekennen, met alle gevolgen van dien, wat hij niet heeft gedaan, is moeilijk voorstelbaar.

Anders gezegd, het idee dat een verdachte een valse bekentenis zou afleggen is tegenintuïtief. Inbau et al. (2001) zeggen daarover: "It is our contention, however, that an innocent suspect operating within normal limits of competency would not accept physical responsibility for an act he knows he did not commit" (p. 286).

Omdat een bekentenis gevoelsmatig zo sterk is, investeert de politie veel in verhoorprotocollen. De politie verhoort mensen niet zomaar als verdachte. Daarvoor moet er een redelijk vermoeden van schuld zijn. Hier tekent zich een tegenstrijdigheid af tussen de juridische fictie van de presumpctie van onschuld enerzijds en de psychologische effecten van het hebben van een vermoeden anderzijds. Het omvangrijke onderzoek naar *confirmation bias* maakt namelijk aannemelijk dat het hebben van een vermoeden van schuld, de verhoorder ongevoelig maakt voor de mogelijkheid dat hij het toch verkeerd heeft ingeschat en de verdachte onschuldig is. De psycholoog Nickerson ziet dat trouwens als een algemeen en wijdverbreid probleem en zegt het aldus: "If one were to attempt to identify a single problematic aspect of human reasoning that deserves attention above all others, the confirmation bias would have to be among the candidates for consideration" (1998, p. 175).

Uit onderzoek van Kassin et al. (2003) blijkt dat verhoorders die zonder vooraf uitgeschreven verhoorplan een verdachtenverhoor uitvoeren, offensievere vragen stellen wanneer ze sterk overtuigd zijn van de schuld van de verdachte (bijvoorbeeld: "How did you find the key that was hidden behind the VCR?"), dan wanneer hun overtuiging is verminderd doordat de proefleider hen heeft verteld dat de kans dat de verdachte schuldig is, klein is. In dat laatste geval stellen ze een meer open variant zoals "Do you know anything about the key that was hidden behind the VCR?". Uit hun onderzoek blijkt ook dat de overtuiging van schuld bij de verhoorder op een ongunstige manier interacteert met de volgehouden (terechte) onschuld van

de verdachte. Als de verhoorder merkt dat de verdachte blijft ontkennen, zal hij namelijk nog meer druk gaan uitoefenen, waardoor de verdachte steeds defensiever wordt. Het verhoor wordt gaandeweg onaangener en paradoxaal genoeg maakt de in zijn ontkening volhardende, onschuldige verdachte uiteindelijk een ongeloofwaardige en zelfs irritante indruk – ook in de ogen van onpartijdige observatoren.

In de Engels wordt het verschil tussen het neutralere getuigenverhoor en het meer vooringenomen verdachtenverhoor mooi gevat door naar het eerste te verwijzen met de term *interview*, terwijl het laatste een *interrogation* is. Daarbij zij opgemerkt dat men er bijna dogmatisch vanuit gaat dat alleen schuldige verdachten als verdachte worden verhoord. Volgens Kassin (2005) is de meest gehoorde reactie op de vraag aan een verhoorder of hij zich zorgen maakt over valse bekentenissen: “No, because I do not interrogate innocent people” (p. 216). Het spreekt tegen deze achtergrond voor zich dat het verdachtenverhoor een ander protocol volgt dan het getuigenverhoor. Een verhoormethode die in de VS veel wordt gebruikt en onder wetenschappers berucht is, is de zogenoemde negenstappenmethode van Reid. Deze methode werd voor het eerst beschreven door Reid in 1962 in het boek *Criminal Interrogation and Confessions*. Dit boek is in de loop van de jaren uitgegroeid tot het basisboek van de Amerikaanse politie. In de uitgave van de eerder genoemde Inbau et al. (2001) wordt een overzicht gegeven van Reid’s negenstappenplan. De stappen worden samengevat in Tabel 18. Merk op dat niet alle stappen in dezelfde volgorde hoeven te worden doorlopen, behalve dat het verhoor altijd zal beginnen met een confrontatie, waarbij de verdachte eigenlijk wordt voorgehouden dat het verhoor pas stopt als hij bekent. En idealiter zal het altijd eindigen met de bekentenis. De tussenliggende stappen zijn eerder te overwegen interventies dan vastliggende fasen. Bedacht moet worden dat een deel van de verdachten niet

tot medewerking bereid is bij aanvang van het verhoor. In de woorden van Inbau et al. (2001): "Indeed, it is a rare occurrence when a guilty person, after being presented with a direct confrontation of guilt says: 'Okay, you've got me; I did it.' Almost always, the suspect, whether innocent or guilty, will initially make a denial" (p. 303).

Tabel 18. Overzicht van de negenstappen verhoormethode (Inbau et al., 2001).

	Stap	Kernpunten
1.	Directe confrontatie	Verdenking uitleggen en intimideren door te spreken van overweldigend bewijs.
2.	Thema-ontwikkeling	Achterhalen waarvoor de verdachte gevoelig is, zodat hij in gesprek gaat.
3.	Omgaan met ontkenningen	Ontkenningen moeten worden omzeild bijvoorbeeld door een ander onderwerp aan te snijden.
4.	Tegenargumenten neutraliseren	Welles-nietes situatie voorkomen door tegenwerpingen om te draaien.
5.	Aandacht vasthouden	Voorkomen dat de verdachte inzakt en in zichzelf gekeerd raakt.
6.	Inactiviteit duiden	Goed "lezen" van het nonverbale gedrag om te achterhalen of de verdachte klaar is om te bekennen.
7.	Uitweg bieden	Het vergemakkelijken van bekennen door bijvoorbeeld een eervol motief voor te stellen.
8.	Mondelingen bekentenis	Zoveel mogelijk specifieke details uitvragen.
9.	Schriftelijke bekentenis	Snelle verbalisering.

Wat de tweede stap betreft: een voorbeeld van een thema dat volgens Inbau et al. (2001) bij een emotionele verdachte kan worden uitgeprobeerd, is het praten in de derde persoon. Hierbij stopt de verhoorder met het direct confronteren en gaat hij over op het spreken in algemeenheden: "Een man in jouw leeftijd, die heeft

bepaalde problemen en ik kan me voorstellen dat zo'n man oplossingen zoekt voor zijn problemen (...)." Hiermee wordt beoogd de verdachte te verleiden tot een bekentenis, omdat het besproken strafbare feit wellicht ineens een stuk verder weg lijkt te staan. Een ander thema is te stellen dat iedereen in een vergelijkbare situatie zoals die van de verdachte, zich zou hebben overgegeven aan de verleiding om het ten laste gelegde feit te plegen. Zo ook kan de ernst van het feit worden gebagatelliseerd (minimaliseren), opdat de drempel om te bekennen verlaagd wordt. De verhoorder kan sympathie tonen voor de verdachte en de schuld externaliseren: "De maatschappij kan niet verwachten dat een drugsverslaafde zoals jij stopt met inbreken als je geen alternatief hebt; hoe moet je anders aan je geld komen?" Verder kunnen de nadelen van blijven ontkennen worden overdreven (derden zullen de verdachte niet respecteren, de rechter zal een hogere straf opleggen), hetgeen ook wel maximaliseren wordt genoemd.

Als het oplaten van zulke ballonnetjes geen effect heeft, is de verdachte er waarschijnlijk een van het niet-emotionele type. Voor dergelijke verdachten zijn, aldus Inbau et al. (2001), ietwat zwaardere thema's geschikt. Op de eerste plaats kan de verhoorder trachten de verdachte te betrappen op een leugen, ongeacht hoe futiel die is. Het feit dat de verdachte één keer heeft gelogen, wordt vervolgens uitgebuit in redeneringen als "Je loog al eerder, dus waarom zouden we je nu geloven?" Het is verder van belang om de waarschijnlijkheid van zijn schuld te blijven benadrukken op basis van een toegegeven verband met het eventuele slachtoffer dan wel de plaats delict. Een niet-emotionele verdachte moet verder ervan worden overtuigd dat het bewijs tegen hem sowieso al rond is en dat bekennen alleen maar gunstig is, omdat hij daarmee karakter toont. Desnoods mag de omvang van het andere bewijs overdreven worden. Zo ook mogen verdachten tegen elkaar worden uitgespeeld, bijvoorbeeld door te zeggen dat medeplegers al bekend hebben – ook als daarvan in

werkelijkheid geen sprake is. Er staan de verhoorder kortom verschillende thema's ter beschikking. In de loop van het verhoor moet de verhoorder ontdekken welke thema's het meest aanslaan. Daartoe zal hij vriendelijke momenten moeten afwisselen met strenge momenten. Dergelijke stemmingswisselingen kunnen op zich tot gevolg hebben dat de verdachte praatgrager wordt zodra de verhoorder een vriendelijke aanpak hanteert. Overigens is het zo dat de afwisseling van vriendelijkheid en norsheid ook kan worden bereikt door twee verhoorders de afzonderlijke rollen te laten spelen. Er is dan een vriendelijke en een norse verhoorder, en de gedachte is dat de verdachte door dit contrast ertoe zal neigen in gesprek te komen met de vriendelijke verhoorder. Deze techniek wordt de *good cop, bad cop*-techniek genoemd, oftewel de *Mutt en Jeff*-techniek (Gudjonsson, 2003).

Stap drie behelst eenvoudigweg dat ontkenningen vermeden moeten worden. Dit kan door nonverbale signalen, zoals een stopgebaar wanneer de verdachte tegenstribbelt. Verbale overstemming is een andere manier om te voorkomen dat ontkenningen worden uitgesproken. Mocht het de verdachte toch lukken om te ontkennen, dan dient de verhoorder pijlsnel te reageren door te zeggen dat er ontzettend veel bewijs is waarover nog gesproken moet worden. Overigens zegt de inhoud van een ontkenning soms iets over de schuld van de verdachte, althans volgens Inbau et al. (2001). Een onschuldige verdachte zal namelijk heel kordaat blijven zeggen dat hij niets te maken heeft met het ten laste gelegde feit. Daarbij kan hij zelfs geagiteerd raken: "You're wrong, You've got to be crazy if you think I did something like that!" (Inbau et al., 2001, p. 314). Een schuldige zal echter afwijken van de regelrechte ontkenning en zal tijdtrekken: "Why do you think I did that?", "Honestly sir, I've been telling you the truth", "I knew this would happen", "You're just out to get me. I'm framed" (p. 315). Ontkenningen van schuldige verdachten zijn verder meestal ofwel bijzonder specifiek ("Dát heb ik niet gedaan"), of ze

vormen een tegenargument (marchanderen). Dergelijke tegenstribbelingen (*objections*) zijn, gesprekstechnisch beschouwd, storende zijpaden en moeten worden gepareerd door zo snel mogelijk terug te keren naar het hoofdonderwerp ("Ik begrijp wat je bedoelt, maar het feit blijft dat in deze zaak sprake is van een lijk en van 50.000 vermiste euro's; laten we het daarover hebben"). Het wegwerken van tegenwerpingen is stap vier uit het negenstappenplan.

Stap vijf draait erom de aandacht van de verdachte vast te houden. Dit kan door met verschillende verhoorders tegelijkertijd het gesprek met de verdachte aan te gaan, of door de fysieke afstand tussen verhoorder en verdachte te verkleinen. Als de aandacht van de verdachte dreigt te verslappen is het verder zaak om oogcontact te houden. De verdachte kan ook bij de les worden gehouden door allerlei hypothetische, retorische vragen te stellen ("We maken allemaal fouten, toch?", "Vind jij ook dat het zelf erkennen van fouten belangrijk is?"). Tot slot kunnen visuele stimuli (zoals foto's van de plaats delict) of (valse) bewijsmiddelen worden ingezet. Dit alles om de verdachte bij de les te houden.

Stap zes betreft het moment waarop de verdachte enigszins inzakt. Het heetst van de strijd is voorbij en de sfeer wordt rustiger. De verhoorder moet op dit cruciale moment bewerkstelligen dat de verdachte gaat bekennen. Dit kan hij doen door verschillende opmerkingen te maken. Bij sommige verdachten geldt een opmerking als "Dat je zo stil bent, duidt erop dat je een gevoelig mens bent en dat het plegen van dit feit iets is wat niet bij jou past" als laatste duwtje in de rug. Anderen hebben het juist nodig om te horen: "Wees nou maar niet zo stil, want toen je die overval pleegde was je ook niet zo verlegen." Welke aanpak het best zal werken in een bepaald geval leert de verhoorder door ervaring.

Bij de uitleg van stap zeven (het presenteren van een alternatieve vraag) gebruiken Inbau et al. (2001) de

analogie van de serveerster. Een serveerster die de gast vraagt "Kan ik u interesseren in een dessert, of bent u voldaan?", zal niet veel desserts slijten. Haar collega die de dessertopties smakelijk opnoemt en uiteindelijk vraagt "Kiest u voor de taart of het ijsje?" zal daarentegen succesvoller zijn. Op een vergelijkbare manier moet de verhoorder de verdachte uitnodigen om toe te geven dat hij het ten laste gelegde feit heeft gepleegd. Dat vereist tact en emotionele ondersteuning. De alternatieve vraag wordt bij voorkeur in de vorm van een suggestieve vraag gesteld ("It was for your family's sake, wasn't it Joe?"; Inbau et al., 2001, p. 360).

Stap acht houdt in dat een eenvoudige bekentenis moet worden aangevuld met allerlei detailverklaringen. Een uitgebreide bekentenis is namelijk overtuigender dan een eenvoudig "Ja, ik heb het gedaan". Daarom is het goed om het incident gedetailleerd door te nemen met de verdachte en te bewerkstelligen dat hij zich aan zo veel mogelijk details committeert. Het is van belang dat hetgeen wordt besproken in deze fase opgenomen wordt, of dat er tenminste een getuige bij aanwezig is, bijvoorbeeld een extra politiefunctionaris. Dit verkleint de kans dat de verdachte zijn bekentenis later terugtrekt.

Tot slot dient er een schriftelijke bekentenis te worden samengesteld ter ondertekening door de verdachte. Die bekentenis moet met zorg worden geschreven. Alle relevante aspecten van het ten laste gelegde feit moeten worden opgenomen in de bekentenis. Daarnaast moet worden vermeld dat de verdachte voorafgaand aan het verhoor de cautie heeft ontvangen, opdat later niet kan worden betoogd dat de bekentenis onder druk tot stand is gekomen. Tot slot hebben de auteurs de volgende tip: "it is a good practice to purposefully arrange for the inclusion, on each page of the confession, one or two errors, such as an incorrect name of a person or street, which will be subject to later correction by the confessor when the document is read by or to him. Any such correction should be in the

confessor's own handwriting, accompanied by his initials or signature in the margin alongside the corrections. When confronted at the trial with a confession bearing corrections of this nature, the confessor will encounter considerable difficulty in denying having read the document before signing it." (Inbau et al., 2001, p. 383-384).

Waar Inbau et al. (2001) zweren bij de kracht van de negenstappenmethode om zwijgende, onwelwillende verdachten alsnog aan de praat te krijgen, is er vanuit de wetenschap forse kritiek. Zo is het protocol eenzijdig gericht op het verkrijgen van een bekentenis en wordt de mogelijkheid dat de verdachte onschuldig is goeddeels genegeerd (Gudjonsson, 2003; Kassin, 1997, 2005). Verder propageert de methode dat de politie tegen de verdachte liegt, bijvoorbeeld door te beweren meer bewijs te hebben dan er daadwerkelijk is. Inbau et al. zelf zeggen daarover: "None of what is recommended is apt to induce an innocent person to offer a confession" (p. 313) en: "The important question to answer is whether it is human nature to accept responsibility for something we did not do in the face of contrary evidence. Upon checking a sales receipt a customer may discover that the clerk rang up the same item twice. Under this circumstance, certainly the customer would challenge the evidence (the sales receipt) rather than pay for something not purchased ... The ordinary citizen is outraged and indignant when presented with supposed "evidence" of an act he knows he did not commit" (p. 428). De auteurs gaan er dus vanuit dat hun methode niet zal leiden tot een toename van het aantal valse bekentenissen. Recentelijk weerlegden Vrij et al. (2017) echter deze en andere claims over vermeende voordelen van intense verhoormethoden zoals het negenstappenplan.

Betoogd kan worden dat de Nederlandse politie anders te werk gaat dan de Amerikaanse. Dat neemt niet weg dat ook hier duidelijk onderscheid wordt gemaakt tussen protocollen ter verhooring van getuigen versus verdachten. Waar het in de *Handleiding verhoor* (Van

Amelsvoort & Rispens, 2025) voor de Nederlandse politie gaat om dingen als verkrijgen van een spontaan verhaal, geleide herinnering, cognitief interview en forensische hypnose, voor getuigenverhoren, klinken bij het verdachtenverhoor minder vrolijke termen als persoonsgericht verhoren, bevorderen verklaringsbereidheid, spanningsniveau normaliseren en direct stapelen. In het verleden zijn ook in Nederland dubieuze verhoormethoden aan het licht gekomen. Een daarvan was de zogenoemde Zaanse verhoormethode, officieel CASE 36 genaamd (CASE staat voor *Communication Analysing System Europe*). Uit een arrest van de Hoge Raad van 22 september 1998 (NJ 1999, 104) kan worden afgeleid welke technieken deel uitmaakten van de Zaanse verhoormethode. Zo werd de verdachte in de betreffende casus drie dagen lang zeer intensief verhoord, door meer dan twee verhoorders simultaan. Ook werden hem foto's van het stoffelijk overschot van het slachtoffer getoond en werd hem voorgehouden dat hij naar huis zou mogen als hij meewerkte met de politie. Tot dusver achtte de Hoge Raad de verhoormethode toelaatbaar. De volgende elementen werden echter gewraakt: het tonen van lijkfoto's in combinatie met foto's van de kinderen van de verdachte, het ten onrechte vertellen dat een medeverdachte had bekend, het ten onrechte meedelen dat de verdachte in Turkije werd gezocht in verband met 16 kilo heroïne, het voorhouden van de redenering dat de verdachte zich zou moeten verantwoorden tegenover de maffiaorganisatie waarvan hij deel uitmaakte en het suggereren dat de raadsman van de verdachte niet diens belangen maar die van de maffia zou behartigen. Deze elementen behelsden volgens de Hoge Raad een ongeoorloofde psychische druk om de verdachte tot een bekentenis te brengen. Dat geldt ook voor het wijsmaken van de verdachte dat diens partner dreigtelefoontjes zou hebben gehad, maar dat van politiebescherming pas sprake kon zijn als de verdachte zou bekennen (zie Mul & Pols, 1999).

In een ander arrest over de Zaanse verhoormethode (13 mei 1997, *NJ* 1998, 152) ging het om de wegens doodslag tot veertien jaar gevangenis veroordeelde H.J. De verdachte had de doodslag bekend na, in februari 1995, twee dagen te zijn verhoord onder leiding van Hoenderdos, de grondlegger van de Zaanse verhoormethode. Deze communicatiedeskundige sloeg het verhoor gade vanuit een naastgelegen ruimte en voorzag de verhoorders via oortelefoontjes van tips. Tijdens het verhoor werden ook fotocollages in de kamer getoond waarop foto's van het lijk van het slachtoffer waren gekoppeld aan foto's van de dochters en vrouw van de verdachte. De raadsman vond dit veel te ver gaan en zette daarom hoog in: hij eiste dat het Openbaar Ministerie niet-ontvankelijk zou worden verklaard. Hoewel de Hoge Raad de besproken technieken in strijd achtte met een eerlijk proces, vond hij de geëiste sanctie te ver gaan. Daarom werd het cassatieberoep verworpen.

Een andere verhoormethode die ook in Nederland wordt ingezet en die kritiek vanuit de wetenschap te verduren krijgt, is de *mister Big* methode. Bij deze aanpak wordt de op vrije voeten zijnde verdachte benaderd door undercoveragenten die hem in de waan brengen dat hij in de race is om een lucratieve klus te mogen doen voor een grote gangsterbaas (*mister Big*). Om bij die baas in de smaak te vallen zou het helpen, zo maken de infiltranten de verdachte wijs, om enkele zware delicten op zijn naam te hebben. Op die manier proberen de agenten de verdachte te verleiden tot het bekennen van delicten. Vrees vanuit de wetenschap is dan dat de tegenover de infiltranten afgelegde bekentenis vals is en alleen is uitgesproken uit strategische overwegingen (stoer praten over eigen misdrijven om bij de grote baas in een positief daglicht te komen; zie Smith et al., 2009).

Het uitoefenen van enige druk op de verdachte is begrijpelijk, omdat de verhoorders overtuigd zijn van zijn schuld, terwijl niet elke verdachte onmiddellijk bekend. Dat

neemt niet weg dat het uitoefenen van druk risico's met zich meebrengt, waarvan het uitlokken van valse bekentenissen het meest bedreigend is.

Overigens zijn Inbau et al. (2001) niet onder de indruk van het alt-toets experiment van Kassin en Kiechel (1996): "However, it is a tremendous leap in logic to go from persuading someone that he accidentally pushed a computer key when he did not to persuading a criminal suspect that he intentionally killed his neighbor or sexually molested his child when he did not" (p. 426). En: "Even the naïve observer should recognize the inherent motivational differences between a laboratory subject who is unable to recall which words from a list were crossed off and an actual criminal suspect who may be facing life in prison if he acknowledges committing a crime ... Perhaps because of ethical considerations, many laboratory studies approach the issue of criminal interrogation in such a removed sense that the phenomenon being studied is so remotely related to the actual event that regardless of the statistical significance reported, one is left with the question, What does this prove?" (p. 444). Zij hebben een goed punt. Tegelijkertijd is de druk die in experimentele verhoren van vijf minuten wordt uitgeoefend op proefpersonen evenmin te vergelijken met wat de politie doet met verdachten van een misdrijf. Scherr et al. (2020) maakten een scherpe analyse van de verschillende omstandigheden in een politieel onderzoek die lijken samen te werken tegen de verdachte. Daarbij valt te denken aan tunnelvisie bij de politie met als gevolg eenzijdig onderzoek, maar ook aan de naïeve gedachte van de verdachte zelf dat "de waarheid" uiteindelijk wel aan het licht zal komen en hij zich dus (als onschuldige verdachte) geen zorgen hoeft te maken.

Valse bekentenissen met als ultieme consequentie rechterlijke dwalingen moeten worden voorkomen. Een radicale manier om valse bekentenissen te voorkomen zou zijn om te stoppen met typische verdachtenverhoren. De verhoren met verdachten gaan dan meer lijken op

getuigenverhoren, waarbij de verdachte geen “vijand” van de politie is die onder druk moet worden gezet om de waarheid te vertellen, maar een objectief te benaderen bron van informatie. Er is dan geen sprake meer van interrogation met als belangrijkste doel het lospeuteren van een bekentenis, maar van een investigative interview met als doel informatieverzameling en waarheidsvinding. De eerste bevindingen over deze softe aanpak zijn veelbelovend, niet alleen in termen van een reductie van valse bekentenissen, maar zelfs ook in een toename van verkregen informatie (Catlin et al., 2024; Meissner et al., 2014).

Een minder radicale, maar wel nog steeds preventieve interventie is het opnemen van verdachtenverhoren. Audiovisuele registraties maken niet alleen controle achteraf mogelijk, maar lijken verhoorders ook iets terughoudender te maken als het gaat om de inzet van verhoortruks (Kassin et al., 2014). Kassin (1997) waarschuwde overigens al jaren geleden voor mogelijke fouten bij het gebruik van videofragmenten. Zo bespreekt hij de *point of view* bias: een goede opname kenmerkt zich erdoor dat de gehele verhoorkamer wordt gefilmd inclusief verdachte en verhoorders, maar als alleen de verdachte in beeld wordt gebracht, voegt de opname niets toe aan een audio-opname. Daarnaast is er de *recap* bias, die inhoudt dat alleen het laatste stukje verhoor wordt opgenomen. Dat laatste stuk is relevant, omdat daarin veelal de bekentenis plaatsvindt, maar het is nu juist interessant om de aanloop naar die bekentenis te zien. Bovendien bestaat de kans dat de verdachte er onvoordelig uitziet op het laatst van het verhoor. Hij is dan vermoeid en lijkt zijn bekentenis wellicht emotioneel te doen, hetgeen hem onsympathiek doet overkomen. Tot slot is er de *framing* bias, waarmee wordt bedoeld dat ambigue fragmenten soms zowel ten voordele als ten nadele van de verdachte kunnen worden geïnterpreteerd en gebruikt.

Een preventieve werking mag ook worden verwacht van de aanwezigheid van de advocaat van de verdachte tijdens het verhoor (Verhoeven & Stevens, 2012).

Tot slot wordt in de literatuur het belang benadrukt van het in kaart brengen van kwetsbaarheden bij de verdachte. Daarbij kan worden gedacht aan de eerder genoemde kenmerken zoals suggestibiliteit (Gudjonsson, 1984) en compliance (Gudjonsson, 1989), die volgens recente inzichten samenhangen met de neiging om een valse bekentenis af te leggen (Otgaar et al., 2021). Maar er zijn ook andere kwetsbaarheden. Geijssen et al. (2018) observeerden de verhoren van 37 verdachten en ontdekten dat 69% van deze verdachten een of andere kwetsbaarheid had, zoals een verstandelijke beperking of een psychiatrische aandoening, terwijl in slechts 14% van de verhoren rekening werd gehouden met die kwetsbaarheid, bijvoorbeeld door het tempo of taalgebruik aan te passen. In ieder geval kan met Posthumus (2005) worden geconcludeerd: "In opleidingen moet aandacht worden besteed aan de theorie en praktijk van het verdachtenverhoor en aan het fenomeen van de valse bekentenis" (p. 171).

SYMPTOM VALIDITY TESTING

Merckelbach et al. (2002) gaven twintig psychologie-studenten de opdracht om, in het kader van een experiment, een bepaalde kamer binnen te gaan, daar vijf minuten goed om zich heen te kijken, en tot slot een blauwe envelop met zeven euro erin van een tafel te pakken en mee naar buiten te nemen. Eenmaal terug uit de kamer zouden zij vijftien vragen beantwoorden over hun verblijf in de kamer, maar dan op een manier alsof zij helemaal geen herinnering hadden ooit in die kamer te zijn geweest. Zij moesten amnesie simuleren. Daarna kregen zij dezelfde vijftien vragen nog een keer, maar deze keer moesten de proefpersonen proberen het correcte antwoord te geven. De vijftien vragen waren tweekeuzevragen, bijvoorbeeld "Was de envelop blauw of groen? Zat er vijf of zeven euro in de envelop? Hing er een schilderij aan de muur?".

Tijdens de tweede vragenronde, toen de deelnemers hun best deden om de vragen goed te beantwoorden, scoorden dertien van hen (65%) minstens 11 punten (van de maximale 15), en zeven mensen scoorden tussen 4 en 11 punten. Belangrijk was de prestatie tijdens de eerste ronde, toen de proefpersonen deden alsof ze amnesie hadden voor hun *mock crime*. Zes studenten (30%) scoorden binnen tussen 4 en 11. Zes (30%) hadden meer dan 11 goede antwoorden. Acht studenten (40%) scoorden lager dan 4 punten. Hierbij moet het volgende worden opgemerkt. Iemand die echt niet weet hoe de betreffende kamer eruit ziet, zal per toeval toch ongeveer de helft van de 15 vragen goed gokken. Iemand die amnesie simuleert zal dus moeten proberen om ongeveer de helft van de vragen goed te beantwoorden. Dat lukte 30%. Sommige proefpersonen overdreven hun amnesie en scoorden slechter dan op grond van toeval mag worden verwacht. Dat gold voor 40%. De 30% die meer dan 11 vragen goed beantwoordde deed het te goed voor een amnestisch

persoon. Aldus bleek de sensitiviteit van deze *symptom validity test* (SVT) 70% te zijn, omdat 70% van de deelnemers ofwel te goed (30%), ofwel te slecht (40%) presteerde voor een amnestisch persoon. Met deze toegepaste test kon dus 70% van de simulanten, overdrijvers, of zo men wil leugenaars worden betrapt.

De kritieke grenzen voor onderpresteren (< 4) en overpresteren (> 11) kunnen worden vastgesteld met een formule beschreven door Denney (1996), die als een van de eersten betoogde dat dit principe kan worden toegepast in delictsituaties. Die formule staat in Figuur 11.

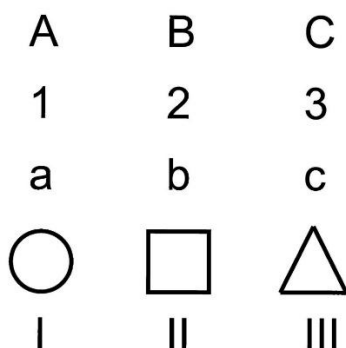
$$-Z = \frac{x - Np}{\sqrt{Np(1-p)}}$$

Figuur 11. SVT-formule (Denney, 1996).

In deze formule staat x voor het aantal goed beantwoorde vragen, N voor het totale aantal vragen en p voor de gokkans. Een score van 3 levert dan een Z -score op van $(-)2,32$ en een score van 12 levert een Z -waarde van $2,32$ op. In het algemeen ligt de 5% waarschijnlijkheid bij een Z -score van $(-)1,96$ en de 1% waarschijnlijkheidsgrens bij een Z van $(-)2,33$. Dat betekent dat iemand die echt niets weet van de kamer in het onderzoek van Merckelbach et al. (2002) bijna slechts 1% kans loopt om minder dan 4 vragen goed te gokken.

Het idee van deze *performance* SVT komt uit de neuropsychologie. In de klinische context is het van belang om het overdrijven van symptomen te detecteren. Dat kan met *tailor-made* tests zoals die in het onderzoek van Merckelbach et al. (2002), maar ook met meer gestandaardiseerde malingertests. En hoewel de eerstgenoemde soort enige tijd in de wetenschappelijke aandacht stond, lijkt hij inmiddels te zijn ingehaald door de tweede soort (Merten, 2025). Een voorbeeld van een

gestandaardiseerde test voor het overdrijven van geheugenproblemen, is Rey's 15-itemsgeheugentest (Lezak, 1983). De onderzochte krijgt een kaart te zien waarop de symbolen staan die worden weergegeven in Figuur 12. Hij krijgt deze stimulus tien seconden te zien onder de instructie dat hij alle vijftien weergegeven items moet onthouden, terwijl dat best moeilijk is. Korte tijd later moet hij de items reproduceren.



Figuur 12. Rey's 15-items test.

In werkelijkheid is dit een vrij eenvoudige test, omdat de vijftien items slechts variaties zijn van hetzelfde thema (een oplopende reeks). De gedachte achter de test is dat mensen met echte geheugenproblemen best goed presteren, terwijl simulanten het te slecht doen. De *cut-off* score is 8 of lager. De specificiteit van deze test is erg hoog: bijna niemand met geheugenklachten scoort lager dan 8. De sensitiviteit is echter teleurstellend: slechts 18% van de simulanten valt door de mand. Waarschijnlijk komt dit doordat simulanten gemakkelijk door de rationale van de test heen prikken, de test is te doorzichtig (Ponds et al., 1995). Om de sensitiviteit te vergroten, voegden Boone et al. (2002) een herkenningstest toe aan de Rey, waardoor zij een sensitiviteit van 71% bereikten, met een specificiteit van 91% (zie ook Orthey et al., 2022).

Een andere gestandaardiseerde malingertest is de *Structured Inventory of Malingered Symptomatology* (SIMS; Smith & Burger, 1997). Deze vragenlijst bestaat uit 75 items, die allemaal klachten betreffen op vijf psychologische deelgebieden: geheugen (bijvoorbeeld: "Ik heb moeite met het onthouden van mijn adres"), depressie (bijvoorbeeld: "Wanneer mijn depressie te erg wordt, ga ik lange wandelingen maken of doe ik aan gymnastiek om zo spanningen kwijt te raken"), psychose (bijvoorbeeld: "Als ik stemmen hoor, dan is het alsof mijn tanden mijn mond verlaten"), cognitieve problemen (bijvoorbeeld: "De hoofdstad van Italië is Hongarije") en neurologische klachten (bijvoorbeeld: "Ik heb gemerkt dat mijn schaduw wild beweegt, ook al houd ik me stil"). Alle beschreven klachten worden echter verwacht niet daadwerkelijk indicatief te zijn voor pathologie. Het gaat met andere woorden om symptomen die lijken op echte klachten, maar die in werkelijkheid niet of zelden voorkomen bij de betreffende aandoeningen. Respondenten die hoog scoren op deze en vergelijkbare tests, lijken hun klachten dus voor te wenden en worden simulanten of *malingers* genoemd, hoewel tegenwoordig ook wordt gesproken van het minder verwijtende "vertonen van een responsbias".

Waar de SIMS zoekt naar het overdrijven van klachten (*fake bad*), kan het ook zo zijn dat iemand zich juist beter wil voordoen dan dat hij is (*fake good*). Daar wordt soms voor gecorrigeerd met zogenoemde leugenschalen. Dat zijn items die in vragenlijsten zitten en waarop sommige antwoorden zozeer atypisch zijn dat ze mogelijk responsbias indiceren. Een voorbeeld van een *stand alone* leugenschaal is de *Social Desirability Scale* (SDS; Crowne & Marlowe, 1960). In die test staan stellingen van het type "Voordat ik bij een verkiezing mijn stem uitbreng, ga ik altijd uitgebreid de kwalificaties van alle kandidaten na", "Ik ben altijd bereid mijn fouten toe te geven", "Als ik iets niet weet, vind ik het helemaal niet erg om dat toe te geven" en "Ik heb mij nog nooit geërgerd

wanneer mensen ideeën verkondigden die erg verschilden van de mijne". Het idee is dat mensen die opvallend veel (dat wil zeggen, significant meer dan gemiddeld) van deze stellingen bevestigen, zich al te sociaal wenselijk presenteren. Die respondenten dienen dan vervolgens juist niet helemaal te worden vertrouwd. De SDS probeert de respondent als het ware te verleiden om zich onwaarschijnlijk goed voor te doen. Als de respondent zich inderdaad laat provoceren, weet de beoordelaar dat er sprake is van onoprechte, althans onwaarschijnlijke antwoorden. Ook de SDS kan daarom een provocatietest worden genoemd.

De psychometrische kwaliteiten van gestandaardiseerde SVT's is grofweg bekend, maar die van een toegepaste SVT zou per geval moeten worden geschat, hetgeen een moeilijke taak is. Dat neemt niet weg dat een toegepaste SVT zeer sterk diagnostisch kan zijn, mits de vragen goed zijn geformuleerd en de antwoordopties vergelijkbaar zijn. Tot slot kan worden opgemerkt dat symptom validity testing welbeschouwd een speciale vorm van leugendetectie is. Speciaal omdat anders dan in de meeste gevallen, de ground truth bekend is.

THE BIG-EFFECT-BIG-CAUSE HEURISTIC

Remijn en Crombag (2007) deden twee onderzoeken tegen de achtergrond van een incident in de Amsterdamse Warmoesstraat in 1997. Bij dat incident ging het om een fysieke confrontatie tussen een politieagent en een dakloze man. De laatste wilde bij herhaling het politiebureau binnentreden om onderdak en eten te eisen. Toen de politieagent de man buiten de deur probeerde te houden, kwam deze ten val en hij overleed later, waarschijnlijk als gevolg van zijn verwondingen opgelopen tijdens deze val. Vraag is of de politieagent disproportioneel geweld had gebruikt bij het weren van de dakloze man. In hun eerste onderzoek gaven Remijn en Crombag 87 proefpersonen een gefingeerd krantenartikel. Sommigen kregen een versie met als titel *Onderzoek naar de dood van een dakloze man*, verwijzend naar de daadwerkelijke afloop. Anderen kregen een versie met de titel *Klacht ingediend door dakloze man*, hintend naar een alternatieve afloop. De proefpersonen beantwoordden enkele vragen, bijvoorbeeld naar hoe hard ze dachten dat de politieagent had geduwd op een schaal van 0 tot 100, en of ze de politieagent schuldig vonden aan de dood van de dakloze man. Hoewel de geschatte kracht van de duw niet verschilde tussen de twee groepen ($M = 39,4$ en $38,2$, respectievelijk), achtten meer mensen in de eerste conditie (59%) de agent schuldig dan in de tweede conditie (29%).

In het tweede onderzoek, dat een licht aangepaste replicatie was, dachten proefpersonen in de eerste conditie wel dat de duw harder was, vergeleken bij de proefpersonen in de tweede conditie. De scores waren deze keer (op een schaal van 1 tot 6) $3,1$ versus $2,5$. Net als in het eerste onderzoek achtten meer mensen in de eerste (58%) dan in de tweede (28%) conditie de politieagent schuldig aan het overlijden van de dakloze man.

Remijn en Crombag (2007) verwijzen naar het werk van Einhorn en Hogarth (1986) over subjectieve *cues to causality: covariation, precedence, contiguity, congruency* en *lack of alternative explanations*. De resultaten passen goed bij congruentie of similariteit: als het gevolg groot is (overlijden), zal de oorzaak ook groot moeten zijn (een hardere duw en dus een schuldige politieagent). Remijn en Crombag spreken van een *major-event-major-cause* heuristiek

Wetenschappelijk gezien is het onderzoek alleen al interessant omdat de auteurs niet alleen resultaten vonden die passen bij hun hypothese. Het is bijvoorbeeld opmerkelijk dat in het eerste onderzoek wel een verschil in schuldattributie maar niet in verwachte disproportionaliteit van het geweld werd gevonden. Interessanter nog is dat het onderzoek een inherent groot probleem van juridisch beslissen aansnijdt, te weten dat in het strafrecht altijd sprake is van achterafkennis. Strafrecht is namelijk repressief: er komt pas actie als er een strafbaar feit is gepleegd. Daardoor beoordeelt de rechter het gedrag van de verdachte, met kennis van de gevolgen van dat gedrag, terwijl de verdachte (of beter dader) die kennis op het moment van het gedrag nog niet met zekerheid had. Daarom is de rechter per definitie behept met *hindsight* bias of *outcome* bias. Uit het onderzoek van Remijn en Crombag (2007) volgt dat dergelijke achterafkennis soms een onbedoeld sturende invloed kan hebben op het oordeel.

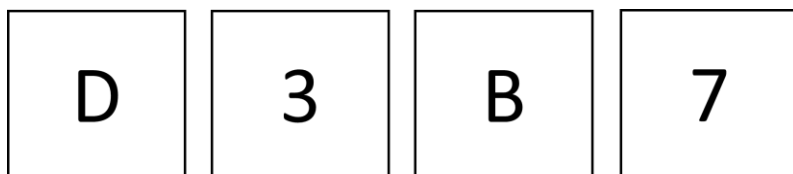
Wagenaar (1977) geeft een voorbeeld van de *hindsight* bias. Een groep proefpersonen las een kort verhaal over een Cubaanse krijgsheer, Cortés, die in 1519 ten strijde wilde trekken tegen Montezuma, de heerser van Mexico. Cortés' vloot kreeg de opdracht om Mexico binnen te vallen. De 600 zeelieden hadden daar echter geen zin in. Toen besloot Cortés om zijn eigen schepen te doen zinken, opdat zijn zeelui geen andere optie meer zouden hebben dan Mexico binnen te vallen. De deelnemers werd vervolgens gevraagd de kans te schatten op de volgende

aflopen: 1) de vloot werd tot zinken gebracht en Cortés' mannen trokken inderdaad Mexico in, 2) de soldaten ontdekten de plannen van Cortés en er ontstond muiterij en 3) de handlangers van Cortés weigerden de schepen te saboteren en daarom moest de versterking uit Cuba worden afgewacht alvorens Mexico werd aangevallen. De geschatte respectievelijke kansen bedroegen 30%, 30% en 40%. Een tweede groep proefpersonen las dezelfde casus, maar deze keer werd de afloop erbij vermeld: de schepen werden tot zinken gebracht en de soldaten vielen Mexico binnen. Als in dit geval de kans op de drie aflopen wordt geschat, zegt een meerderheid (60%) inderdaad optie 1 het waarschijnlijkst te vinden, terwijl de andere opties nog maar door 20% van de respondenten worden gekozen. Achteraf is de afloop van een gebeurtenis altijd overduidelijk en vraagt men zich af waarom de betrokken personen deze uitkomst vooraf niet konden voorzien.

Hofvind et al. (2005) beschrijven hoe achterafkennis kan leiden tot oneerlijke uitkomsten van *second opinions* in medische geschillen. Zij toonden uitkomsten van radiologisch onderzoek (mammografie) aan een aantal radiologen. In totaal waren er honderden mammografieën en de vraag was steeds of er aanwijzingen waren voor borstkanker. In een deel van de gevallen waren die aanwijzingen er wel, in een ander deel niet. Als de instructie luidde dat een te beoordelen mammografie diende te worden onderworpen aan een second opinion, kwamen de proefpersonen in 36% van de gevallen tot de conclusie dat er aanwijzingen waren voor borstkanker. Zonder deze instructie kwam echter slechts 20% van de radiologen tot de conclusie dat dergelijke aanwijzingen er waren. Blijkbaar had de informatie dat het ging om een second opinion de proefpersonen alerter gemaakt. Extra alertheid is uiteraard prima, maar volgens de onderzoekers kan dit ertoe leiden dat deskundigen die wordt gevraagd om een second opinion te geven, sneller tot een positieve uitkomst komen (daarmee de aanvankelijke arts impliciet verwijtend dat hij

de aandoening heeft gemist) dan wanneer ze op de voor hen gebruikelijk manier naar de testresultaten kijken.

Achterafkennis is ook problematisch omdat terugredeneren het moeilijker maakt om logisch correct te redeneren. De rechter die het dossier bestudeert en zich dan afvraagt of de verdachte het tenlaste gelegde feit heeft gepleegd, moet ook terugredeneren (van bewijsmiddelen naar daderschap, in plaats van andersom). In de woorden van Aben (2010, p. 226): "Het is redeneren van gevolgen naar oorzaken". Een sterk voorbeeld van een paradigma dat onze beperking in logisch redeneren blootlegt, is Wason's Kaartselectietest (Wason, 1968). In dat paradigma liggen vier kaarten op tafel. De kaarten zijn aan twee kanten bedrukt, maar uiteraard is slechts één kant zichtbaar. Op een kant staat een letter, en op de andere kant een cijfer. De kaarten worden weergegeven in Figuur 13. Vraag is welke kaarten moeten worden omgedraaid om de volgende hypothese te toetsen: "Als er op de ene kant een D staat, dan staat op de andere kant een 3".



Figuur 13. Wason's Kaartselectietest.

In deze kaartselectietest vertegenwoordigen de eerste en de laatste kaart de logische gevolgtrekkingen (modus ponens en modus tollens, respectievelijk). Als het antecedent aanwezig is, moet het consequent er ook zijn (eerste kaart) en als de consequent er niet is, mag het antecedent er ook niet zijn (laatste kaart). Interesse in de tweede en derde kaart is onlogisch en wordt aangeduid met de termen bevestiging door het consequent en ontkenning van het antecedent, respectievelijk. De kaart met de 3 hoeft dus niet te worden omgedraaid. Immers, dat achter een D

een 3 moet staan, betekent niet per se dat achter elke 3 een D moet staan. Toch is dat precies wat de reguliere manier van strafrechtelijke bewijsverzameling typeert. Het syllogisme "Een dader laat sporen achter" wordt omgedraaid naar "Er zijn sporen van de verdachte, dus dat is de dader" (Rassin, 2018).

We hebben in het algemeen moeite met correct logisch redeneren. Maar de problemen worden nog groter als de logisch correcte conclusie in strijd is met wat onze intuïtieve kennis over de werkelijkheid ons influistert. Evans et al. (1983) legden vier soorten syllogismen voor aan deelnemers. Sommige waren relatief gemakkelijk. Bijvoorbeeld: "Alle drugs zijn duur. Sommige verslavende producten zijn niet duur. Daarom zijn sommige verslavende producten geen drugs". Dit syllogisme is in zoverre gemakkelijk dat de conclusie niet alleen logisch volgt uit de premissen, maar dat het qua inhoud ook logisch, geloofwaardig lijkt. Wat ook gemakkelijk is, is een syllogisme dat onlogisch is, en tevens inhoudelijk ongeloofwaardig is. Bijvoorbeeld: "Alle drugs zijn duur. Sommige verslavende producten zijn niet duur. Daarom zijn sommige drugs niet verslavend". Het wordt moeilijk als de logische correctheid van het syllogisme afwijkt van de inhoudelijke geloofwaardigheid. Bijvoorbeeld: "Alle verslavende producten zijn duur. Sommige drugs zijn goedkoop. Daarom zijn sommige drugs niet verslavend". Hoewel de conclusie in dit geval logisch voortvloeit uit de premissen, stuit de inhoud ervan ons tegen de borst. Tot slot een syllogisme dat onlogisch, maar wel geloofwaardig is: "Alle verslavende producten zijn duur. Sommige sigaretten zijn goedkoop. Daarom zijn sommige verslavende producten geen sigaretten". Deze conclusie klopt empirisch gezien (er zijn immers andere verslavende dingen dan sigaretten), maar hij volgt niet logisch uit de premissen. Evans et al. Vroegen aan de deelnemers om de logische correctheid van de verschillende syllogismen te beoordelen. De logische en geloofwaardige syllogismen

werden door 92% van de deelnemers logisch correct geacht. De onlogische en ongeloofwaardige syllogismen waren volgens slechts 8% van de mensen correct. Maar de logische en ongeloofwaardige conclusies werden door slechts 46% van de mensen als logisch valide beoordeeld. De logisch incorrecte maar geloofwaardige conclusies waren volgens maar liefst 92% van de deelnemers toch logisch valide. Hieruit blijkt dat logisch redeneren moeilijk is, en dat we ons meer door gevoelsmatige geloofwaardigheid laten leiden dan door logica, hetgeen door Evans et al. *belief bias* werd genoemd.

Terugredeneren is, hoewel onvermijdelijk in strafrechtelijke procedures, problematisch vanwege de invloed van achterafkennis, alsook vanwege strijdigheid met logisch redeneren, en tot slot werkt het tunnelvisie in de hand. Bedacht moet worden dat de Nederlandse rechter al voor de zitting het dossier heeft doorgenomen. Als hij de verdachte voor het eerst ziet, heeft hij dus al veel informatie over hem geconsumeerd. Hoewel het dossier ook ontlastende stukken kan bevatten, is de algemene teneur belastend. Sterker nog, het hele idee van het vervolgingsonderzoek is dat de verdachte schuldig is. Het zou naïef zijn om te veronderstellen dat de strafrechter onbevooroordeeld aan een zaak begint en van daaruit alle informatie tot zich laat komen (de zogenoemde *bottom up*-aanpak). In plaats daarvan lijkt er eerder sprake te zijn van een proces waarbij de rechter de tenlastelegging als uitgangspunt neemt en vervolgens het dossier doorspint op zoek naar bewijs dat dat uitgangspunt bevestigt. Deze gang van zaken is ongunstig voor de verdachte, omdat diens schuld eigenlijk van meet af aan de werkhypothese van de rechter vormt. In plaats van een *bottom up*-aanpak, is er sprake van een *top down*-besluitvorming: het begint met de tenlastelegging en de verdere informatieverwerking is hypothesegestuurd. Dit alles brengt mee dat de rechter wel eens geneigd zou kunnen zijn om vooral aandacht te schenken aan bewijs voor de schuld van de verdachte en

minder interesse aan de dag te leggen voor ontlastende informatie. Dat wordt tenminste gesuggereerd door de wettelijke motiveringsplicht. Artikel 359 lid 3 Sv. luidt bijvoorbeeld: "De beslissing dat het feit door de verdachte is begaan, moet steunen op de inhoud van in het vonnis opgenomen bewijsmiddelen" Over expliciete aandacht voor ontlastende informatie wordt hier niet gerept. Dat gebeurt overigens wel in lid 2, zij het op een ietwat voorzichtige manier: "Het vonnis geeft, indien de beslissing afwijkt van door de verdachte dan wel door de officier van justitie uitdrukkelijk onderbouwde standpunten, in het bijzonder de redenen op die daartoe hebben geleid".

Schünemann en Bandilla (1989) onderzochten de potentiële tunnelvisie bij strafrechtelijke beslissingen. Duitse rechters ($N = 35$) werden bekendgemaakt met een casus 'in which the material of a real case was employed so as to increase external validity; the facts of the case were such that the question of guilt could not be decided definitely' (Schünemann & Bandilla, 1989, p. 185). De deelnemers kregen de informatie van de zitting via de computer. Sommige rechters konden zelf bepalen welke informatie ze opriepen en in welke volgorde. Anderen hadden geen invloed op de presentatie. Dit onderscheid werd gemaakt om het eventuele effect van activiteit versus passiviteit tijdens de zitting te onderzoeken. De belangrijkste variabele was dat een deel van de rechters voorafgaand aan de computersessie het dossier had gelezen, terwijl de overige deelnemers onbevraagd aan de computersessie begonnen. Ervan uitgaande dat het bewijs niet volledig overtuigend was, zou men verwachten dat de rechters de verdachte niet zouden veroordelen. Gevraagd naar hun beslissing bleek 100% van de rechters die het dossier hadden gelezen, aan te geven dat ze de verdachte zouden veroordelen. Er was hier geen verschil tussen actieve en passieve rechters. Van de passieve rechters zonder dossierkennis kwam 71% tot een veroordeling. De actieve én naïeve rechters deden het het best: slechts 27%

van hen gaf aan de verdachte te zullen veroordelen. De rechters in deze laatste conditie bleken tijdens een free recall-test ook de informatie die gedurende de zitting ter tafel kwam, het best te hebben onthouden. Schünemann en Bandilla (1989) concluderen "that judgmental perseverance can best be reduced if the judge lacks knowledge of the pretrial investigation and possesses an active role in investigation carried out during the trial" (p. 190). In andere woorden, de rechter zou niet moeten blijven steken in het zoeken naar voldoende bewijs om te concluderen dat de verdachte het tenlastegelegde feit heeft gepleegd.

Tot slot zij opgemerkt dat het niet gemakkelijk is om achterafkennis, en andere contextinformatie, uit te schakelen tijdens de besluitvorming. Onderzoek duidt erop dat dat moeilijk, zo niet onmogelijk is (zie Edwards & Bryan, 1997; Kassin & Sukel, 1997; Wegner et al., 1987).

ABDUCTED BY A UFO

Otgaar et al. (2008) toonden aan dat ook kinderen pseudoherinneringen kunnen oplopen, niet alleen aan alledaagse gebeurtenissen zoals zich verslikken in een snoepje, maar ook aan bizarre gebeurtenissen zoals een ontvoering door buitenaardse wezens. Zij deden dat door kinderen iets wijs te maken in de volgende trant: "Your mother told me that when you were four years old, you were abducted by a UFO. This happened when you were alone outside. Your mother was inside the house. Then she suddenly saw through the window that a UFO took you" (p. 118). Sommige kinderen kregen daarbij ook nog wat de onderzoekers noemden *prevalence information*. Zo maakten zij een krantenartikel over het verschijnen van UFO's in de periode dat de ontvoering zou hebben plaatsgevonden. Zij lieten hun suggesties los op veertig 7-8-jarigen en veertig 11-12-jarigen. In de jongste groep bleek het aanpraten van een pseudoherinnering aan een niet-bizarre gebeurtenis (bijna stikken in een snoepje) vrij succesvol, met (45%) en zonder (40%) prevalence information. De bizarre gebeurtenis werd echter vaker geaccepteerd, met (78%) dan zonder (21%) prevalence information. In de oudere groep lagen alle percentages lager. De niet bizarre herinnering werd in 23% van de gevallen zonder, en in 45% van de gevallen met prevalence information geaccepteerd. Voor de ontvoering door de buitenaardse wezens waren de respectievelijke percentages 8% (zonder) en 18% (met). In een tweede interview, een week later, bleken de pseudoherinneringen te zijn beklijfd. Gemiddeld over alle leeftijden en precieze manipulaties, geloofde 36% van de kinderen dat de gebeurtenis had plaatsgevonden.

Dit onderzoek is om meerdere redenen interessant. Om te beginnen pareert het de kritiek inhoudend dat het in dit soort onderzoek veelal gaat om triviale details waaraan pseudoherinneringen ontstaan. De ontvoering door

buitenaardse wezens is geen detail. Als kinderen pseudoherinneringen kunnen hebben daaraan, kunnen ze die ook hebben aan mishandelingen of misbruik. Door de ontvoering als onderwerp te kiezen, kunnen de onderzoekers iets zeggen over pseudoherinneringen aan "heftige" gebeurtenissen zonder over de ethische grenzen te treden. Tegelijkertijd is het interessant om de vraag te stellen in hoeverre experimentele manipulaties leiden tot langer durende effecten, aangezien de pseudoherinneringen in het tweede interview, na een week, nog aanwezig bleken te zijn.

Het onderzoek is ook interessant omdat het bijdraagt aan een reeks waarin het bijzondere karakter van het kinderlijke geheugen centraal staat. Otgaar et al. (2008) toonden aan dat het geheugen van kinderen gevoelig is voor suggestie door volwassenen, maar het geheugen van kinderen verdient om andere redenen speciale aandacht. Er zijn namelijk meerdere manieren waarop het kinderlijke geheugen afsteekt tegen dat van volwassenen. Op de eerste plaats bestaat er zoiets als infantiele amnesie, waarmee wordt bedoeld dat volwassenen zich niets of nauwelijks iets kunnen herinneren van hun eerste levensjaren. Multhaup et al. (2005) bedachten een manier om te onderzoeken over welke jaren deze amnesie zich uitstrekt. Zij legden aan een groep studenten een lijst voor met gebeurtenissen en vroegen hen om aan te geven of ze aan deze gebeurtenissen herinneringen hadden, of dat ze slechts wisten dat ze ze hadden meegemaakt (de zogenoemde *remember-know distinction*). Aangezien er algemene normen zijn voor wanneer de gepresenteerde gebeurtenissen plaatsvinden in een mensenleven (denk aan je eerste tand krijgen, gaan staan, lopen, eerste woordjes, zwemlessen, leren fietsen), konden de onderzoekers schatten vanaf welke leeftijd we herinneringen koesteren. Dat bleek grofweg vanaf viereneenhalf jaar te zijn.

Een tweede opmerkelijkheid is dat kinderen minder gevoelig zijn voor het DRM-effect. Brainerd et al. (2008) bestudeerden 17 publicaties over 26 experimenten in verschillende leeftijdscategorieën tussen de drie en begin twintig, en ontdekten dat het percentage herinnerde critical lures verdrievoudigt tussen vijf jaar en jongvolwassenheid. Valse herkenningen verdubbelen in dezelfde periode. Mogelijk doorzien kinderen de onderliggende associatie in de DRM-lijsten niet waardoor ze geen last hebben van associaties. De auteurs spreken van een paradoxale *developmental reversal*.

Op de derde plaats blijken kinderen dus, soms nog meer dan volwassenen, gevoelig voor suggestie. Dat is niet alleen aangetoond door Otgaar et al, (2008). In een inmiddels klassiek onderzoek van Leichtman en Ceci (1995) werden in totaal 176 kinderen in de leeftijd van 3 tot 6 jaar in hun klas bezocht door een onbekende man. Deze man stelde zich voor als Sam Stone, waarna hij een rondje door het klaslokaal liep. Tien weken later werden de kinderen op een correcte en neutrale manier ondervraagd over dit bezoek ("Remember the day Sam Stone visited the school? Well, I wasn't there that day, and I'd like you to tell me everything that happened when he visited"). Tot zover de controleconditie. Een deel van de kinderen zat echter in de stereotypeconditie, wat inhield dat ze voorafgaand aan het bezoek van Sam, vier keer negatieve informatie over hem hadden gekregen. Zo werd Sam beschreven als een aardige maar onhandige man die nogal eens iets laat vallen. Verder was er een suggestieconditie waarin de kinderen na Sam's bezoek vier weken lang wekelijks op een suggestieve manier werden ondervraagd. In die gesprekken werd bijvoorbeeld gesuggereerd dat Sam tijdens zijn bezoek aan de klas iets stuk of vies had gemaakt. Tot slot was er een gecombineerde stereotype- en suggestieconditie, waarin kinderen aan beide manieren van beïnvloeding werden blootgesteld. Bij de analyse van de verklaringen die de kinderen na tien weken aflegden, werd gekeken of ze

negatieve, belastende informatie over Sam gaven. Als dat zo was, was dat per definitie vals positief, omdat Sam zich in werkelijkheid keurig had gedragen in de klas. De resultaten van de analyse worden samengevat in Tabel 19. Om die tabel goed te lezen is het van belang om te weten dat in het finale interview na tien weken, na een eerste open vraag, nog verschillende soorten vragen werden gesteld. Zo was er een suggestieve vraag waarin een ballonnetje werd opgelaten ("Have you heard something about a book or teddy bear?"). als deze vraag bevestigend werd beantwoord, stelde de verhoorder een kritische vraag, namelijk of ze dat echt zelf hadden gezien. Als deze vraag ook bevestigend werd beantwoord, volgde nog een kritische opmerking ("You didn't really see him do this, did you?").

Tabel 19. Percentage valse beschuldigingen als functie van leeftijd en coaching (Leichtman & Ceci, 1995).

Conditie		Percentage kinderen dat Sam vals beschuldigde		
Vraag tijdens interview na 10 weken		3-4 jaar	5-6 jaar	Totaal
Controle	Open	0	0	0
	Ballonnetje	10	3,5	6,2
	Echt gezien?	5	0	2
	Niet echt gezien, hè?	2,5	0	1
Stereotype	Open	0	0	0
	Ballonnetje	37	19	27
	Echt gezien?	18	9	13
	Niet echt gezien, hè?	10	2,5	7,4
Suggestie	Open	21	14	17,6
	Ballonnetje	53	38	45,1
	Echt gezien?	35	10	22
	Niet echt gezien, hè?	12	9,1	9,3
Stereotype + suggestie	Open	46	30	38,5
	Ballonnetje	72	38	57,7
	Echt gezien?	44	14	31,8
	Niet echt gezien, hè?	21	5	14,4

Het valt op dat kinderen die niet werden gecoacht, in reactie op de openingsvraag in het finale interview geen spontane valse beschuldigingen tegen Sam uitten. Tegelijkertijd blijkt enige vorm van beïnvloeding en zeker

combinaties daarvan, een dramatisch effect op de prevalentie van valse beschuldigingen te hebben.

Waar in het Sam Stone-onderzoek geen data zijn verzameld over de kans op een terechte beschuldiging, gebeurde dat wel in een handvol onderzoeken van Principe et al. In die onderzoeken keken kinderen naar een goochelshow van *Magic Mumfry*. Aan het eind van de show probeerde Mumfry een konijn uit zijn hoed te toveren, maar deze truc mislukte. Later hoorde sommige kinderen twee volwassen (waaronder een leraar) praten over de mislukking. Volgens hen was het konijn ontsnapt en liep het door de school. Daarnaast was er een groep kinderen die alleen maar in dezelfde klas zaten als die die het gesprek hadden gehoord. Tot slot was er een groep die het konijn zelf zag lopen in hun klaslokaal. Een of enkele weken later werden de kinderen ondervraagd over de mislukte truc van Magic Mumfry. Maar in de tussentijd hadden zij veelvuldig de mogelijkheid gehad om met elkaar te spreken over het (vermeende) incident met het konijn. Principe et al. herhaalden dit paradigma enkele keren met steeds kleine variaties. De resultaten worden samengevat in Tabel 20.

Het is opmerkelijk hoe gevoelig de herinnering is voor sociale beïnvloeding. Zo bleek het effect niet alleen na een week op te treden, maar ook nog na vier weken. Verder bleek de verspreiding van het gerucht over het loslopend konijn snel te gaan. Als de kinderen een half uur na het overhoren van het gerucht vrij mochten spelen, verklaarde 38% van hen een week later dat ze het konijn zelf hadden gezien. Als de kinderen het eerste half uur na de show en het gerucht niet met elkaar konden praten, bleek "slechts" 11% ten prooi te vallen aan de verwarring. Zo ook bleek het gerucht funester als de kinderen elkaar kenden (41% vals positieven), dan wanneer ze elkaar niet kenden (7%). De overtuiging dat het konijn ontsnapt was bleef zelfs overeind als Magic Mumfry zelf zei dat het konijn in de hoed was blijven zitten omdat hij ziek was. Uiteraard bleek ook in deze onderzoeken dat er meer vals positieven waren

naarmate de kinderen jonger waren en naarmate de verhoorder meer suggestieve vragen stelde.

Tabel 20. Overzicht van bevindingen van Principe et al.

	Terechte vermelding van ontsnapt konijn	Spontane valse vermelding van ontsnapt konijn	Valse vermelding van ontsnapt konijn na enige vorm van suggestie
Principe et al. (2002)	78%	0%	36%
Principe et al. (2006)	100%	0%	56%
Principe et al. (2007)	-	12%	35%
Principe et al. (2010)	-	0%	22%
Principe et al. (2012)	-	0%	29%
Gemiddeld	89%	2,4%	36%

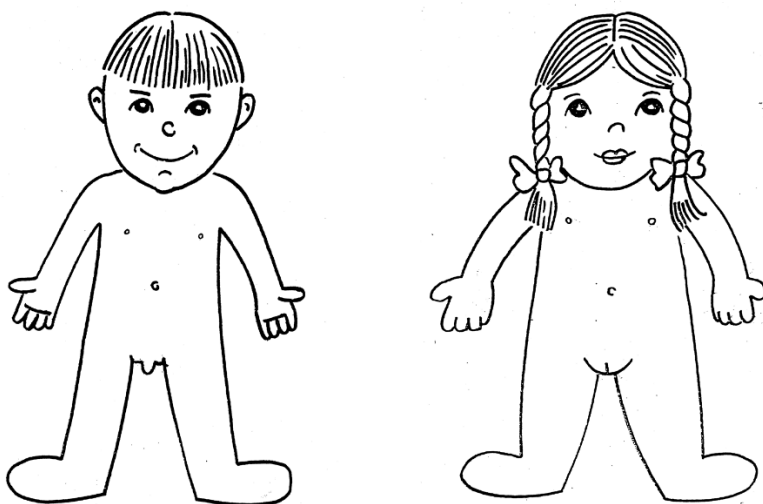
Principe et al. (2013) brachten ook in kaart dat ouders (allicht onbedoeld) een funeste invloed kunnen hebben op de betrouwbaarheid van de verklaring van hun kind. In dit onderzoek zagen kinderen weer hoe Magic Mumfry's truc mislukte, maar hoorde niemand een gerucht over de reden van de mislukking. In plaats daarvan kregen de moeders van sommige kinderen een brief met het verzoek het kind te bevragen over de truc. Andere moeders kregen een vergelijkbare brief, maar dan met de suggestie dat het konijn wel eens ontsnapt zou kunnen zijn geweest. Een week later bleek deze suggestie zijn werk te hebben gedaan. Van de kinderen wier moeder geen brief had gehad zei 5% dat het konijn ontsnapt was. Dat percentage bedroeg 8% in de groep waarin een neutrale, suggestievrije brief was gestuurd. In de conditie waarin de suggestieve brief aan de moeders was gegeven, bleek 60% van de kinderen te weten dat het konijn was ontsnapt.

Gegeven hun gevoeligheid voor suggestie is het bij het verhoren van kinderen nog belangrijker dan toch al om geen suggesties te doen. Mede daarom worden kind-verhoren waar mogelijk uitgevoerd door gespecialiseerde verhoorders in een kindvriendelijke verhoorstudio. De studio bevat een verhoorkamer die is ingericht als een soort

kinderkamer: er is allerlei speelgoed aanwezig. De door de inrichting gecreëerde huiselijke sfeer beoogt het kind zich op zijn gemak te laten voelen, opdat het vrijelijk zal durven praten. Naast de verhoorkamer bevindt zich een regiekamer van waaruit het verhoor via een onewayscreen en/of camera's kan worden gevolgd. In de regiekamer bevinden zich ten minste een regisseur en de onderzoeksleider, die de verhoorder tijdens een eventuele pauze in het verhoor van advies kunnen dienen. De regisseur draagt zorg voor de audio- en video-opname van het verhoor. Deze opname dient verschillende doelen. Ten eerste wordt de opname gebruikt om de verklaring van het kind achteraf woordelijk uit te schrijven. Verder kan de opname worden geraadpleegd in geval van onduidelijkheden, hetgeen opnieuw horen van het kind overbodig maakt. Op de derde plaats maakt de opname het mogelijk (voor de rechter of een deskundige) om de kwaliteit van het verhoor te beoordelen. En ten slotte lukt het met name jongere kinderen vaak slecht om iets verbaal te beschrijven. Zij doen het dan voor. Op enkel een geluidsband of in een proces-verbaal is dat uiteraard niet te zien, maar in een visuele registratie wel. De combinatie van speciaal opgeleide verhoorders en kindvriendelijke verhoorstudio's beoogt te bewerkstelligen dat kinderen enerzijds volledige verklaringen afleggen, die anderzijds geen – door de verhoorder gesuggereerde – onjuiste elementen bevatten. Vergeleken bij het verhoren van volwassenen dient de kindverhoorder veel tijd te investeren in de aanloop naar het daadwerkelijke verhoor. Het ijs moet worden gebroken en de verhoorder moet het kind duidelijk maken dat het alleen moet vertellen over wat het zelf heeft waargenomen (en dus niet over wat iemand hem heeft ingefluisterd). Het tonen van geduld, neutraliteit en acceptatie als het kind niet wil verklaren is een belangrijke vuistregel voor de verhoorder. In geval van een zwijgend kind liggen enkele verleidelijke interventies voor de hand, die echter gevaarlijk kunnen zijn. Op de eerste plaats is het

uitoefenen van druk uit den boze. Een opmerking als "Dat moet je je toch nog herinneren!?" kan al aanzetten tot het confabuleren van herinneringen. Ook geleide imaginatie (dat wil zeggen dat het kind wordt uitgenodigd om zich voor te stellen hoe een en ander zou kunnen zijn gebeurd) kan niet alleen de functie van drempelverlager vervullen, maar kan ook uitmonden in het ontwikkelen van pseudoherinneringen. Verder moet worden opgepast met het verhoren van kinderen in aanwezigheid van hun ouders. Het lijkt logisch dat kinderen zich in dat gezelschap vrijer voelen om iets te vertellen, maar ze zullen dan extra terughoudend zijn om iets te vertellen wat de aanwezige volwassene belast. Ook bestaat het risico dat het kind dan inderdaad netjes gaat vertellen wat de aanwezige volwassene vooraf heeft gedicteerd, terwijl in afwezigheid van die persoon nog de kans bestaat dat het kind zijn eigen herinnering zou prijsgeven. Selectieve bekrachtiging is een andere valkuil voor de verhoorder. Als een kind na enig stilzwijgen eindelijk iets belastends zegt over de verdachte, dan moet de verhoorder niet al te enthousiast reageren. Als hij dat wel doet, bestaat de kans dat het kind de indruk krijgt dat het – alleen dan – beloning krijgt (in de vorm van complimentjes) wanneer het iets belastends vertelt. Daarmee ontstaat het risico dat het kind belastende details gaat verzinnen om vooral in de gratie van de verhoorder te komen. Het verhoren van kinderen is dus moeilijk en er bestaat dan ook een speciaal leerboek voor (Dekens & Van der Sleen, 2010).

Een interventie die in het verleden mogelijk tot problemen heeft geleid, is de anatomisch correcte pop en tekening (*anatomically detailed doll/drawing*; ADD; zie Figuur 14). Dergelijke stimuli werden vooral gebruikt in gevallen waarin kinderen vermoed werden seksueel te zijn misbruikt of mishandeld. De verhoorder probeerde dan te achterhalen wat het kind had meegemaakt door te bekijken hoe het kind met de ADD speelde.



Figuur 14. Anatomisch correcte tekeningen.

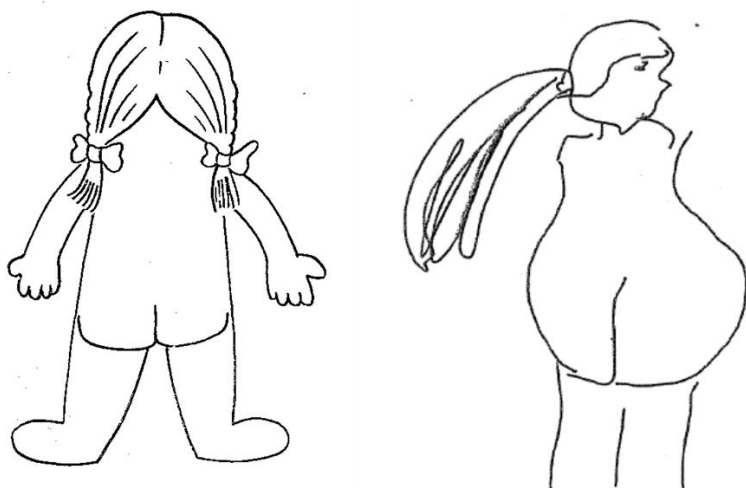
De Hoge Raad (*NJ* 1989, 748) heeft geoordeeld dat het gebruik van verklaringen die zijn ontsproten aan een poppenverhoor extra motivering behoeft. Het ging in de betreffende zaak om de 36-jarige Kees S. die was veroordeeld wegens ontucht met verschillende kinderen. Het bewijs bestond uit twee middelen. Enerzijds waren er belastende verklaringen van de kinderen. De vermoede slachtoffers (de zoon van de verdachte, twee bevriende jongens en een meisje) waren met zijn vieren onderworpen aan een gezamenlijk spelinterview met poppen. Tijdens dat interview kwamen termen boven tafel als "piemeltje likken", "anaal geslachtsverkeer" en "manueel genitaal contact". Anderzijds was er een zelfbelastende verklaring van de verdachte. Toen deze werd geconfronteerd met het feit dat zijn zoon hem beschuldigde van verkrachting, reageerde hij namelijk met: "Nu u mij vertelt dat Michel ook heeft verklaard dat ik hem zodanig heb misbruikt dat ik mijn penis in zijn anus heb gebracht, moet ik zeggen: hij kan dat zo niet verzinnen." Deze reactie was blijkbaar niet opgevat als een blijk van verschrikking, maar als een bekenenis.

Als Michel dat niet kon verzinnen, moest het dus wel waar zijn, zo lijkt de redenering te hebben geluid. De raadsman voerde een aanzienlijke hoeveelheid literatuur aan waaruit bleek dat het spelinterview mogelijk voedingsbodem was geweest voor valse beschuldigingen, maar de Rechtbank en het Hof Amsterdam verwierpen dit verweer impliciet. Volgens de Hoge Raad was dat onterecht. In casu was op zijn minst een verdere motivering van de verwerping nodig geweest.

ADD's werden na het arrest van de Hoge Raad nog wel gebruikt, maar niet meer in het kader van een spel. Kinderen konden ze wel nog gebruiken om aan te wijzen wat ze bedoelden, als ze bijvoorbeeld spraken over hun *pipi* of *poepoe*. Maar ook die toepassing is niet zonder gevaren. Willcock et al. (2006) rapporteerden een onderzoek naar de effectiviteit van anatomisch correcte tekeningen. In dat onderzoek namen ze 125 kinderen van vijf en zes jaar mee naar een brandweerkazerne. Daar werden de kinderen onder meer in een poncho gehesen, waarbij ze aan hun hoofd, schouders en oksels werden aangeraakt door een brandweerman. Een maand later werd de kinderen gevraagd om op een tekening aan te geven waar ze waren aangeraakt. Het percentage correcte aanwijzingen bedroeg 37,6%. Gemiddeld kruisten de kinderen ook 3,6 plaatsen of de tekening aan waar ze in werkelijkheid niet waren aangeraakt. Uiteindelijk was slechts 47,8% van alle aanwijzingen correct. 11,3% Van de kinderen zette een kruisje in de genitale zone (zie ook Poole et al., 2011).

Inmiddels worden ADD's door de Nederlandse politie ook niet meer ingezet als aanwijsinstrument, dit vanwege onderzoek zoals dat van Willcock et al. (2006). Enige tijd geleden zag ik een verhoorfragment waarin er onduidelijkheid was over waar op haar lichaam de kindgetuige was aangeraakt. Omdat er inmiddels geen anatomisch correcte tekeningen meer aanwezig waren in de verhoorstudio, probeerde de verhoorder (die klaarblijkelijk nog niet ervan op de hoogte was dat dat niet meer

gebruikelijk is) er zelf een te tekenen. Het resultaat daarvan staat rechts in Figuur 15. Het valt op dat de intieme zone in de zelfgemaakte versie nogal omvangrijk is, hetgeen een "toevalstreffer" in de hand werkt.



Figuur 15. ADD van achterkant meisje (links) en zelfgemaakte versie (rechts).

THE USE-THE-BEST HEURISTIC

Verschuere et al. (2023) bundelden negen onderzoeken naar onze accuratesse om leugens te detecteren. Allereerst vroegen zij aan tientallen studenten om op te schrijven wat zij op een bepaald moment hadden gedaan. Sommigen schreven de waarheid. Anderen verzonnen een alibi, omdat zij in werkelijkheid, in opdracht van de onderzoekers, een mock crime hadden gepleegd, die zij moesten verbloemen. Op die manier hadden de onderzoekers stimulusmaterialen die ze gebruikten voor hun verdere onderzoeken. Daarin werd proefpersonen gevraagd om meerdere alibi's te beoordelen op waarheid. Sommige proefpersonen pakten dat op een vrije manier aan, anderen kregen de instructie om daarbij op bepaalde kenmerken te letten, waarvan de belangrijkste de gedetailleerdheid van de verklaring was. Over de verschillende onderzoeken heen, kwamen de volgende bevindingen naar voren. Proefpersonen die ongestructureerd tewerk gingen, presteerden op gokkansniveau. Dat wil zeggen, zij konden in ongeveer 50% van de gevallen goed onderscheid maken tussen ware en gelogen alibi's. Proefpersonen die letten op de gedetailleerdheid van de narratieven deden het in 65-70% van de gevallen goed. Dat was beter dan proefpersonen die niet alleen op details letten, maar ook op andere kenmerken zoals emoties en spontaniteit. Dit zijn interessante bevindingen, omdat relatief ongetrainde beoordelaars blijkbaar een significant grotere accuratesse kunnen bereiken bij leugendetectie (70% is indrukwekkend), alleen maar door de letten op gedetailleerdheid. Tegen de achtergrond van de uitgebreide literatuur over leugendetectie en de daarin beschreven soms complexe methoden is dat een frisse boodschap.

Om de publicatie van Verschuere et al. (2023) in perspectief te plaatsen, moet allereerst worden opgemerkt dat er verschillende aanvliegroutes zijn bij het detecteren van leugens. Dubelaar (2024) spreekt van: "... de

verklaring, de bron, de presentatie en de totstandkoming” (p. 162). Het onderzoek van Verschuere et al. (2023), waarin het ging om gedetailleerdheid, valt onder de eerste categorie, de inhoud van verklaring, ook wel de verbale kenmerken genoemd. Als er wordt gekeken naar de manier waarop iemand zijn verhaal vertelt, wordt gesproken van de nonverbale kenmerken. Dubelaar noemt dat de presentatie. Denk aan oogcontact, beweeglijkheid, stotteren en dergelijke. Kortom alles wat niet de inhoud betreft, maar de manier waarop het wordt verteld. Op de derde plaats kan ook rekening worden gehouden met de persoon van wie de verklaring afkomstig is. Misschien vertrouwen we de ene persoon categorisch meer dan de andere. Tot slot is er de totstandkoming: wat is er allemaal gebeurd in de aanloop naar het te beoordelen narratief? Al deze aanvliegroutes hebben hun sterke en zwakke kanten. Geen enkele leidt tot opmerkelijk hoge accuratesse.

Om te beginnen met de laatste aanvliegroute, kan in een concreet geval de vraag worden gesteld wat de a priori kans is dat iemand zou liegen? Het lijkt erop dat deze base rate best aanzienlijk is. Uiteraard zijn er individuele verschillen, maar er wordt, afgaand op het wetenschappelijke onderzoek veel gelogen, door veel mensen. Niet iedereen liegt, en niet iedereen is goed in liegen. Het vergt bijvoorbeeld een goed ontwikkelde *theory of mind*, dat wil zeggen het vermogen om zich in te leven in andere mensen (Premack & Woodruff, 1978). Niettemin ontdekten DePaulo et al. (1996) in hun steekproef van 147 respondenten dat die gemiddeld één of twee leugens per dag vertelden, naar eigen zeggen althans. Slechts zeven respondenten beweerden nooit te liegen. Bedacht moet daarbij worden dat er verschillende soorten leugens zijn. Altruïstische leugens zijn wellicht beter te verdedigen dan hedonistische. Zo ook is het gevoelsmatig iets anders om je emoties verborgen te houden dan te verzwijgen dat je een bepaalde handeling hebt uitgevoerd.

Er zijn ook onderzoeken gedaan waarin proefpersonen in de positie werden geplaatst dat zij konden liegen voor persoonlijk gewin en dachten dat hun leugen onopgemerkt zou blijven. In dat soort onderzoek kunnen deelnemers bijvoorbeeld meer kans maken op een geldbedrag naarmate zij vaker zes ogen gooien met een dobbelsteen. Stel dat de proefpersoon vervolgens alleen wordt gelaten met de dobbelsteen om 100 keer te gooien, en dat achteraf wordt gevraagd hoe vaak hij een zes heeft gegooid. Als hij dan een hoog aantal noemt dat statistisch afwijkt van ($100 * 1/6 =$) 17, dan kan worden vermoed dat er wordt overdreven met het oog op persoonlijk gewin. In dat soort onderzoek blijkt bijna 50% van de deelnemers inderdaad te liegen (Gerlach et al., 2019). Het is dus van belang om op zijn minst de vraag te stellen of er factoren in de aanloop zijn aan te wijzen die de kans op deceptie vergroten.

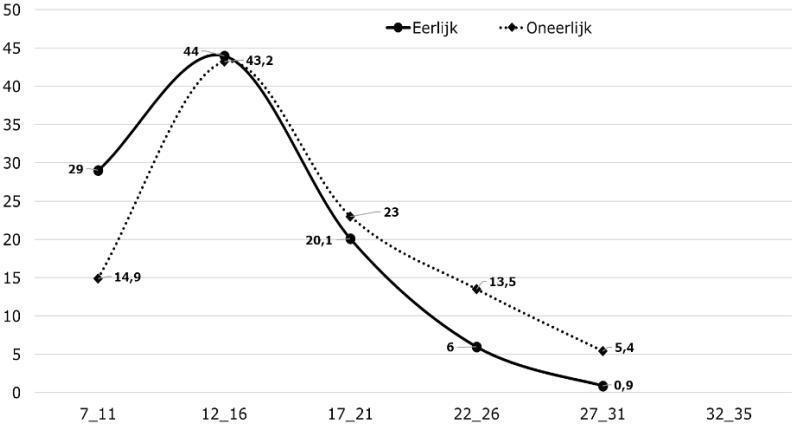
Nu is het een grote en dogmatische stap om een individu van leugens te verdenken, alleen al vanwege de grote a priori kans dat mensen in bepaalde situaties nu eenmaal liegen. Een meer op het individu afgestemde aanpak is om de persoon wiens verklaring ter discussie staat, te verdisconteren in de deceptiedetectie. Sommige mensen neigen meer naar deceptie dan andere. Rassin et al. (2024) lieten 751 proefpersonen een spelletje doen, waarin ze moesten detecteren of er in een matrix van twaalf cellen twee getallen aanwezig waren die samen precies tot tien optellen. Het spel bestond uit twintig trials. Vooraf werd gezegd de prestatie van de proefpersonen niet gecontroleerd zou worden en dat degene die de meeste combinaties vond, kans maakte op een extra geldbedrag. Figuur 16 bevat twee voorbeelden van matrices. In werkelijkheid zaten er in zeven matrices geen getallen die samen tien zijn. De maximaal haalbare score was dus dertien. In dit onderzoek werd ervan uitgegaan dat proefpersonen die meer dan dertien combinaties zeiden te hebben gevonden valsspeelden om hun kans op extra geld

te vergroten. Van de 751 proefpersonen beweerden er 74 (10%) dat ze meer dan dertien combinaties hadden gevonden. Omdat alle proefpersonen vooraf een persoonlijkheidsvragenlijst hadden ingevuld, kon worden gezien of deze 74 valsspelers verschilden van proefpersonen die hun prestatie niet overdreven.

2.38	0.72	9.41	8.64	6.34	4.93
6.51	3.59	1.23	2.93	5.68	7.05
6.36	7.62	4.93	7.68	4.52	9.12
5.05	4.04	9.18	1.67	3.64	0.59

Figuur 16. Voorbeeld van een oplosbare (links) en onoplosbare (rechts) matrix.

De valsspelers bleken hoger te scoren dan de eerlijke spelers op alle vier gemeten zogenaamde donkere trekken: psychopathie, Machiavellianisme, Narcisme en Sadisme (Paulhus et al., 2021). De procentuele psychopathiescores van de eerlijke en oneerlijke spelers wordt weergegeven in Figuur 17.



Figuur 17. Psychopathiescores van eerlijke en oneerlijke spelers.

De scoreverdelingen overlappen sterk, maar op basis van deze resultaten kan bijvoorbeeld worden gesteld dat een lage score (< 12) twee keer beter past bij eerlijke dan bij oneerlijke spelers. Ook past een hoge score (> 26) zes keer beter bij oneerlijke dan bij eerlijke spelers. Op deze manier kan de persoonlijkheid van de zender bijdragen aan de beoordeling van de geloofwaardigheid van een narratief. Ook hier is voorzichtigheid geboden. Hoewel iemands persoonlijkheid iets kan zeggen over de kans dat hij liegt, zit er een stap tussen de persoon en hetgeen die verklaart. Bovendien kan een persoon vandaag eerlijk zijn, maar morgen bedrog plegen. Als we aan onze neiging toegeven om stabiele kenmerken te attribueren, zelfs als dat op basis van wetenschappelijk onderzoek gebeurt (persoon A is eerlijk en persoon B is oneerlijk), lopen we het risico op het onderschatten van intra-individuele verschillen, wat door Vrij (2004) de fundamentele attributiefout wordt genoemd.

Naast de aanloop en de persoon is een derde aanvliegroute het analyseren van de inhoud van de verklaring. Het onderzoek van Verschuere et al. (2023) naar de gedetailleerdheid past in deze traditie. Inhoudelijke (verbale) kenmerken zijn veelvuldig onderzocht. *Criteria Based Content Analysis* (CBCA) "is probably the most widely used veracity assessment technique in the world". (Oberlader et al., 2016, p. 441). Het idee achter deze techniek is terug te voeren naar de Duitse psycholoog Undeutsch (1967), die postuleerde dat ware verklaringen een aantal structurele kenmerken bezitten die ontbreken in gelogen verklaringen. Dit postulaat staat te boek als de *Undeutsch-hypothese*. Undeutsch' idee betrof met name de verklaringen van kinderen die mogelijk slachtoffer zijn geworden van een zedendelict. In de loop van de jaren hebben Duitse psychologen de Undeutsch-hypothese uitgewerkt tot negentien criteria (Steller & Köhnken, 1989), die worden weergegeven in Tabel 21. De criteria zijn zodanig geformuleerd dat aanwezigheid ervan duidt op geloofwaardigheid. Hoewel er geen universele consensus

bestaat over de precieze scoring, kan als regel worden genoemd dat de te beoordelen verklaring nul punten per criterium krijgt toegekend in geval van afwezigheid, één punt indien het criterium enigszins aanwezig is, en twee punten indien het criterium duidelijk en in sterke mate aanwezig is. Op die manier kan een verklaring maximaal 38 punten vergaren, waarbij hogere scores duiden op sterkere geloofwaardigheid.

Tabel 21. De 19 criteria van de Criteria Based Content Analysis (CBCA).

1.	Logische structuur.
2.	Ongestructureerde, spontane reproductie.
3.	Groot aantal relevante details.
4.	Inbedding in ruimere context.
5.	Nonverbale interacties.
6.	Verbale interacties.
7.	Onverwachte complicaties.
8.	Ongewone details.
9.	Niet-relevante details.
10.	Niet-begrepen details.
11.	Relevante, maar externe gebeurtenissen.
12.	De eigen mentale toestand.
13.	De mentale toestand van de dader.
14.	Spontane correcties.
15.	Toegeven bepaalde dingen niet meer te herinneren.
16.	Twijfel aan de eigen verklaring uiten.
17.	Zichzelf beschuldigen, of neerhalen.
18.	Rechtvaardigen van de dader.
19.	Details die exclusief, specifiek zijn voor de betreffende gebeurtenis.

Bij het toepassen van de criteria staan twee veronderstellingen centraal. Op de eerste plaats dat kinderen, wegens hun beperkte mentale capaciteiten, niet in staat zijn om bepaalde elementen te verzinnen. Bij deze zogenoemde cognitieve factor moet worden bedacht dat een kind dat niet slachtoffer is geworden van een

zedendelict, maar daarover wel een (gelogen) verklaring aflegt, niet in staat zal zijn om een gedetailleerd verhaal te vertellen (criterium 3), waarin allerlei ogenschijnlijk irrelevante zijlijnen zijn opgenomen (criteria 7-11). De verklaring van een liegend kind zal dus laag scoren op dit soort criteria.

De tweede veronderstelling die in de criteria verweven zit, is dat ligende kinderen hun verklaring zo zullen voorbereiden dat deze, naar hun eigen mening, geloofwaardig overkomt. Deze motivationele factor behelst onder meer dat ligende kinderen op een vooraf bestudeerde, gladde (en dus niet-spontane) manier hun verhaal zullen vertellen (criteria 2 en 14). Verder zullen ze beducht zijn om hun eigen relaas op welke manier dan ook in twijfel te trekken – terwijl waarheidsprekers geacht worden dat wél regelmatig te doen. Een liegend kind zal daarom zelden of nooit zeggen dat het iets vergeten is, dat het betwijfelt of zijn herinnering op alle punten correct is, dat het enig begrip kan opbrengen voor de dader, of dat het zelf meer had kunnen doen om het delict te voorkomen (criteria 15-18).

Hoewel CBCA in de praktijk nogal eens als zelfstandig instrument wordt ingezet, maakt het samen met de *Validity Checklist* (VC) eigenlijk deel uit van *Statement Validity Analysis* (SVA). De VC bevat onderwerpen waar de deskundige rekening mee moet houden, maar die buiten de inhoud van de verklaring zelf liggen, te weten: de cognitieve en emotionele ontwikkeling van het kind, de gepastheid van de gebezigde taal, de gepastheid van emoties, suggestibiliteit, de kwaliteit van het verhoor (idealiter voert de psycholoog zelf een verhoor, in plaats van dat er gebruik wordt gemaakt van het proces-verbaal van het reeds afgenomen politieverhoor), mogelijke motieven om een valse verklaring af te leggen, *disclosure* omstandigheden en grote inconsistenties, mogelijk beïnvloeding door derden, realiteitsgehalte, conflicterend bewijs of andere informatie. Welbeschouwd is de VC een opdracht om alternatieve

scenario's te onderzoeken. Daarom is het jammer dat deskundigen CBCA veelal toepassen zonder de VC (Hauch et al., 2017; Ruby & Brigham, 1997; Vrij, 2005, 2016).

De diagnostische kwaliteit van CBCA is veelvuldig onderzocht en blijkt rond de 70% te liggen (Oberlader et al., 2016; Ruby & Brigham, 1997; Vrij, 2005). Dat is wetenschappelijk gezien interessant, maar bij toepassing op individuele gevallen in de praktijk is een foutenmarge van 30% best groot. Sommige auteurs pleiten dan ook tegen toepassing van CBCA (Hauch et al., 2017; Oberlader et al., 2016; Ruby & Brigham, 1997; Vrij, 2005). Volgens Horowitz (1991) zijn de foute diagnoses niet gelijkelijk verdeeld over vals-positieven en vals-negatieven: "When CBCA does err, it will tend to make false positive errors (calling false statements true)" (p. 295). Mogelijk zou het de accuratesse ten goede komen als CBCA, zoals oorspronkelijk bedoeld, wordt toegepast in combinatie met de VC. Ook zou het instrument alleen moeten worden toegepast op de verklaringen van kinderen over (seksueel) misbruik. Ook dat was oorspronkelijk de bedoeling, maar in de praktijk wordt CBCA breder ingezet. Verder is duidelijke consensus vereist over wanneer er punten worden gescoord per criterium, alsook aan hoeveel criteria (of hoeveel punten) moet zijn voldaan om een verklaring valide te achten. Het betrachten van de beperkingen van CBCA is belangrijk. Volbert en Steller (2014, p. 208) zeggen daarover: "a non-critical application of CBCA may well lead to a false result in these cases ... the central diagnostic question when performing a credibility assessment in a single case is whether this particular witness could have or would have produced this testimony on this particular alleged event if it were not based on real experience" (2014, p. 212). Bedacht moet ook worden dat CBCA uitsluitend bedoeld is ter differentiatie tussen ware en gelogen verklaringen. Maar CBCA is niet gevoelig voor verklaringen die berusten op pseudoherinneringen: "CBCA characteristics ... would seem to be less appropriate for

distinguishing between truthful reports and accounts based on false memories" (Volbert & Steller, 2014, p. 215), en: "If the fabrication hypothesis is falsified because of high content quality, this does not necessarily confirm the veracity of the testimony. As pointed out above, it is also necessary to test whether it could be based on a false memory" (p. 214), en tot slot: "To examine the hypothesis that a statement is based on a false memory, it is necessary to focus on the way in which the statement has emerged and evolved" (p. 207). Wat de accuratesse van CBCA wellicht ook ten goede zou komen, is dat de getuige, wiens verklaring het onderwerp van analyse is, gevraagd wordt om een aantoonbaar waar verhaal te vertellen, dat vervolgens als een baseline meting ter vergelijking kan worden gebruikt.

In Nederland wordt CBCA met enige regelmaat toegepast door psychologen. Er zijn zelfs ook enkele arresten van de Hoge Raad waarin eraan wordt gerefereerd. Op 30 maart 1999 (*NJ* 1999, 451) behandelde de Hoge Raad een nadere motiveringsplicht in geval van CBCA. Het ging in deze zaak om F.J. de B. die tot tien maanden gevangenisstraf was veroordeeld wegens misbruik van een minderjarige jongen. De belastende verklaring van het slachtoffer was positief beoordeeld door deskundige Bullens, die daarbij gebruik had gemaakt van CBCA. Twee contra-experts (Wagenaar en Merckelbach) hadden echter betoogd dat CBCA vaktechnisch en methodisch onbetrouwbaar is. Daarom hadden de Rechtbank en het Hof Den Haag de verklaring van Bullens niet mogen gebruiken zonder de verwerping van het verweer van de raadsman nader te motiveren.

Behalve CBCA wordt ook *Reality Monitoring* (RM) gebruikt als instrument voor verbale deceptiedetectie. RM is eigenlijk niets anders dan de eerder genoemde MCQ van Johnson et al. (1988). Dat is frappant, want de MCQ betreft de kenmerken van herinneringen, terwijl RM een instrument is om leugens en waarheden te onderscheiden.

De 39 MCQ-criteria worden veelal geclusterd tot zes categorieën, te weten: 1) contextinformatie (bijvoorbeeld ruimte en tijd), 2) sensorische informatie (vormen, kleuren, geluiden), 3) semantische informatie (verbale interacties), 4) emotionele informatie, 5) helderheid, levendigheid en 6) cognitieve operaties (reconstructies en invullingen). Voor de eerste vijf categorieën geldt dat hun aanwezigheid wordt geacht te duiden op geloofwaardigheid, terwijl dat voor de laatste categorie juist andersom is. Anders dan bij CBCA is het toepassingsdomein van RM onbeperkt. Er is geen speciale training voor vereist. Ondertussen doet de accuratesse van RM met 70% niet onder voor die van CBCA (Masip et al., 2005).

De afgelopen jaren zijn er enkele nieuwe inzichten opgedaan die vooral betrekking hebben op gedetailleerdheid als criterium voor geloofwaardigheid. Zo kan het best gebeuren dat leugenaars een gedetailleerd narratief presenteren, maar omdat zij een *keep it simple* (in plaats van een *tell it all*) strategie zullen aanhangen, zullen zij vooral verifieerbare details vermijden (dat wil zeggen, details die in theorie zijn te controleren). Vandaar dat men spreekt van een *verifiability approach*. Ook zullen zij complicaties (in de zin van details die niet per se noodzakelijk zijn om te vertellen, en die het vertellen van het verhaal bemoeilijken) vermijden. *Self-handicapping* opmerkingen en details die eigenlijk slechts algemene kennis reflecteren, zullen leugenaars wel graag opnemen in hun narratief. Inmiddels zijn er enkele onderzoeken gedaan naar verifiability en complicaties, en blijken deze criteria inderdaad te kunnen helpen bij het discrimineren tussen ware en gelogen narratieven (Nahari & Vrij, 2014; Nahari et al., 2014). Sommigen propageren zelf het gebruik van een formule waarin de proportie van echte complicaties wordt berekend tussen de algemene kennis details en self-handicapping details (Nahari et al., 2019). Al met al veegt het onderzoek van Verschuere et al. (2023) de vloer aan met deze complexe nuances.

De laatste aanvliegroute in de beoordeling van de geloofwaardigheid van narratieven is een analyse van de presentatie, waarin de nonverbale kenmerken worden betrokken. Daarvoor is een persoonlijke ontmoeting of het een audiovisuele registratie van een verhoor nodig. Daarbij moet worden bedacht dat een analyse van nonverbale communicatie bijna per definitie bovenop een analyse van de inhoud komt, tenzij de beoordelaar zich op de een of andere manier afsluit van het geluid. In eerste instantie lijkt dat goed, omdat het aantal mogelijke indicatoren daarmee wordt vergroot. Bij nader inzien is het echter de vraag of een audiovisuele analyse superieure resultaten oplevert ten opzichte van een verbale analyse. Wiseman (1995) voerde een elegant onderzoek uit waaruit blijkt dat meer niet altijd beter is. Hij liet een bekende Britse politicus voor de camera twee keer vertellen over zijn favoriete film: één keer sprak hij de waarheid en de ander keer loog hij; hij had het toen over een film die hij helemaal niet leuk vond. In sommige delen van Engeland liet Wiseman de twee fragmenten uitzenden op de lokale tv. In een ander deel van het land werden de fragmenten uitgezonden via de radio. In weer een ander deel werden de uitgetypte verhalen in de krant afgedrukt. Op deze manier kon Wiseman grote groepen respondenten rekruteren (in totaal meer dan 41.000). Bovendien wist hij vrij zeker dat de krantenlezers alleen verbale informatie kregen, radioluisteraars verbale en vocale, en tv-kijkers alle potentiële aanwijzingen (verbaal en nonverbaal). Het bleek dat 64% van de krantenlezers het goed hadden, 73% van de radioluisteraars, en slechts 52% van de tv-kijkers. Blijkbaar was het in dit geval onvoordelig om zoveel mogelijk indicatoren te gebruiken. Vooral de visuele indicatoren drukten de accuratesse.

In het onderzoek van Wiseman (1995) deden geïnteresseerde leken mee als beoordelaars. Ekman en O'Sullivan (1991) en Ekman et al. (1999) vroegen aan verschillende professionele steekproeven om meerdere filmpjes te beoordelen. Ook deze speciale steekproeven

bleken nauwelijks het niveau van gokkans (50% correct) te kunnen overstijgen (zie Tabel 22). Bond en DePaulo (2006) schreven een overzichtsartikel over 206 onderzoeken waarin in totaal 24.483 proefpersonen meermaals hadden getracht leugens en waarheden te onderscheiden. Van de waarheden bleek 61% te worden herkend, van de leugens 47%, en de gemiddelde accuratesse was 54%. Gongola et al. (2017) herhaalden het onderzoek, maar dan met betrekking tot het differentiëren tussen leugens en waarheden in de verhalen van kinderen. Zij includeerden 45 experimenten met 7.892 proefpersonen die de verhalen van 1.858 kinderen beoordeelden, in hun meta-analyse en vonden een gemiddelde detectie van 64% van de waarheden, 48% van de leugens, en een gemiddelde accuratesse van 54%.

Tabel 22. Deceptiedetectie accuratesse in verschillende steekproeven.

	Ekman & O'Sullivan (1991)	Ekman et al. (1999)
Secret service	64,1%	73%
Polygraphers	55,7%	-
Politieagenten	55,8%	50,8%
Rechters	56,7%	62%
Psychiaters	57,6%	-
Studenten	52,8%	-
Sheriffs	-	66,7%
Klinisch psychologen	-	62,1%

Aamodt en Custer (2006) keken naar 108 onderzoeken met in totaal 16.537 proefpersonen en concludeerden dat leeftijd, ervaring, opleiding, geslacht, noch beroep zijn geassocieerd met accuratesse in het discrimineren tussen leugen en waarheid. Bond en DePaulo (2008) concludeerden: "When judging deception, people differ less in ability than in the inclination to regard others' statements as truthful ... Moreover, some people are more credible than others whether lying or truth telling ... the outcome of a deception judgment depends more on the

liar's credibility than any other individual difference" (p. 477).

De vraag dringt zich op waarom mensen niet goed zijn in het doorzien van leugens. Op die vraag zijn meerdere antwoorden mogelijk. Allereerst is er geen duidelijk, sterk nonverbaal kenmerk dat leugenaars verraaft. De beschikbare indicatoren treden immers niet altijd op. Een leugenaar die niet zenuwachtig is, zal zich waarschijnlijk niet zenuwachtig gedragen. En dan nog geldt dat mensen verschillend reageren als ze zenuwachtig zijn. Daar komt bij dat indicatoren veelal klein in amplitude zijn. Als Pinocchio's neus groeit tijdens liegen, groeit hij geen centimeter maar hooguit een millimeter. Als ze al te zien zijn, zijn de nonverbale indicatoren uiterst moeilijk waarneembaar. Paul Ekman spreekt van "micro-expressies". Een volgende reden voor de tegenvallende prestaties in onderzoeken naar bedrogdetectie is dat mensen ertoe neigen op de verkeerde indicatoren te letten. DePaulo et al. (2003) verzamelden 116 studies die de afgelopen decennia werden verricht naar nonverbale kenmerken van bedrog. Uit hun meta-analyse kwamen verschillende kenmerken naar voren die in het algemeen bleken te differentiëren tussen waarheid en leugen. Deze kenmerken worden weergegeven in Tabel 23. Voor nummers 1 tot en met 9 geldt dat hun aanwezigheid samenhangt met geloofwaardigheid, terwijl indicatoren 10 tot en met 24 juist duiden op bedrog.

Het resultaat van DePaulo et al. (2003) lijkt wellicht indrukwekkend, maar bedacht moet worden dat hier alleen kenmerken worden genoemd die bleken te verschillen. Deze stammen echter uit een poel van 158 kenmerken, waarvan verondersteld was dat ze zouden differentiëren tussen waarheid en leugen. In werkelijkheid viel het dus tegen met de verschillen tussen waarheid en leugen: slechts 15% van de onderzochte kenmerken bleek te differentiëren. Dat betekent dat veel van de vermeende indicatoren (zoals afdrukken oogcontact tijdens liegen) in

werkelijkheid niet kloppen. Er blijkt dus een verschil te zijn tussen door mensen veronderstelde (non)verbale kenmerken van bedrog (de zogenoemde subjectieve kenmerken) en de kenmerken die blijkens wetenschappelijk onderzoek inderdaad differentiëren (de objectieve kenmerken).

Tabel 23. Indicatoren van bedrog (DePaulo et al., 2003).

1.	Hoeveelheid informatie
2.	Hoeveelheid details
3.	Plausibiliteit
4.	Logische structuur
5.	Hoeveelheid illustratoren (ondersteunende handbewegingen)
6.	Directe spreekstijl
7.	Medewerkingsgezindheid
8.	Spontane correcties
9.	Toegegeven geheugenhiaten
10.	Versmalde, samengetrokken lippen
11.	Discrepantie en ambivalentie
12.	Verminderde betrokkenheid
13.	Onzekere spreekstijl
14.	Hoofd naar boven richten en onderlip uitsteken
15.	Herhalingen
16.	Negativiteit en klagen
17.	Onaangename gezichtsuitdrukking (fronsen, weinig glimlachen)
18.	Nervositeit
19.	Trillende stem.
20.	Stemhoogte
21.	Pupildilatatie
22.	Friemelen aan eigen lichaam, kleren, of objecten
23.	Afdwalen naar externe gebeurtenissen/informatie
24.	Vertraagd antwoord (in afwezigheid van voorbereiding)

Nu moet worden onderkend dat de bulk van de wetenschappelijke kennis in dit geval voortkomt uit simulatieonderzoek. Deelnemers wordt dan gevraagd om een waar en een onwaar verhaal te vertellen en worden tijdens die opdracht gefilmd of geobserveerd. Bij dit soort simulatieonderzoek kan de kanttekening worden geplaatst dat het liegen tijdens een serieus politieverhoor niet zomaar te simuleren valt in een labonderzoek. Vrij en Mann (2001) rapporteerden een uniek onderzoek, waarin zij de

gedragskenmerken beschrijven van een moordenaar tijdens een echt verhoor. De betreffende man werd aangehouden op grond van zijn gelijkenis met een compositietekening. Tijdens het eerste verhoor ontkende de man iets met de moord te maken te hebben. Hij zei het slachtoffer zelfs niet te kennen. In de loop van dit verhoor vertelde hij de waarheid voor zover het zijn alibi voor de ochtend betrof, en loog hij over wat hij 's middags had gedaan. Enkele weken later bleek onder meer uit DNA-onderzoek dat het slachtoffer inderdaad in de auto van de verdachte had gezeten (een haar en kledingsporen werden aangetroffen). Geconfronteerd met deze bevinding bekende de man de moord, maar ook tijdens deze bekentenis loog hij aantoonbaar op sommige punten. Het unieke aan dit onderzoek is dat wegens uitgebreide getuigenverklaringen en sporenonderzoek precies is komen vast te staan hoe de moord zich heeft afgespeeld. De *ground truth* was met andere woorden bekend en dat komt niet vaak voor in een echte zaak. Nu de ground truth bekend was, kon ook precies worden achterhaald wanneer de verdachte loog en de waarheid sprak. De nonverbale indicatoren van bedrog die deze moordenaar liet zien kwamen grofweg overeen met datgene wat uit wetenschappelijk simulatieonderzoek blijkt (zie ook Mann et al., 2004, voor een replicatie met 14 verdachten in zware strafzaken).

Hoewel de leugendetector (poligraaf) in Nederland niet wordt gebruikt als bewijsmiddel, verdient hij hier een korte bespreking, omdat hij stoelt op een analyse van nonverbale gedragingen, en omdat hij in andere landen wel enige populariteit geniet. Met de poligraaf kunnen subtiele lichamelijke reacties worden gemeten, bijvoorbeeld in hartslagfrequentie, bloeddruk, pupildilatatie, ademhaling, hersenactiviteit (met behulp van een elektro-encefalogram, EEG), huidgeleiding (lees: zweetsecretie) en lichaams-temperatuur (Pavlidis et al., 2002). Minstens zo belangrijk als de fysiologische maten, is welke vragen de interviewer stelt aan de geïnterviewde. In de literatuur worden twee

poligraafprotocollen beschreven: de Controlevragen-techniek (CVT) en de Schuldige-Kennistest (SKT). De controlevragentechniek (in het Engels *Control Question Technique*; CQT) is de meest gebruikte verhoorvariant. De rationale achter de CQT is dat liegen een stressreactie veroorzaakt, die onder meer valt af te lezen aan een verhoogde hartslag, verhoogde bloeddruk en een toegenomen huidgeleiding. Bij de CQT worden relevante vragen (bijvoorbeeld: "Weet u iets van de inbraak?"), afgewisseld met controlevragen (bijvoorbeeld: "Heeft u in het verleden wel eens iets gestolen?"). Zowel schuldige als onschuldige verdachten zullen de relevante vragen ontkennend willen beantwoorden. Daartoe zullen de schuldigen, in tegenstelling tot de onschuldigen, moeten liegen, en dat zal een autonome fysiologische reactie met zich meebrengen. Om vals positieven (dat wil in deze context zeggen, de onterechte conclusie dat de onderzochte persoon liegt) te voorkomen, wordt de reactie op de relevante vragen vergeleken met de reactie op de controlevragen. Hoewel beide soorten vragen mogelijk bij zowel leugenaars als waarheidsprekers een stressreactie veroorzaken, mag worden verwacht die reactie op relevante vragen bij leugenaars nog sterker is, omdat zij op die vragen moeten liegen. Deze ingebouwde baseline (de reacties op de controlevragen) vereist dat de controlevragen een emotionele impact hebben die niet onderdoet voor die van de relevante vragen. Overigens is het aanbevelenswaardig om de scoring van de fysiologische reacties te kwantificeren volgens een vooraf vastgelegde methode. Men kan bijvoorbeeld koppels van relevante en controlevragen maken. Vervolgens wordt per koppel gekeken welke vraag de sterkste respons uitlokte. Als de controlevraag een sterkere respons uitlokt dan de relevante vraag, dan wordt 1 punt toegekend; als de relevante vraag een sterkere respons uitlokt, wordt -1 toegekend. Op deze manier is een negatieve score indicatief voor bedrog. Als tien koppels vragen zijn gesteld, kan bijvoorbeeld worden

afgesproken dat een score van -5 of lager duidt op bedrog, een score van +5 op waarheid en alles tussen -5 en +5 geen conclusie toelaat (*inconclusive*). Hoewel de precieze diagnostische kwaliteiten van de CQT moeilijk zijn te bepalen, wordt geschat dat de CQT 85% sensitief is en 60% selectief (Gannon et al, 2008; Patrick & Iacono, 1989).

De CQT is internationaal gezien het meest gebruikte poligraafprotocol. Ook in Nederland wordt ermee geëxperimenteerd, niet als bewijsmiddel maar bijvoorbeeld wel als hulpmiddel bij recidiverisicotaxatie (Gannon et al., 2008). Een tweede poligraafprotocol is de Schuldige Kennistest (SKT; in het Engels *Guilty Knowledge Test*; GKT, of *Concealed Information Test*; CIT; Lykken, 1959). Vergeleken bij de CQT, wijkt de CIT op twee manieren af. Op de eerste plaats is het theoretische uitgangspunt anders. Bij de CIT draait het namelijk niet om een stressrespons, maar om een oriëntatierespons. Die is fysiologisch overigens bijna identiek aan de stressrespons, behalve dat er sprake is van een hartslagverlaging in plaats van -verhoging. Op de tweede plaats werkt de CIT niet met relevante en controlevragen, maar met meerkeuzevragen. Een voorbeeld is: "De overvaller liet tijdens zijn vlucht iets vallen; was dat 1) een masker, 2) een sleutel, 3) een wapen, 4) geld, 5) een bivakmuts, 6) een zakdoek?". De ondervraagde dient deze antwoordopties rustig aan te horen. De poligrafist laat ongeveer 25 seconden tijd verstrijken na elke optie, want na zoveel tijd zijn alle fysiologische reacties weer beëindigd (een met EEG te meten P300 is zelfs al na twee seconden afgevlakt; Ben-Shakar et al., 2002). Als de ondervraagde beschikt over misdaadspecifieke kennis, zal hij een herkenningrespons vertonen nadat de verhoorder het relevante antwoordalternatief heeft voorgelezen. Strikt genomen hoeft de ondervraagde niet eens te antwoorden op de voorgelezen antwoordopties. De andere antwoordopties dienen als baseline-meting. Als de ondervraagde niets weet van het misdrijf, is de kans dat hij net op het verkeerde

moment een afwijkende fysiologische reactie laat zien, relatief klein, in dit voorbeeld 1/6. Bij twee vragen is die kans 1/36. Toepassing van de CIT vereist dat de politie over misdaadspecifieke kennis beschikt die verder alleen de dader heeft. Mochten veel details van het misdrijf reeds in de media zijn besproken, of in eerdere verhoren zijn prijsgegeven, dan wordt het moeilijk om goede meerkeuzevragen op te stellen. Volgens MacLaren (2001) is de sensitiviteit van de CIT ongeveer 76% en de selectiviteit 83% (zie ook Matsuda et al., 2012).

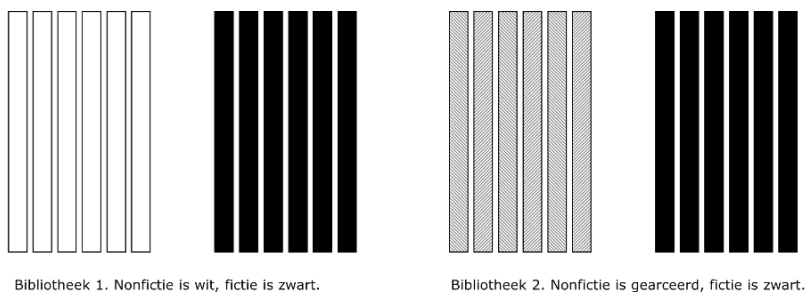
Op het eerste gezicht lijken de diagnostische kwaliteiten van de CQT en de CIT wellicht niet veel te verschillen. Uitgedrukt in likelihood ratio's is dat verschil er echter wel degelijk. Voor de CQT is de algemene likelihood ratio ongeveer ($85\% : 40\% =$) 2,1, terwijl die van de CIT ongeveer ($76\% : 17\% =$) 4,5 is. Los van het interview-protocol zijn er gevalsbeschrijvingen van verdachten, die wetend dat ze aan een poligraaftest onderworpen zouden gaan worden, op voorhand de waarheid opbiechtten. Dit bijeffect staat te boek als het *bogus pipeline effect* (Gannon et al., 2008).

Terugkijkend op de vier aanvliegroutes (aanloop, persoon, inhoud en presentatie), die alle hun beperkingen hebben, kan nog worden opgemerkt dat de eerste twee fundamenteel anders zijn dan de laatste twee. Dat wil zeggen, uitgaande van de ontstaansgeschiedenis en de persoon om te schatten hoe waarschijnlijk het is dat iemand liegt, zijn prospectieve redeneringen, omdat de kans op deceptie wordt voorspeld. Afgaan op de inhoud en de presentatie zijn echter retrospectieve benaderingen. Er wordt gekeken naar het eindresultaat, om te achterhalen of de verklaring oprecht of verzonnen is. Daarin schuilt een gevaar. Wetenschappelijk onderzoek zoals dat van Verschuere et al. (2023) maakt aannemelijk dat ware verklaringen gedetailleerd(er) zijn. Die bevinding wordt in de praktijk omgedraaid in de zin dat gedetailleerde narratieven geloofwaardig worden geacht. Daarmee wordt

de eerder genoemde transposed conditional redeneerfout gemaakt. Immers, dat de meeste verkrachters mannen zijn, betekent niet dat de meeste mannen verkrachters zijn. Ondertussen lijken rechters bij het waarderen van getuigenverklaringen vooral te vertrouwen op inhoudelijke kenmerken, met name consistentie, accuratesse, volledigheid en gedetailleerdheid en nonverbale kenmerken, met name authenticiteit en emotionaliteit (Verschuere et al., 2025).

Welke van de aanvliegroutes dan ook worden gebruikt, ons vermogen om leugens te detecteren is inherent beperkt. Het proces van deceptiedetectie lijkt daarnaast ook het speelveld van biases en andere hindernissen. Een daarvan is goedgelovigheid, ofwel *gullibility*, of *truth bias*. Dat goedgelovigheid een diepgewortelde neiging is, werd aangetoond in enkele onderzoeken van Gilbert en collega's. Deze onderzoekers waren onder meer geïnspireerd door een verschil van inzicht tussen Descartes en Spinoza. Waar Descartes, volgens Gilbert et al. (1990), van mening was dat mensen zich konden beschermen tegen onwaarheden door nieuwe informatie pas te geloven als er bewijs voor is gevonden, dacht Spinoza dat mensen helemaal niet in staat zijn om informatie te begrijpen zonder hem meteen te geloven. "Understanding and believing are merely two words for the same mental operation" zeggen Gilbert et al. (1993, p. 222) daarover. Informatie wordt in deze visie initieel begrepen en dus ook meteen voor juist aangenomen. In een volgend stadium kan de informatie (bijvoorbeeld vanwege strijdigheid met andere bronnen) alsnog worden "on(t)geloofd". Gilbert et al. (1990) baseren deze interpretatie van Spinoza op diens stelling 49 ("Er bestaan in de geest alleen wilsuitingen, met andere woorden, bevestigingen of ontkenningen, die een idee qua idee insluiten. De wil en het verstand zijn één en hetzelfde", 1677/2009, p. 211) uit *Ethica*.

Het lijkt onnatuurlijk om alle informatie aanvankelijk te geloven. Toch betogen Gilbert et al. (1990) dat dat juist efficiënt is. Ze trekken ter verduidelijking van dat standpunt een vergelijking met een fictieve bibliotheek. Stel dat die bibliotheek boeken slechts ordent op grond van één criterium: fictieboeken versus nonfictieboeken. Ter onderscheiding van deze twee soorten boeken, die allemaal standaard een wit omslag hebben, krijgen fictieboeken een zwarte rugband opgeplakt. Wie in deze bibliotheek langs een boekenwand loopt, ziet meteen wat fictieboeken zijn (zwarte rug) en wat nonfictieboeken zijn (witte rug; zie Figuur 18). Neem nu een tweede bibliotheek die ongeveer op dezelfde manier werkt. Fictieboeken krijgen ook hier een zwarte rug, maar nonfictieboeken krijgen in dit categoriseringssysteem een gearceerde rug (Figuur 18). Ook in deze bibliotheek zijn (zwarte) fictieboeken gemakkelijk te onderscheiden van (gearceerde) nonfictieboeken. Verschil is echter dat in deze bibliotheek twee keer zoveel werk is verzet als in de eerste bibliotheek, om hetzelfde onderscheid te maken. Daarom is de eerste bibliotheek efficiënter dan de tweede. Om dezelfde reden is het efficiënter om informatie standaard te geloven (tenzij redenen voor wantrouwen zich aandringen), dan om voor elke nieuwe informatie een doelbewuste afweging over de geloofwaardigheid ervan te maken.



Figuur 18. Analogie van twee bibliotheken met verschillend coderingssysteem.

Stel nu dat er een nieuw boek in de bibliotheek verschijnt, dat ontglipt aan de aandacht van de bibliothecaris. In de tweede bibliotheek zal dat boek (met witte, onbewerkte rug) meteen opvallen. In de eerste, normaliter efficiëntere bibliotheek zal het boek daarentegen niet opvallen, maar worden aangezien voor een nonfictieboek. Om deze reden vallen mensen regelmatig ten prooi aan onjuiste informatie. Dat is de prijs die betaald wordt voor efficiëntie. Gilbert en collega's bedachten een manier om de posities van Descartes en Spinoza te testen. Stramien van het paradigma is dat mensen nieuwe informatie krijgen. Tijdens het verwerken van die informatie worden ze gestoord. Later wordt hen gevraagd of de informatie correct of fout was. Volgens Descartes zullen deelnemers aan een dergelijk onderzoek even vaak een vals positieve (foute informatie geloven) als een vals negatieve fout (correcte informatie verwerpen) maken. Volgens Spinoza zullen ze echter vaker een vals positieve dan een vals negatieve fout maken. Bij binnenkomst wordt de informatie standaard geloofd, en door de interruptie blijft het subsequeute verwerpen uit. In een van hun onderzoeken lieten Gilbert et al. (1990) 35 studenten deelnemen aan een zogenaamd taalonderzoek. Deelnemers kregen 28 woorden te zien uit de niet bestaande taal "Hopi" met de Engels vertaling. Zo zagen ze op het scherm "A monisha is a star" en "A walive is a bear". Meteen na elk scherm volgde informatie over de juistheid van de vertaling; soms verscheen het woord "true", soms het woord "false". Deelnemers vernamen dus onmiddellijk of een vertaling klopte of niet. Daarnaast was het zo dat soms een geluid was te horen in het lab. In dat geval dienden de deelnemers onmiddellijk een bepaalde toets in te drukken. Dit laatste was bedoeld als onderbreking van de informatieverwerking. Het geluid was namelijk net te horen als de feedback over de juistheid van de vertaling op het scherm verscheen. Na dit alles werden de vertalingen nogmaals voorgelegd aan de deelnemers met de vraag of

ze klopten of niet. Kijkend naar de vertalingen waarbij geen onderbreking plaats had gevonden, bleek dat 23 procent van de incorrecte vertalingen toch als correct werden herinnerd (een vals positieve conclusie), terwijl ook 23 procent van de correcte vertalingen als incorrect werden herinnerd (een vals negatieve conclusie). Voor de onderbroken verwerkingen gold dat er 33 procent vals positieven waren en 17 procent vals negatieven. Blijkbaar leidde de onderbreking van de informatieverwerking er dus inderdaad toe dat deelnemers vaker incorrecte informatie gingen geloven. In het tweede onderzoek leerden twintig deelnemers onderscheid te maken tussen een valse, niet gemeende glimlach en een oprechte lach. Dit geschiedde door verschillende glimlachende mensen op het scherm te laten zien, met daarna feedback over de oprechtheid van de getoonde glimlach. Soms werd het verwerken van die feedback wederom verstoord door een concurrente taak (in dit geval differentiëren tussen hoge en lage geluiden). Tijdens de herinneropdracht bleek dat oprechte glimlachen in 65 procent van de gevallen als zodanig werden onthouden, ongeacht of de verwerking van de feedback was verstoord. Valse glimlachen werden echter minder vaak gedetecteerd in geval van interruptie (35 procent) dan in geval van ongestoorde verwerking (65 procent). In het laatste onderzoek kregen 30 deelnemers informatie over het niet bestaande dier Glark. De verwerking van de feedback over die informatie werd deze keer niet gestoord. In dit geval bleek dat 92 procent van de als correct gepresenteerde informatie ook als zodanig werd onthouden, terwijl 83 procent van de incorrecte informatie goed werd gedetecteerd. Als de onderzoekers voor aanvang van deze procedure eerst alle informatie presenteerden onder de uitdrukkelijke instructie dat de deelnemers de geloofwaardigheid ervan niet moesten proberen te schatten, maar dat ze de informatie zakelijk moesten verwerken, bleken de betreffende percentages 96 procent

en 76 procent te zijn. Kortom: de waarschuwing voor de hardnekkige truth bias bleek ineffectief.

Gilbert et al. (1993) lieten 71 deelnemers twee misdaaddossiers lezen, een over Tom die een vreemdeling beroofd had, en een over Kevin die een supermarkt overviel. In de dossiers zaten zeven onjuiste details. Die details waren in rood gedrukt. Het hele dossier verscheen op het computerscherm. Soms kroop een getallenreeks over het scherm en in dat geval moest de deelnemer een toets indrukken indien er een 5 in die reeks zat. Deze interventie beoogde de informatieverwerking te verstoren. Achteraf werden verschillende details voorgelegd aan de deelnemers. Sommige betroffen eerder in het dossier gelezen aspecten, andere betroffen de roodgedrukte details en weer andere waren details die niet in het dossier hadden gezeten. Van al deze details diende de deelnemer aan te geven of ze klopten, incorrect waren, of niet eerder waren getoond. Van de ongestoord verwerkte correcte details werd 93 procent herkend, terwijl dit percentage 89 bedroeg in geval van verstoorde verwerking. Wat de rode details betreft werd 69 procent inderdaad als incorrect herkend in geval van ongestoorde verwerking, maar slecht 34 procent in geval van gestoorde verwerking. Ook bleek verstoring ertoe te leiden dat niet eerder getoonde informatie minder vaak als zodanig werd herinnerd (86 procent, vergeleken bij 94 procent bij ongestoorde verwerking). Deze nieuwe details werden namelijk vaker als correcte details aangewezen.

Hoewel aan het onderzoek van Gilbert en collega's nauwelijks wordt gerefereerd in de leugendetectieliteratuur, levert het wel de meest theoriegestuurde, doordachte en wetenschappelijk onderzochte verklaring op voor goedgelovigheid. Niettemin valt er ook op het Spinoza-effect kritiek te leveren. Wat nu, als blijkt dat mensen niet alleen geloven in zaken die ze begrijpen en waar zich een voorstelling van kunnen maken, maar ook in zaken die ze niet snappen. Voor dat laatste is er geen wetenschap-

pelijk bewijs, maar er zijn wel enkele aanwijzingen. Zo is er het welbekende autoriteitsargument ("ik snap het niet, maar als hij dat zegt, dan zal het wel kloppen, want hij is expert"). Een spannend voorbeeld hiervan is te vinden in wat *the Sokal hoax* is gaan heten (Editors of *Lingua Franca*, 2000). Alan Sokal is een Amerikaans hoogleraar natuurkunde, die zich ergerde aan wetenschappelijke dikdoenerij. Volgens hem zijn er hele wetenschapsgebieden die zich overgeven aan abstract gefilosofeer. Om zijn idee te toetsen, schreef hij een afstotelijke, nauwelijks te snappen tekst die bol staat van de onzinnigheden (zoals de stelling dat zwaartekracht niet heus bestaat, maar slechts een wetenschappelijke fictie is), maar ook doorspekt is van referenties (Sokal, 1996a). Hij stuurde het ter publicatie op naar een vooraanstaand tijdschrift, dat het miraculeus accepteerde en publiceerde. Meteen daarna stuurde hij naar datzelfde tijdschrift een tweede artikel, waarin hij uiteenzette dat het eerste stuk onzinnig was en wat zijn motief was geweest om het te schrijven. Dit tweede stuk werd echter afgewezen door de redactie, om vervolgens door een concurrerend tijdschrift wel te worden gepubliceerd (Sokal, 1996b). Deze actie van Sokal toont aan dat ook wetenschappers ten prooi vallen aan goedgelovigheid en zomaar een artikel accepteren dat ze onmogelijkerwijs snappen. Als het inderdaad zo is dat mensen ook goedgelovig zijn ten aanzien van informatie die ze niet snappen, dan is die variant van goedgelovigheid niet te verklaren met het Spinoza-effect (dat immers het begrijpen van informatie centraal stelt). Dat zou betekenen dat goedgelovigheid nog dieper in de menselijke informatieverwerking is verankerd dan dat het Spinoza-effect veronderstelt.

Een andere bias die ons parten kan spelen bij het onderscheiden tussen leugen en waarheid is de *conjunction* bias. Die komt erop neer dat we narratieven die specifieker zijn veelal waarschijnlijker vinden dan algemenere varianten. Tversky en Kahneman (1983) legden in oktober

1980 aan 93 mensen de volgende tekst voor: "Stel dat Bjorn Borg de finale van Wimbledon 1981 bereikt. Sorteert de volgende scenario's van meest waarschijnlijk naar minst waarschijnlijk. A) Borg wint de wedstrijd. B) Borg verliest de eerste set. C) Borg verliest de eerste set, maar wint de wedstrijd. D) Borg wint de eerste set, maar verliest de wedstrijd". De gemiddelde posities van de vier scenario's waren 1,7 voor A, 2,7 voor B, 2,2 voor C en 3,4 voor D. Daarbij geldt hoe kleiner het getal, des te waarschijnlijker het scenario. Wat vooral in het oog springt, is dat de deelnemers scenario C waarschijnlijker achtten dan scenario B. In werkelijkheid is dat onmogelijk omdat scenario C vereist dat scenario B optreedt. C is een deelverzameling van B en daarom is B meeromvattend en dus waarschijnlijker dan C.

Een ander punt is dat voor veel van de genoemde criteria, de effectgrootte klein is. Ware narratieve bevatten meer details, maar de vraag is hoeveel meer.

Verder is het soms onduidelijk hoever een beoordelaar kan en wil gaan. Willen we echt weten of het goed gaat met een kennis die onze groet "hoe gaat het?" bevestigend beantwoordt? Vrij (2004) spreekt van het struisvogeleffect. En als we inderdaad de waarheid willen achterhalen, zijn we dan bereid om de wetenschappelijke literatuur over nonverbale kenmerken te bestuderen, of doen we het met de kennis die we al hebben? Gaan we met pen en papier het nonverbale gedrag noteren en dat vervolgens wikken en wegen, of beoordelen we de betrouwbaarheid van onze gesprekspartner op ons gevoel? Als we dat laatste doen, is de kans groot dat we automatisch terugvallen op vuistregels (heuristieken). In de literatuur zijn enkele heuristieken beschreven die soms tot correcte, maar soms ook tot incorrecte conclusies leiden. Om te beginnen lijken mensen er nauwelijks rekening mee te houden dat er individuele verschillen bestaan in nonverbaal gedrag. We nemen ons eigen gedrag als ijkpunt en vinden alles wat daarvan afwijkt verdacht (de

zogenoemde *expectancy-violation bias*; Vrij, 2004). In dit verband kan de "Othello-fout" worden genoemd. Shakespeare's Othello verdacht zijn vrouw Desdemona van overspel. Geconfronteerd met die onterechte verdenking werd Desdemona zenuwachtig en wilde ze Cassio (de man met wie ze een verhouding zou hebben) ten tonele roepen om haar onschuld te bewijzen. Toen Othello haar meedeelde dat hij Cassio alvast had vermoord, zag Desdemona dat ze in een uitzichtloze positie was gemanoeuvreed: ze kon haar onschuld nu namelijk niet meer aantonen. Haar daardoor ontstane zenuwachtige gedrag werd vervolgens door Othello gemisinterpreteerd als indicatie van bedrog. Het komt erop neer dat er vaak meerdere redenen zijn om verdacht gedrag te tonen dan liegen alleen. Sommige mensen zijn altijd zenuwachtig, of worden zenuwachtig van de gedachte dat iemand hun verdenkt – ook al zijn ze in werkelijkheid onschuldig. Mensen onderschatten niet alleen interindividuele verschillen, maar ook intra-individuele verschillen. Dat geldt *a fortiori* als we persoonlijkheidskenmerken gebruiken om de geloofwaardigheid van een verklaring te beoordelen, zoals hiervoor betoogd (vgl. de fundamentele attributiefout; Vrij, 2004).

Andere bias zijn de *probing heuristic* (mensen met wie we persoonlijk contact hebben gehad vinden we geloofwaardiger dan mensen die we niet persoonlijk hebben gesproken) en de *facial appearance heuristic* (aantrekkelijke mensen worden eerlijker gevonden dan onaantrekkelijke mensen; Vrij, 2004). Lev-Ari en Keysar (2010) lieten een aantal zinnen op band inspreken van het type "een giraf kan langer zonder water dan een kameel" en "mieren slapen nooit". Deze stellingen werden ingesproken door geboren Amerikanen en immigranten met een Pools, Turks, Duits, Koreaans, of Italiaans accent. Vervolgens moesten dertig in Amerika geboren Amerikanen de geloofwaardigheid van de stellingen beoordelen. Het bleek dat de stellingen geloofwaardiger werden bevonden

indien ze accentloos waren uitgesproken. Een ogenschijnlijk even verrassende bevinding is dat mensen minder goed zijn in bedrogdetectie wanneer ze een goed humeur hebben. Een slecht humeur maakt mensen blijkbaar kritischer (Forgas & East, 2008).

Gegeven de veelvuldig aangetoonde beperkte accuratesse in het onderscheiden tussen leugen en waarheid, zijn er ideeën geopperd om de accuratesse te doen stijgen. Om te beginnen zijn er tal van trainingen in het leren doorzien van bedrog. Driskell (2012) zette de resultaten van 16 onderzoeken naar detectietrainingen op een rij en concludeerde dat een goede training de prestatie van de cursist ietwat opkrikt. Onder een goede training wordt hier overigens verstaan een training waarbij de correcte informatie over indicatoren wordt gedoceerd, waar geoefend kan worden met detectie en waarbij de cursist ook uitgebreide feedback krijgt over zijn prestaties. Dat een leugendetectietraining staat bij correcte informatie over indicatoren lijkt een open deur. Toch is dat nog niet zo vanzelfsprekend als het lijkt. Zo worden in een invloedrijk Amerikaans studieboek voor de politie nonverbale indicatoren genoemd (zoals afgenomen oogcontact) die niet kloppen. Het gaat dan om de handleiding van Inbau et al. (2001) die inmiddels door tienduizenden of zelfs honderdduizenden verhoorders is gelezen (zie Kassin & Fong, 1999). Kassin en Fong deden een onderzoek naar de effecten van het lezen van het Inbau et al.-boek op de diagnostische accuratesse van leugendetectie. In de eerste fase van dat onderzoek begingen acht studenten een mock crime (bijvoorbeeld het stelen van een teddybeer) en werden zij meteen daarna verhoord over die misdaad. Dat verhoor duurde vijf minuten en werd uitgevoerd door een 48-jarige detective McCarthy. Deze man volgde een standaardprotocol waarin ook enige druk voorkwam. Zo zei hij op een gegeven moment: "Dammit, you were caught red-handed, so stop lying to me. I want to know exactly, minute by minute, where you were and what you were

doing before you were picked up" (Kassin & Fong, 1999, p. 504). Behalve de acht mock crime-plegers verhoorde McCarthy ook acht onschuldige studenten. Alle zestien studenten kregen de opdracht om McCarthy ervan te overtuigen dat ze onschuldig waren. Als dat lukte, kregen ze vijf dollar, maar als McCarthy hen schuldig zou achten, werden ze afgevoerd in een politiebusje en vijf minuten vastgehouden. Alle verhoren werden opgenomen op video. In fase 2 werden deze videofragmenten getoond aan veertig beoordelaars (eveneens studenten). Elke beoordelaar kreeg acht fragmenten te zien (vier van onschuldige en vier van schuldige verdachten). Twintig beoordelaars kregen vooraf echter een Inbau-training. Ze moesten de betreffende hoofdstukken uit het boek doornemen en kregen enkele instructievideo's te zien. Uiteindelijk ondergingen alle veertig beoordelaars een test van twintig items over de Inbau et al.-handleiding. Vervolgens beoordeelden ze de acht fragmenten en tot slot gaven ze aan hoe zeker ze waren van hun oordelen (1 = *volledig onzeker*; 10 = *volledig zeker*). De resultaten worden weergegeven in Tabel 24. De getrainde beoordelaars scoorden significant hoger op de test, hetgeen logisch is, omdat ze getraind waren. Zij waren ook zekerder van de juistheid van hun oordelen. Echter, hun accuratesse was significant minder dan die van ongetrainde beoordelaars. Conclusie was dat het bestuderen van de Inbau et al.-handleiding resulteert in een slechtere prestatie in plaats van een betere. Meissner en Kassin (2002) onderwierpen later ook nog eens 44 politiefunctionarissen aan hun test. Ook de prestatie van deze groep staat in Tabel 24. Zoals verwacht, presteerden de politiemensen niet boven kansniveau, hoewel ze behoorlijk zeker van hun kunnen waren.

Vrij (2004) noemt verder de impliciete of indirecte aanpak als een mogelijke manier om beter te worden in deceptiedetectie. Deze aanpak houdt in dat we ons niet moeten afvragen of iemand liegt, maar bijvoorbeeld of hij

ondersteunende handgebaren maakt. Op die manier verplichten we onszelf om op objectieve indicatoren te letten, en hopelijk vermijden we dan de niet-werkende indicatoren. Vrij (2004) noemt ook nog de baseline-methode. Het komt erop neer dat de beoordelaar pas tot een conclusie mag komen wanneer hij over vergelijkingsmateriaal beschikt. Het is hierbij dus onaanvaardbaar om op basis van één fragment te beoordelen of iemand liegt. Het te beoordelen fragment moet worden vergeleken met het gedrag van de beoordeelde persoon op een moment waarvan zeker is dat hij de waarheid sprak. Pas als het te beoordelen gedrag afwijkt van het gedrag tijdens het vertellen van een waarheid, is er reden om verder na te denken over de geloofwaardigheid van het narratief. Dit betekent dat iemand die zenuwachtig overkomt niet verdacht hoeft te worden van liegen indien hij tijdens het controlefragment (de baseline) ook zenuwachtig is. Daarbij kan worden opgemerkt dat het te beoordelen fragment liefst zo authentiek mogelijk moet zijn. Het lijkt er namelijk op dat mensen die vaak eenzelfde leugen vertellen daarmee hun herinnering aan de betreffende gebeurtenis overschrijven en na verloop van tijd dus niet meer liegen, maar hun narratief zelf zijn gaan geloven (Otgaar & Baker, 2018). Inmiddels blijkt uit het overzicht van Bond en DePaulo (2006) dat het gebruik van een baseline-fragment inderdaad de accuratesse doet stijgen, zij het van 52% naar 56%.

Tabel 24. Resultaten van Kassin en Fong (1999), en Meissner en Kassin (2001).

	Ongetraind	Getraind	Politie mensen
Testscore (0-20)	5,8	13,3	-
Accuratesse (0-8)	4,4 (56%)	3,7 (46%)	4,0 (50%)
Zelfvertrouwen (1-10)	5,9	6,6	7,1

Volgens Vrij et al. (2017) zou het onderzoek zich moeten richten op verhoormethoden die nieuwe duidelijke bedrogindicatoren opleveren. Als voorbeeld geven ze dat de getuige of verdachte gevraagd kan worden om zijn verhaal achterstevoren te vertellen. Dit is moeilijk, maar voor een leugenaar nog moeilijker dan voor iemand die oprecht uit zijn herinnering put. Door deze relatief sterke toename van de cognitieve belasting zouden nonverbale indicatoren meer uitgesproken kunnen worden. In diezelfde geest is het voorstel om getuigen een tekening te laten maken van het incident of de plaats delict. Dat zou voor leugenaars een moeilijkere uitdaging zijn dan voor mensen die naar waarheid verklaren. Volgens Vrij et al. (2017) kunnen dergelijke cognitieve interventies tijdens het verhoor de uiteindelijke deceptiedetectie-accuratesse doen toenemen van 56% naar 71%. Volgens Hartwig en Bond (2014) is het aan te bevelen om verschillende soorten indicatoren (verbale en nonverbale) te combineren. Volgens deze onderzoekers kan ook dat de accuratesse doen stijgen naar ongeveer 70%.

Wetenschappers reppen soms over *Ockham's razor*: we snijden alle overbodige informatie weg uit de vergelijking en dan blijft de zuinigste oplossing over. Misschien is dat in het geval van leugendetectie die van Verschuere et al. (2023): door ons te beperken tot de gedetailleerdheid van de verklaring, halen we allemaal, zonder enige speciale training een accuratesse van ongeveer 70%.

CONCLUSIE

In dit boek wordt een bloemlezing gegeven van rechtspsychologische onderwerpen. De kapstok waaraan die onderwerpen zijn opgehangen, zijn specifieke publicaties van Nederlandse rechtspsychologen. Terugblikkend hebben Nederlandse rechtspsychologen over vrijwel alle bekende rechtspsychologische thema's gepubliceerd, zoals besluitvorming, bias, geheugen, verhoren en leugendetectie. De bespreking van de verschillende onderzoeken is uiteraard niet beperkt tot Nederlandse bijdragen. Sterker, de meeste publicaties van Nederlandse rechtspsychologen verschenen in het Engels. Vandaar dat de meeste hoofdstuktitels Engelstalig zijn.

De keuze om de onderwerpen te ordenen aan de hand van specifieke wetenschappelijke publicaties is ingegeven door het doel om duidelijk te maken hoe wetenschappelijke inzichten kunnen worden toegepast in het (straf)recht. Daarbij gaat het niet alleen om het hebben van relevante inhoudelijke kennis over bijvoorbeeld het geheugen, maar ook om het toepassen van wetenschappelijke inzichten (zoals het letten op niets anders dan gedetailleerdheid bij het beoordelen van de betrouwbaarheid van verklaringen), of zelfs om het toepassen van wetenschappelijke procedures, bijvoorbeeld bij herkenningsprocedures, discussies met deskundigen, of het beantwoorden van de vraag naar de bewezenverklaring. *En passant* bieden de verschillende hoofdstukken hopelijk kennis over wat wetenschap wel en niet is. Deze doelen staan in dit boek centraler dan het bieden van een state of the art overzicht. Eerder lijkt de collectie besproken onderzoeken een historisch overzicht van de Nederlandse rechtspsychologie. In dit boek is een belangrijker doel dan op de hoogte zijn van het laatste nieuws, om inzicht te krijgen in de methodologie van de rechtspsychologie. Die verandert niet of nauwelijks, en kennis daarvan biedt een goede basis om inzichten die tot ons komen op hun merites

te beoordelen. Dat gezegd hebbende, niet alle besproken onderzoeken hebben het niveau van een experiment. Zo gebruikte Wagenaar (1988) geen controlegroep in zijn tailor-made onderzoek in de zaak Groscamp. Dat geldt ook voor het besproken onderzoek naar het langetermijngeheugen (Wagenaar & Groeneweg, 1990), crashing memories (Crombag et al., 1996; niemand had de betreffende filmpjes wél gezien), en de ontvoering door buitenaardse wezens (Otgaar et al., 2008; niemand was echt ontvoerd). Desondanks hebben al deze onderzoeken een unieke en sterke bijdrage geleverd aan de literatuur. Het is altijd zaak om goed te bezien welke conclusies een onderzoek, gegeven de opzet ervan, wel en niet toelaat.

LITERATUUR

- Aamodt, M.G. & Custer, H. (2006). Who can best catch a liar? A meta-analysis of individual differences in detecting deception. *The Forensic Examiner*, spring issue, 6-11.
- Aben, D. (2010). De facto. *RM Themis*, 223-232.
- American Psychiatric Association (APA) (2013). *Diagnostic and Statistical Manual of Mental Disorders*, 5th edition. Washington: APA.
- Arce, R., Selaya, A., Sanmarco, J. & Fariña, F. (2023). Implanting rich autobiographical false memories: Meta-analysis for forensic practice and judicial judgment making. *International Journal of Clinical and Health Psychology*.
<https://doi.org/10.1016/j.ijchp.2023.100386>
- Association of Forensic Science Providers (AFSP) (2009). Standards for the formulation of evaluative forensic science expert opinion. *Science and Justice*, 49, 161-164.
- Balcetis, E. & Dunning, D. (2006). See what you want to see: Motivational influences on visual perception. *Journal of Personality and Social Psychology*, 91(4), 612-625.
- Battista, F., Mangiulli, I & Curci, A. (2024). Eyewitness Memory: Factors affecting the formation of false memories. *Rassegna Italiana di Criminologia*, XVIII, 21-31.
- Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances: By the late rev. Mr. Bayes, communicated by mr. Price, in a letter to John Canton. *Philosophical Transactions of the Royal Society of London*, 53, 370-418.
- Bem, D. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100, 407-425.

- Ben-Shakhar, G., Bar-Hillel, M. & Kremnitzer, M. (2002). Trial by polygraph: Reconsidering the use of the guilty knowledge technique in court. *Law and Human Behavior*, 26, 527-541.
- Benton, T.R., Ross, D.F., Bradshaw, E., Thomas, W.N., & Bradshaw, G.S. (2006). Eyewitness memory is still not common sense: Comparing jurors, judges and law enforcement to eyewitness experts. *Applied Cognitive Psychology*, 20, 115-129.
- Billings, F. J., Taylor, T., Burns, J., Corey, D. L., Garven S. & Wood, J. M. (2007). Can reinforcement induce children to falsely incriminate themselves? *Law and Human Behavior*, 31, 125-139.
- Blair, J. P. (2007). The roles of interrogation, perception, and individual differences in producing compliant false confessions, *Psychology, Crime & Law*, 13, 173-186.
- Bloom, P. & Weisberg, D.S. (2007). Childhood origins of adult resistance to science. *Science*, 316, 996-997.
- Boeschen, L.E., Sales, B.D. & Koss, M.P. (1998). Rape trauma experts in the courtroom. *Psychology, Public Policy, and Law*, 4, 414-432.
- Bogaard, G., Meijer, E. H., Vrij, H., Broers, N. J. & Merckelbach, H. (2014). Contextual bias in verbal credibility assessment: Criteria-based content analysis, reality monitoring and scientific content analysis. *Applied Cognitive Psychology*, 28, 79-90.
- Bond, C.F. & DePaulo, B.M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, 10, 214-234.
- Bond, C.F. & DePaulo, B.M. (2008). Individual differences in judging deception: Accuracy and bias. *Psychological Bulletin*, 134, 477-492.
- Boone, K. B., Salazar, X., Lu, P., Warner-Chacon, K. & Razani, J. (2002). The Rey 15-item recognition trial: A technique to enhance sensitivity of the Rey 15-item memorization test. *Journal of Clinical and Experimental Neuropsychology*, 24(5), 561-573.

- Brainerd, C.J., Reyna, V.F. & Ceci, S.J. (2008). Developmental reversals in false memory: A review of data and theory. *Psychological Bulletin*, 134, 343-382.
- Brown, R. & Kulik, J. (1977). Flashbulb memories. *Cognition*, 5, 73-99.
- Calado, B., Luke, T. J., Connolly, D. A., Landstrom, S. & Otgaar, H. (2021). Implanting false autobiographical memories for repeated events. *Memory*, 29(10), 1320-1341.
- Candel, I., Merckelbach, H., Loyen & S. Reyskens, H. (2005). "I hit the shift-key and then the computer crashed": Children and false admissions. *Personality and Individual Differences*, 38, 1381-1387.
- Carmichael, L.C., Hogan, H.P. & Walter, A.A. (1932). An experimental study of the effect of language on the reproduction of visually perceived forms. *Journal of Experimental Psychology*, 15, 73-86.
- Catlin, M., Wilson, D., Redlich, A. D., Bettens, T., Meissner, C., Bhatt, S. & Brandon, S. (2024). Interview and interrogation methods and their effects on true and false confessions: A systematic review update and extension. *Campbell Systematic Reviews*, 20(4), e1441. doi: 10.1002/cl2.1441.
- Chambers, C. (2017). *The 7 deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton: University press.
- Clark, S.E., Howell, R.T. & Davey, S.L. (2008). Regularities in eyewitness identification. *Law and Human Behavior*, 32, 187-218.
- Clark, S.E. & Wells, G.L. (2008). On the diagnosticity of multiple-witness identifications. *Law and Human Behavior*, 32, 406-422.
- Cleiren, C.P.M. (2010). De rechterlijke overtuiging: Een sprong met hindernissen. *RM Themis* 5/6, 259-267.
- Coursol, A. & Wagner, E.E. (1986). Effect of positive findings on submission and acceptance rates: A note on meta-

- analysis bias. *Professional Psychology: Research and Practice*, 17, 136-137.
- Crombag, H.F.M. (1999). Collaborative story-telling: A hypothesis in need of experimental testing. *Psychology, Crime & Law*, 5, 279-289.
- Crombag, H.F.M., Van Koppen, P.J. & Wagenaar, W.A. (1994). *Dubieuze zaken: De psychologie van strafrechtelijk bewijs*. Amsterdam: Contact.
- Crombag, H.F.M., Wagenaar, W.A., & Van Koppen, P.J. (1996). Crashing memories and the problem of 'source monitoring', *Applied Cognitive Psychology*, 10, 95-104.
- Crowne, D.P. & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 24, 349-354.
- Darley, J.M. & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology*, 8, 377-383.
- De Bock, R. H. (2011). *Tussen waarheid en onzekerheid: over het vaststellen van feiten in de civiele procedure*. Deventer: Kluwer.
- De Bruijn, C. (2010). *Van Tofu krijg je geheugenverlies: Gekonkel en gestuntel met statistiek in media, politiek en reclame*. Den Haag: Het Tweede Gezicht.
- De Jong, M., Wagenaar, W.A., Wolters, G. & Verstijnen, I.M. (2005). Familiar face recognition as a function of distance and illumination: A practical tool for use in the courtroom. *Psychology, Crime & Law*, 11, 87-97.
- De Vries, P. E. (1993). A draft of Iethe: A neglected statement from the works of Sigmund Freud. *Psychotherapy*, 30(3), 524-530.
- Deffenbacher, K.A. (1985). Eyewitness recall, the Yerkes-Dodson law, and optimal-level theory. *Medicine and Law*, 4, 361-372.
- Dekens, K. & Van der Sleen, J. (2010). *Handleiding Het kind als getuige*. Amsterdam: Stapel & De Koning.

- Denney, R.L. (1996). Symptom validity testing of remote memory in a criminal forensic setting. *Archives of Clinical Neuropsychology*, 11, 589-603.
- DePaulo, B.M., Kashy, D.A., Kirkendol, S.E., Wyer, M.M. & Epstein, J.A. (1996). Lying in everyday life. *Journal of Personality and Social Psychology*, 70, 979-995.
- DePaulo, B.M., Lindsay, J.J., Malone, B.E., Muhlenbruck, L., Charlton, K. & Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, 129, 74-118.
- Dodier, O. & Denault, V. (2018). The Griffiths question map: A forensic tool for expert witnesses' assessments of witnesses and victims' statements. *Journal of Forensic Sciences*, 63, 266-274.
- Doob, A.N. & Kirshenbaum, H.M. (1973). Bias in police lineups: Partial remembering. *Journal of Police Science and Administration*, 1, 287-293.
- Driskell, J.E. (2012). Effectiveness of deception detection training: A meta-analysis. *Psychology, Crime & Law*, 18, 713-731.
- Dror, I.E., Péron, A.E., Hind, S.L., & Charlton, D. (2005). When emotions get the better of us: The effect of contextual top-down processing on matching fingerprints. *Applied Cognitive Psychology*, 19, 799-809.
- Dubelaar, M. J. (2014). *Betrouwbaar getuigenbewijs: Totstandkoming en waardering van strafrechtelijke getuigenverklaringen in perspectief*. Deventer: Kluwer.
- Dunning, D. & Perretta, S. (2002). Automaticity and eyewitness accuracy: A 10-to-12-seconds rule for distinguishing accurate from inaccurate positive identifications. *Journal of Applied Psychology*, 87, 951-962.
- Editors of Lingua Franca (2000). *The Sokal hoax: The sham that shook the academy*. Nebraska: University of Nebraska.
- Edwards, K. & Bryan, T.S. (1997). Judgmental biases produced by the instructions to disregard: The

- (paradoxical) case of emotional information. *Personality and Social Psychology Bulletin*, 23, 849-864.
- Einhorn, H.J. & Hogarth, R.M. (1986). Judging probable cause. *Psychological Bulletin*, 99, 3-19.
- Ekman, P. & O'Sullivan, M. (1991). Who can catch a liar? *American Psychologist*, 46, 913-920.
- Ekman, P., O'Sullivan, M. & Frank, M.G. (1999). A few can catch a liar. *Psychological Science*, 10, 263-266.
- Engelhard, I.M., van den Hout, M.A. & McNally, R.J. (2008). Memory consistency for traumatic events in Dutch soldiers deployed to Iraq. *Memory*, 16, 3-9.
- Evans, J.S.B.T., Barston, J.L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition*, 11, 295-306.
- Faust, D. & Ziskin, J. (1988). The expert witness in psychology and psychiatry. *Science*, 241, 31-35.
- Fenton, N., Neil, M., & Berger, D. (2016). Bayes and the law. *Annual Review of Statistics and its Application*, 3, 51-77.
- Fenton, N., Neil, M. & Lagnado, D.A. (2013). A general structure for legal arguments about evidence using Bayesian networks. *Cognitive Science*, 37, 61-102.
- Fisher, J.S. & Radvansky, G.A. (2018). Patterns of forgetting. *Journal of Memory and Language*, 102, 130-141.
- Forer, B.R. (1949). The fallacy of personal validation: A classroom demonstration of gullibility. *Journal of Abnormal and Social Psychology*, 44, 118-123.
- Forgas, J.P. & East, R. (2008). On being happy and gullible: Mood effects on skepticism and the detection of deception. *Journal of Experimental Social Psychology*, 44, 1362-1367.
- Forrest, K. D., Wadkins, T. A. & Larson, B. A. (2006). Suspect personality, police interrogations, and false confessions: Maybe it is not just the situation. *Personality and Individual Differences*, 40, 621-628.

- Frenda, S. J., Berkowitz, S. R., Loftus, E. F. & Fenn, K. M. (2016). Sleep deprivation and false confessions. *PNAS*, *113*, 2047-2050.
- Friedrich, J. (1993). Primary error detection and minimization (PEDMIN) strategies in social cognition: A reinterpretation of confirmation bias phenomena. *Psychological Review*, *100*, 298-319.
- Gannon, T.A., Beech, A.R. & Ward, T. (2008). Does the polygraph lead to better risk prediction for sexual offenders? *Aggression and Violent Behavior*, *13*, 29-44.
- Garb, H.N. & Boyle, P.A. (2003). Understanding why some clinicians use pseudoscientific methods: Findings from research on clinical judgment. In S.O. Lilienfeld, S.J. Lynn & J.M. Lohr (red.) *Science and pseudoscience in clinical psychology* (pp. 17-38). London: The Guilford press.
- Garb, H.N., Wood, J.M., Lilienfeld, S.O., & Nezworski, M.T. (2005). Roots of the Rorschach controversy. *Clinical Psychology Review*, *25*, 97-118.
- Geijssen, K., Vanbelle, S., Kop, N., & de Ruiter, C. (2018). The interrogation of vulnerable suspects in the Netherlands: An exploratory study. *Investigative Interviewing: Research and Practice*, *9*, 34-51.
- Geiselman, R.E., Fisher, R.P., MacKinnon, D.P. & Holland, H.L. (1986). Enhancement of eyewitness memory with the cognitive interview. *American Journal of Psychology*, *99*, 385-401.
- Gerlach, P., Teodorescu, K., & Hertwig, R. (2019). The truth about lies: A meta-analysis on dishonest behavior. *Psychological Bulletin*, *145*, 1-44.
- Gilbert, D.T., Krull, D.S. & Malone, P.S. (1990). Unbelieving the unbelievable: Some problems in the rejection of false information. *Journal of Personality and Social Psychology*, *59*, 601-613.
- Gilbert, D.T., Tafarodi, R.W. & Malone, P.S. (1993). You can't not believe anything you read. *Journal of Personality and Social Psychology*, *65*, 221-233.

- Godden, D.R. & Baddeley, A.D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology*, 66, 325-331.
- Gongola, J., Scurich, N., & Quas, J.A. (2017). Detecting deception in children: A meta-analysis. *Law and Human Behavior*, 41, 44-54.
- Granhag, P. A., Stromwall, L. A., & Billings, J. F. (2003). "I'll never forget the sinking ferry": How social influence makes false memories surface. In: M. Vanderhallen, G. Vervaeke, P. J. van Koppen, & J. Goethals (ed.). *Much ado about crime: Chapters on psychology and law* (pp. 129-240). Brussel: Politeia.
- Greenberg, S.A. & Shuman, D.W. (1997). Irreconcilable conflict between therapeutic and forensic roles. *Professional Psychology: Research and Practice*, 28, 50-57.
- Greenspan, R. & Scurich, N. (2016). The interdependence of perceived confession voluntariness and case evidence. *Law and Human Behavior*, 40, 650-659.
- Greenwald, A.G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, 82, 1-20.
- Gronlund, S.D., Wixted, J.T., & Mickes, L. (2014). Evaluating eyewitness identification procedures using receiver operating characteristic analyses. *Current Directions in Psychological Science*, 23, 3-10.
- Grove, W.M., Zald, D.H., Lebow, B. S., Snitz, B.E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, 12, 19-30.
- Gudjonsson, G.H. (1984). A new scale of interrogative suggestibility. *Personality and Individual Differences*, 5, 303-314.
- Gudjonsson, G.H. (1989). Compliance in an interrogative situation: A new scale. *Personality and Individual Differences*, 10, 535-540.
- Gudjonsson, G.H. (1997). *The Gudjonsson Suggestibility Scales*. East Sussex: Psychology Press.

- Gudjonsson, G.H. (2003). *The psychology of interrogations and confessions: A handbook*. Chichester: John Wiley & sons.
- Gudjonsson, G.H., Sigurdsson, J.F., & Sigfusdottir, I.D. (2009). Interrogation and false confessions among adolescents in seven European countries: What background and psychological variables best discriminate between false confessors and non-false confessors? *Psychology, Crime & Law*, 15, 711-728.
- Haack, S. (2003). *Defending science – within reason: Between scientism and cynicism*. New York: Prometheus Books.
- Harris, W.S., Gowda, M., Kolb, J.W., Strychacz, C.P., Vacek, J.L., Jones, P.G., Forker, A., O'Keefe, J.H., McCallister, B.D. (1999). A randomized, controlled trial of the effects of remote, intercessory prayer on outcomes in patients admitted to the coronary care unit. *Archives of Internal Medicine*, 159, 2273-2278.
- Hartwig, M. & Bond, C.F. (2014). Lie detection from multiple cues: A meta-analysis. *Applied Cognitive Psychology*, 28, 661-676.
- Hauch, V., Sporer, S.L., Masip, J., & Blandón-Gitlin, I. (2017). Can credibility criteria be assessed reliably? A meta-analysis of criteria-based content analysis. *Psychological Assessment*, 29, 819-834.
- Hofvind, S. et al. (2005). Influence of review design on percentages of missed interval breast cancers: Retrospective study of interval cancers in a population-based screening program. *Radiology*, 237, 437-443.
- Hope, L., Gabbert, F., & Fisher, R.P. (2011). From laboratory to the street: Capturing witness memory using the self-administered interview. *Legal and Criminological Psychology*, 16, 211-226.
- Hordijk, M.L. & Sjerps, M.J. (2005). De bewijswaarde van forensisch vergelijkend glasonderzoek. In M.J. Sjerps & J.A. Coster van Voorhout (red.) *Het onzekere bewijs*:

- Gebruik van statistiek en kansrekening in het strafrecht* (pp. 211-228). Deventer: Kluwer.
- Horgan, A. J., , Russano, M. B., Meissner, C. A. & J.R. Evans, J. R. (2012). Minimization and maximization techniques: Assessing the perceived consequences of confessing and confession diagnosticity', *Psychology, Crime & Law*, 18, 65-78.
- Horowitz, S.W. (1991). Empirical support for statement validity assessment. *Behavioral Assessment*, 13, 293-313.
- Horselenberg, R., Merckelbach, H. & Josephs, S. (1999). Individual Differences and False Confessions: A Conceptual Replication of Kassin and Kiechel (1996). *Psychology, Crime & Law*, 9(1), 1-8. DOI: 10.1080/10683160308141
- Horselenberg, R., Merckelbach, H., Smeets, T., Franssens, D., Peters G.J.Y. & Zeles, G. (2006). False confessions in the lab: Do plausibility and consequences matter? *Psychology, Crime and Law*, 12, 61-75.
- Houben, S.T.L., Otgaar, H., Roelofs, J., Merckelbach, H., & Muris, P. (2020). The effects of eye movements and alternative dual tasks on the vividness and emotionality of negative autobiographical memories: A meta-analysis of laboratory studies. *Journal of Experimental Psychopathology*, 11, 1-14.
- Inbau, F.E., Reid, J.E., Buckley, J.P. & Jayne, B.C. (2001). *Criminal interrogation and confessions*. Gaithersburg: Aspen publishers.
- Jacoby, L.L., Kelley, C.M., Brown, J., & Jasechko, J. (1989). Becoming famous overnight: Limits on the ability to avoid unconscious influences of the past. *Journal of Personality and Social Psychology*, 56, 326-338.
- Jarosz, A.F. & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *Journal of Problem Solving*, 7, 2-9.
- Jelicic, M., Smeets, T., Peters, M.J.V., Candel, I., Horselenberg, R., & Merckelbach, H. (2006).

- Assassination of a controversial politician: Remembering details from another non-existent film. *Applied Cognitive Psychology*, 20, 591-596.
- Jenkins, J.G. & Dallenbach, K.M. (1924). Oblivescence during sleep and waking. *American Journal of Psychology*, 35, 605-612.
- Johnson, M.K., Foley, M.A., Suengas, A.G., & Raye, C.L. (1988). Phenomenal characteristics of memories for perceived and imagined autobiographical events. *Journal of Experimental Psychology: General*, 117, 371-376.
- Jordan, S., Hartwig, M., Wallace, B., Dawson, E & Xhahani, A. (2012). Early versus late disclosure of evidence: Effects on verbal cues to deception, confessions, and lie catchers' accuracy. *Journal of Investigative Psychology and Offender Profiling*, 9, 1-12.
- Kaasa, S.O., Cauffman, E., Clarke-Stewart, K.A. & Loftus, E.F. (2013). False accusations in an investigative context: Differences between suggestible and non-suggestible witnesses. *Behavioral Sciences and the Law*, 31, 574-592.
- Kalven, H. & Zeisel, H. (1966). *The American jury*. Boston: Little brown.
- Kassin, S.M. (1997). The psychology of confession evidence. *American Psychologist*, 52, 221-233.
- Kassin, S.M. (2005). On the psychology of confessions: Does innocence put innocents at risk? *American Psychologist*, 60, 215-228.
- Kassin, S.M., Dror, I.E., & Kukucka, J. (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of Applied Research in Memory and Cognition*, 2, 42-52.
- Kassin, S.M. & Fong, C.T. (1999). "I'm innocent!": Effects of training on judgments of truth and deception in the interrogation room. *Law and Human Behavior*, 23, 499-516.

- Kassin, S.M., Goldstein, C.J. & Savitsky, K. (2003). Behavioral confirmation in the interrogation room: On the dangers of presuming guilt. *Law and Human Behavior*, 27, 187-203.
- Kassin, S.M. & Kiechel, K.L. (1996). The social psychology of false confessions: Compliance, internalization, and confabulation. *Psychological Science*, 7, 125-128.
- Kassin, S.M., Kukucka, J., Lawson, V.Z. & DeCarlo, J. (2014). Does video recording alter the behaviour of police during interrogation? A mock crime-and-investigation study. *Law and Human Behavior*, 38, 73-83.
- Kassin, S.M. & Neumann, K. (1997). On the power of confession evidence: An experimental test of the fundamental difference hypothesis. *Law and Human Behavior*, 21, 469-484.
- Kassin, S.M. & Sukel, H. (1997). Coerced confessions and the jury: An experimental test of the "harmless error" rule. *Law and Human Behavior*, 21, 27-46.
- Kassin, S.M., Tubb, V.A., Hosch, H.M., & Memon, A. (2001). On the "general acceptance" of eyewitness testimony research. *American Psychologist*, 56, 405-416.
- Kebbell, M.R. & Wagstaff, G.F. (1998). Hypnotic interviewing: The best way to interview eyewitnesses? *Behavioral Sciences and the Law*, 16, 115-129.
- Kelley, M.R. & Nairne, J.S. (2003). Remembering the forgotten? Reminiscence, hypermnnesia and memory for order. *The Quarterly Journal of Experimental Psychology*, 56A, 577-599.
- Klaver, J. R. Lee, Z, & Rose, V. G. (2008). Effects of personality, interrogation techniques and plausibility in an experimental false confession paradigm. *Legal and Criminological Psychology*, 13, 71-88.
- Kokshoorn, B., Blankers, B.J., de Zoete, J., & Berger, C.E.H. (2017). Activity level DNA evidence evaluation: On propositions addressing the actor of the activity. *Forensic Science International*, 278, 115-124.

- Kuhn, D., Phelps, E. & Walters, J. (1985). Correlational reasoning in an everyday context. *Journal of Applied Developmental Psychology*, 6, 85-97.
- Kukucka, J., Kassin, S.M., Zapf, P.A., & Dror, I.E. (2017). Cognitive bias and blindness: A global survey of forensic science examiners. *Journal of Applied Research in Memory and Cognition*, 6, 452-459.
- La Rooy, D., Pipe, M.E. & Murray, J.E. (2005). Reminiscence and hypermnesia in children's eyewitness memory. *Journal of Experimental Child Psychology*, 90, 235-254.
- Leichtman, M.D. & Ceci, S.J. (1995). The effects of stereotypes and suggestions on preschoolers' reports. *Developmental Psychology*, 31, 568-578.
- Lev-Ari, S. & Keysar, B. (2010). Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology*, 46, 1093-1096.
- Lewandowsky, S. Oberauer, K. & Gignac, G.E. (2013). NASA faked the moon landing – therefore, (climate) science is a hoax: An anatomy of the motivated rejection of science. *Psychological Science*, 24(5), 622-633.
- Lewis, D.O., Yeager, C.A., Swica, Y., Pincus, J.H., & Lewis, M. (1997). Objective documentation of child abuse and dissociation in 12 murderers with dissociative identity disorder. *American Journal of Psychiatry*, 154, 1703-1710.
- Lezak, M.D. (1983). *Neuropsychological assessment*. New York: Oxford university press.
- Liebman, J.S., Blackburn, S., Mattern, D., & Waisnor, J. (2012-2013). The evidence of things not seen: Non-matches as evidence of innocence. *Iowa Law Review*, 98, 577-688.
- Lilienfeld, S.O. & Landfield, K. (2008). Science and pseudoscience in law enforcement: A user-friendly primer. *Criminal Justice and behavior*, 35, 1215-1230.

- Lilienfeld, S.O., Wood, J.M. & Garb, H.N. (2006). Why questionable psychological tests remain popular. *The Scientific Review of Alternative Medicine*, 10, 6-15.
- Loftus, E.F. (1976). Unconscious transference in eyewitness identification. *Law and Psychology Review*, 2, 93-98.
- Loftus, E.F. & Palmer, J.C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, 13, 585-589.
- Luus, C.A.E. & Wells, G.L. (1994). The malleability of eyewitness confidence. Co-witness and perseverance effects, *Journal of Applied Psychology*, 79, 714-723.
- Lykken, D.T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology*, 43, 385-388.
- MacLaren, V.V. (2001). A quantitative review of the guilty knowledge test. *Journal of Applied Psychology*, 86, 674-683.
- Malle, B.F., Knobe, J.M. & Nelson, S.E. (2007). Actor-observer asymmetries in explanations of behavior: New answers to an old question. *Journal of Personality and Social Psychology*, 93, 491-514.
- Mamede, S. & Schmidt, H. G. (2023). Deliberate reflection and clinical reasoning: Founding ideas and empirical findings. *Medical Education*, 57, 76-85.
- Mann, S., Vrij, A. & Bull, R. (2004). Detecting true lies: Police officers' ability to detect suspects' lies. *Journal of Applied Psychology*, 89, 137-149.
- Martire, K.A., Kemp, R.I., Watkins, I., Sayle, M.A. & Newell, B.R. (2013). The expression and interpretation of uncertain forensic science evidence: Verbal equivalence, evidence strength, and the weak evidence effect. *Law and Human Behavior*, 37, 197-207.
- Masip, J.M., Sporer, S.L., Garrido, E. & Herrero, C. (2005). The detection of deception with the reality monitoring approach: A review of the empirical evidence. *Psychology, Crime & Law*, 11, 99-122.

- Matsuda, I., Nittono, H. & Allen, J.J.B. (2012). The current and future status of the concealed information test for field use. *Frontiers in Psychology*, 3, 1-11.
- McGrayne, S.B. (2011). *The theory that would not die: How Bayes' rule cracked the enigma code, hunted down Russian submarines, & emerged triumphant from two centuries of controversy*. London: Yale university press.
- McNally, R.J. (2001). On the scientific status of cognitive appraisal models of anxiety disorder. *Behaviour Research and Therapy*, 39, 513-521.
- McNally, R.J., Lasko, N.B., Clancy, S.A., Macklin, M.L., Pitman, R.K, & Orr, S.P. (2004). Psychophysiological responding during script-driven imagery in people reporting abduction by space aliens. *Psychological Science*, 15, 493-497.
- Meehl, P.E. (1956). Wanted – a good cookbook. *American Psychologist*, 11, 263-272.
- Meissner, C.A. & Brigham, J.C. (2001). A meta-analysis of the verbal overshadowing effect in face identification. *Applied Cognitive Psychology*, 15, 603-616.
- Meissner, C.A. & Kassin, S.M. (2002). "He's guilty!": Investigator bias in judgments of truth and deception. *Law and Human Behavior*, 26, 469-480.
- Meissner, C.A., Redlich, A.D., Michael, S.W., Evans, J.R., Camiletti, C.R., Bhatt, S., & Brandon, S. (2014). Accusatorial and information-gathering interrogation methods and their effects on true and false confessions: A meta-analytic review. *Journal of Experimental Criminology*, 10, 459-486.
- Memon, A., Meissner, C.A., & Fraser, J. (2010). The cognitive interview: A meta-analytic review and study space analysis of the past 25 years. *Psychology, Public Policy, and Law*, 16, 340-372.
- Merckelbach, H., Hauer, B., & Rassin, E. (2002). Symptom validity testing of feigned dissociative amnesia: A simulation study. *Psychology, Crime & Law*, 8, 311-318.

- Merckelbach, H., Otgaar, H., & Lynn, S.J. (2022). Empirical research on fantasy proneness and its correlates 2000-2018: A meta-analysis. *Psychology of Consciousness: Theory, Research, and Practice*, 9(1), 2-26.
- Merckelbach, H. & van de Ven, V. (2001). Another white Christmas: Fantasy proneness and reports of 'hallucinatory experiences' in undergraduate students. *Journal of Behavior Therapy and Experimental Psychiatry*, 32, 137-144.
- Merten, T. (2025). A case of claimed complete blindness: Forced-choice testing in forensic neuropsychological examinations. *Applied Neuropsychology: Adult*. DOI: 10.1080/23279095.2025.2506114
- Mickes, L., Flowe, H.D., & Wixted, J.T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous versus sequential lineups. *Journal of Experimental Psychology: Applied*, 18, 361-376.
- Miller, G.A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63, 81-97.
- Monds, L.A., Paterson, H.M., Kemp, R.I., & Bryant, R.A. (2013). Individual differences in susceptibility to false memories for neutral and trauma-related words. *Psychiatry, Psychology, and Law*, 20, 399-411.
- Mul, V. & Pols, E.J.V. (1999). Zaans verhoren: Een vervolghverhaal. *Tijdschrift voor de Rechterlijke Macht*, (4), 129-134.
- Multhaup, K.S., Johnson, M.D. & Tetirick, J.C. (2005). The wane of childhood amnesia for autobiographical and public event memories. *Memory*, 13, 161-173.
- Murrie, D.C., Boccaccini, M.T., Guarnera, L.A., & Rufino, K.A. (2013). Are forensic experts biased by the side that retained them? *Psychological Science*, 24, 1889-1897.
- Nahari, G., Ashkenazi, T., Fisher, R.P., Granhag, P.A., Hershkowitz, I., Masip, J. Meijer, E.H., Nisin, Z., Sarid, N., Taylor, P.J., Verschuere, B., & Vrij, A. (2019).

- 'Language of lies': Urgent issues and prospects in verbal lie detection research. *Legal and Criminological Psychology*, 24, 1-23.
- Nahari, G. & Vrij, A. (2014). Can I borrow your alibi? The applicability of the verifiability approach to the case on an alibi witness. *Journal of Applied Research in Memory and Cognition*, 3, 89-94.
- Nahari, G., Vrij, A., & Fisher, R.P. (2014). The verifiability approach: Countermeasures facilitate its ability to discriminate between truths and lies. *Applied Cognitive Psychology*, 28, 122-128.
- Narchet, F. M., Meissner, C. A. & Russano, M. B. (2011). Modeling the influence of investigator bias on the elicitation of true and false confessions. *Law and Human Behavior*, 35, 452-463.
- Nash, R. A. & Wade, K. A. (2009). Innocent but proven guilty: Eliciting internalized false confessions using doctored-video evidence. *Applied Cognitive Psychology*, 23, 624-637.
- Nederlands Forensisch Instituut (2017). *De reeks waarschijnlijkheidstermen van het NFI en het Bayesiaanse model voor interpretatie van bewijs*. Den Haag: NFI.
- Nickerson, R.S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2, 175-220.
- Nisbett, R.E. & Wilson, T.D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Bulletin*, 84, 231-259.
- Oberlader, V.A., Naefgen, C., Koppehele-Gossel, J., Quinten, L., Banse, R., & Schmidt, A.F. (2016). Validity of content-based techniques to distinguish true and fabricated statements: A meta-analysis. *Law and Human Behavior*, 40, 440-457.
- Orthey, R., Meijer, E., Kooistra, E. & Broers, N. (2022). How to detect concealed crime knowledge in situations with

- little information using the forced choice test. *Collabra: Psychology*. Doi.org/10.1525/collabra.37483.
- Ost, J., Granhag, P. A., Udell, J., & af Hjelmsater, E. R. (2008). Familiarity breeds distortion: The effect of media exposure on false reports concerning media coverage of the terrorist attacks in London on 7 July 2005. *Memory*, 16(1), 76-85.
- Ost, J., Hogbin, I., & Granhag, P. A. (2006). Altering false reports via confederate influence. *Social Influence*, 1(2), 105-116.
- Ost, J., Vrij, A., Costall, A., & Bull, R. (2002). Crashing memories and reality monitoring: Distinguishing between perceptions, imaginations and 'false memories'. *Applied Cognitive Psychology*, 16, 125-134.
- Otgaar, H. & Baker, A. (2018). When lying changes memory for the truth. *Memory*, 26, 2-14.
- Otgaar, H., Candel, I., Merckelbach, H. & Wade, K.A. (2008). Abducted by a UFO: Prevalence information affects children's false memories for an implausible event. *Applied Cognitive Psychology*, 23, 115-125.
- Otgaar, H., Schell-Leugers, J.M., Howe, M.L., De La Fuente Vilar, A., Houben, S.T.L. & Merckelbach, H. (2021). The link between suggestibility, compliance, and false confessions: A review using experimental and field studies. *Applied Cognitive Psychology*, 35(2), 445-455.
- Patihis, L., Frenda, S.J., & Loftus, E.F. (2018). False memory tasks do not reliably predict other false memories. *Psychology of Consciousness: Theory, Research, and Practice*, 5, 140-160.
- Patihis, L. & Loftus, E. F. (2016). Crashing memory 2.0: False memories in adults for an upsetting childhood event. *Applied Cognitive Psychology*, 30, 41-50.
- Patrick, C.J. & Iacono, W.G. (1989). Psychopathy, threat, and polygraph test accuracy. *Journal of Applied Psychology*, 74, 347-355.
- Paulhus, D. L., Buckels, E. E., Trapnell, P. D., & Jones, D. N. (2021). Screening for Dark Personalities: The Short

- Dark Tetrad (SD4). *European Journal of Psychological Assessment*, 37(3), 208–222.
- Pavlidis, I., Eberhardt, N.L. & Levine, J.A. (2002). Seeing through the face of deception: Thermal imaging offers a promising hands-off approach to mass security screening. *Nature*, 415, 35.
- Pennington, N. & Hastie, R. (1992). Explaining the evidence: Test of the story model for juror decision making. *Journal of Personality and Social Psychology*, 62, 189-206.
- Perfect, T.J., Wagstaff, G.F., Moore, D., Andrews, B., Cleveland, V., Newcombe, S., Brisbane, K.A. & Brown, L. (2008). How can we help witnesses to remember more? It's an (eyes) open and shut case. *Law and Human Behavior*, 32, 314-324.
- Perillo, J. T. & Kassin, S. M. (2011). Inside interrogation: The lie, the bluff, and false confessions, *Law and Human Behavior*, 35, 327-337.
- Pimentel, P. S., Arndorfer, A., Malloy, L. C. (2015). Taking the blame for someone else's wrongdoing: The effects of age and reciprocity. *Law and Human Behavior*, 39(3). 219-231.
- Pinizzotto, A.J. (1989). Memory and hypnosis: Implications for the use of forensic hypnosis. *Professional Psychology: Research and Practice*, 20, 322-328.
- Ponds, R., De Lugt, M., Verhey, F. & Jolles, J. (1995). Malingeren bij neuropsychologisch onderzoek. *De Psycholoog*, 30, 357-362.
- Poole, D.A., Bruck, M., & Pipe, M.E. (2011). Forensic interviewing aids: Do props help children answer questions about touching? *Current Directions In Psychological Science*, 20, 11-15.
- Popper, K. (1963). *Conjectures and refutations*. London: Routledge.
- Popper, K. (1959). *The logic of scientific discovery*. Routledge: London.

- Posthumus, F. (2005). *Evaluatieonderzoek in de Schiedammer parkmoord*. Den Haag: Openbaar ministerie.
- Premack, D. & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *The Behavioral and Brain Sciences*, 4, 515-526.
- Principe, G.F. & Ceci, S.J. (2002). "I saw it with my own ears": The effects of peer conversations on preschoolers' reports of nonexperienced events. *Journal of Experimental Child Psychology*, 83, 1-25.
- Principe, G.F., Cherson, M., DiPuppo, J. & Schindewolf, E. (2012). Children's natural conversations following exposure to a rumor: Linkage to later false reports. *Journal of Experimental Child Psychology*, 113, 383-400.
- Principe, G.F., Daley, L. & Kauth, K. (2010). Social processes affecting the mnemonic consequences of rumors on children's memory. *Journal of Experimental Child Psychology*, 107, 479-493.
- Principe, G.F., DiPuppo, J. & Gammel, J. (2013). Effects of mothers' conversation style and receipt of misinformation on children's event reports. *Cognitive development*, 28, 260-271.
- Principe, G.F., Kanaya, T., Ceci, S.J. & Singh, M. (2006). Believing is seeing: How rumors can engender false memories in preschoolers. *Psychological Science*, 17, 243-248.
- Principe, G.F. & Schindewolf, E. (2012). Natural conversations as a source of false memories in children: Implications for the testimony of young witnesses. *Developmental Review*, 32, 205-223.
- Principe, G.F., Tinguely, A. & Dobkowski, N. (2007). Mixing memories: The effects of rumors that conflict with children's experiences. *Journal of Experimental Child Psychology*, 98, 1-19.
- Rassin, E. (2007). *Waarom ik altijd gelijk heb: Over tunnelvisie*. Schiedam: Scriptum.

- Rassin, E. (2009). *Tussen sofa en toga: Een inleiding in de rechtspsychologie*. Tweede druk. Den Haag: Boom juridische uitgevers.
- Rassin, E. (2018). Fundamental failure to think logically about scientific questions: An illustration of tunnel vision with the application of Wason's Card Selection Test to criminal evidence. *Applied Cognitive Psychology*, 32, 506-511.
- Rassin, E. (2022). Suggested False Memories of a Non-Existent Film: Forensically Relevant Individual Differences in the *Crashing Memories* Paradigm. *Memory*, 30(9), 1205-1211.
- Rassin, E. (2023). *Case Studies in Legal Psychology: Psychological Perspectives on Criminal Evidence*. London: Springer. <https://doi.org/10.1007/978-3-031-44931-4>
- Rassin, E., Arbiyah, N., Boskovic, I., Otgaar, H., & Merckelbach, H. (2022). Likelihood Ratios in Psychological Expert Opinion, and Their Reception by Professional Judges. *International Journal of Evidence and Proof*, 26(4), 325-341.
- Rassin, E., De Roos, M. & Van Dongen, J. (2024). Dark personality traits and deception, and the Short Dark Tetrad (SD4) as integrity screening instrument. *Scientific Reports*, 14, 311. Doi.org/10.1038/s41598-023-50968-7
- Rassin, E. & Merckelbach, H. (1999). The potential conflict between clinical and judicial decision making heuristics. *Behavioral Sciences and the Law*, 17, 237-248.
- Rassin, E., Merckelbach, H., & Spaan, V. (2001). When dreams become a royal road to confusion: Realistic dreams, dissociation, and fantasy proneness. *Journal of Nervous and Mental Disease*, 189, 478-481.
- Rassin, E. & Van der Heiden, S. (2006). Swearing voices: An experimental investigation of the suppression of hostile hallucinations. *Behavioural and Cognitive Psychotherapy*, 35, 355-360.

- Rassin, E. & Verschuere, B. (2025). Getraumatiseerde getuigen vergissen zich niet: Misverstanden bij het bepalen van de betrouwbaarheid van getuigenverklaringen. *Expertise en Recht* (6), 161-166.
- Redlich, A. D. & Goodman, G. S. (2003). Taking responsibility for an act not committed: The influence of age and suggestibility. *Law and Human Behavior*, 27, 141-156.
- Reinders, A.A.T.S., Nijenhuis, E.R.S., Paans, A.M.J., Korf, J., Willemsen, A.T.M., & den Boer, J.A. (2003). One brain, two selves. *Neuroimage*, 20, 2119-2125.
- Remijn, C.A.C., & Crombag, H.F.M. (2007). Heuristics in causal reasoning and their influence on eyewitness testimony. *Psychology, Crime and Law*, 13, 201-211.
- Rensink, R.A. (2002). Change detection. *Annual Review of Psychology*, 53, 245-277.
- Richards, P.M., Geiger, J.A., & Tussey, C.M. (2015). The dirty dozen: 12 sources of bias in forensic neuropsychology with ways to mitigate. *Psychological Injury and Law*, 8, 265-280.
- Robins, R.W., Gosling, S.D. & Craik, K.H. (1999). An empirical analysis of trends in psychology. *American Psychologist*, 54, 117-128.
- Roediger, H.L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 24, 803-814.
- Rosenthal, R. (1979). The 'file drawer problem' and tolerance for null results. *Psychological Bulletin*, 86, 638-641.
- Royall, R. (1997). *Statistical evidence: A likelihood paradigm*. London: Chapman & Hall.
- Ruby, C.L. & Brigham, J.C. (1997). The usefulness of the criteria-based content analysis technique in distinguishing between truthful and fabricated allegations: A critical review. *Psychology, Public Policy, and Law*, 3, 705-737.

- Russano, M.B., Meissner, C.A., Narchet, F.M. & Kassin, S.M. (2005). Investigating true and false confessions within a novel experimental paradigm. *Psychological Science*, 16, 481-486.
- Saks, M.J. & Koehler, J.J. (2005). The coming paradigm shift in forensic identification science. *Science*, 309, 892-895.
- Schacter, D.L. (1999). The seven sins of memory. *American Psychologist*, 54, 182-203.
- Scherr, K.C., Redlicht, A.D. & Kassin, S.M. (2020). Cumulative disadvantage: A psychological framework for understanding how innocence can lead to confession, wrongful conviction, and beyond. *Perspectives on Psychological Science*, 15(2), 353-383.
- Schmidt, H.G., Peeck, V.H., Paas, F. & Van Breukelen, G.J.P. (2000). Remembering the street names of one's childhood neighbourhood: A study of very long-term retention. *Memory*, 8, 37-49.
- Schünemann, B. & Bandilla, W. (1989). Perseverance in courtroom decisions. In H. Wegener, F. Lösel & J. Haisch (red.). *Criminal behavior and the justice system: Psychological perspectives* (pp. 181-192). New York: Springer.
- Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, 54, 93-105.
- Schwarz, N., Knäuper, B., Hippler, H.J., Noelle-Neumann, E. & Clark, F. (1991). Rating scales: Numeric values may change the meaning of scale labels. *Public Opinion Quarterly*, 55, 570-582.
- Scoboria, A., Wade, K. A., Lindsay, D. S., Azad, T., Strange, D., Ost, J. & Hyman, I. E. (2017). A mega-analysis of memory reports from eight peer-reviewed false memory implantation studies. *Memory*, 25(2), 146-163.
- Shuman, D.W., Greenberg, S., Heilbrun, K., & Foote, W.E. (1998). An immodest proposal: should treating mental health professionals be barred from testifying about

- their patients? *Behavioral Science and the Law*, 16, 509-523.
- Simmons, J.P., Nelson, L.D. & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22, 1359-1366.
- Simon, D. (2004). A third view of the black box: Cognitive coherence in legal decision making. *The University of Chicago Law Review*, 71, 511-586.
- Simons, D.J., & Chabris, C.F. (1999). Gorilla's in our midst: Sustained inattentive blindness for dynamic events. *Perception*, 28, 1059-1074.
- Simons, D.J. & Levin, D.T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin & Review*, 4, 644-649.
- Sjoden, B., Granhag, P. A., Ost, J., & af Hjelmsater, E. R. (2009). Is the truth in the details? Extended narratives help distinguishing false "memories" from false "reports". *Scandinavian Journal of Psychology*, 50, 203-210.
- Smeets, T., Jelicic, M., Peters, M. J. V., Candel, I., Horselenberg, R., & Merckelbach, H. (2006). 'Of course I remember seeing that film': How ambiguous questions generate crashing memories. *Applied Cognitive Psychology*, 20, 779-789. DOI: 10.1002/acp.1205
- Smeets, T., Telgen, S., Ost, J., Jelicic, M., & Merckelbach, H. (2009). What's behind crashing memories? Plausibility, belief and memory in reports of having seen non-existent images. *Applied Cognitive Psychology*, 23, 1333-1341. DOI: 10.1002/acp.1544
- Smith, G.P. & Burger, G.K. (1997). Detection of malingering: Validation of the structured inventory of malingered symptomatology (SIMS). *Journal of the Academy of Psychiatry and the Law*, 25, 183-190.
- Smith, S.M., Stinson, V. & Patry, M.W. (2009). Using the "Mr. Big" technique to elicit confessions: Successful

- innovation or dangerous development in the Canadian legal system? *Psychology, Public Policy, and Law*, 15, 169-193.
- Sokal, A (1996b). Revelation: A physicist experiment with cultural studies. *Lingua Franca*, 62-64.
- Sokal, A. (1996a). Transgressing the boundaries: Toward a transformative hermeneutics of quantum gravity. *Social Text*, 46/47, 217-252.
- Spinoza, B. (1677/2009). *Ethica*. Amsterdam: Bert Bakker.
- Sporer, S.L., Penrod, S.D., Read, J.D. & Cutler, B.L. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence-accuracy relation in eyewitness identification studies. *Psychological Bulletin*, 118, 315-327.
- Stebay, N.M. (1992). A meta-analytic review of the weapon focus effect. *Law and Human Behavior*, 16, 413-424.
- Stebay, N., Dysart, J., Fulero, S. & Lindsay, R.C.L. (2003). Eyewitness accuracy rates in police showup and lineup presentations: A meta-analytic comparison. *Law and Human Behavior*, 27, 523-540.
- Stebay, N.K., Dysart, J.E. & Wells, G.L. (2011). Seventy-two tests of the sequential lineup superiority effect: A meta-analysis and policy discussion. *Psychology, Public Policy, and Law*, 17, 99-139.
- Steller, M. & Köhnken, G. (1989). Criteria-based statement analysis: Credibility assessment of children's statements in sexual abuse cases. In D.C. Raskin (red.). *Psychological Methods for Investigating Evidence* (pp. 217-245). New York: Springer.
- Sternberg, R.J. (1996). *Cognitive psychology*. London: Harcourt Brace.
- Strack, F., Schwarz, N. & Gschneidinger, E. (1985). Happiness and reminiscing: The role of time perspective, mood, and mode of thinking. *Journal of Personality and Social Psychology*, 49, 1460-1469.
- Suengas, A.G. & Johnson, M.K. (1988). Qualitative effects of rehearsal on memories for perceived and imagined

- complex events. *Journal of Experimental Psychology: General*, 117, 377-389.
- Thomas, D. (1997). Hidden messages and the Bible code. *Skeptical Inquirer*, 21(6).
- Thomas, A.K. & Loftus, E.F. (2002). Creating bizarre false memories through imagination. *Memory & Cognition*, 30, 423-431.
- Thompson, W.C., Vuille, J., Taroni, F., & Biedermann, A. (2018). After uniqueness: The evolution of forensic science. *Judicature*, 102, 18-27.
- Tredoux, C. (1998). Statistical inference on measures of lineup fairness. *Law and Human Behavior*, 22, 217-237.
- Tversky, A. & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90, 293-315.
- Tversky, B. & Marsh, E.J. (2000). Biased retellings of events yield biased memories. *Cognitive Psychology*, 40, 1-38.
- Undeutsch, U. (1967). Beurteilung der Glaubhaftigkeit von Aussagen. In: U. Undeutsch (red.). *Handbuch der Psychologie*, 11: *Forensische Psychologie* (pp. 26-181). Göttingen: Hogrefe.
- Van Amelsvoort, A. (2023). *Handleiding confrontatie*. Den Haag: SDU.
- Van Amelsvoort, A. & Rispens, I. (2025). *Handleiding verhoor*. Den Haag: SDU.
- Van de Ven, V. & Merckelbach, H. (2003). The role of schizotypy, mental imagery, and fantasy proneness in hallucinatory reports of undergraduate students. *Personality and Individual Differences*, 35, 889-896.
- Van den Bersselaar, V. (2007). *Wetenschapsfilosofie in veelvoud: Fundamenten voor onderzoek en professioneel handelen*. Bussum: Coutinho.
- Van den Hout, M. & Kindt, M. (2003). Repeated checking causes memory distrust. *Behaviour Research and Therapy*, 41, 301-316.

- Van Lommel, P., Van Wees, R., Meyers, V. & Elfferich, I. (2001). Near-death experience in survivors of cardiac arrest: A prospective study in the Netherlands. *The Lancet*, 358, 2039-2045.
- Verhey, B., Bex, F., Timmer, S. J., Vlek, C. S., Meyer, J., Renooij, S., & Prakken, H. (2016). Arguments, scenarios and probabilities: Connections between three normative frameworks for evidential reasoning. *Law, Probability and Risk*, 15, 35-70.
- Verhoeven, W.J. & Stevens, L. (2012). The lawyer in the Dutch interrogation room: Influence on police and suspect. *Journal of Investigative Psychology and Offender profiling*, 9, 69-92.
- Verschuere, B., Kleinberg, B., Van Kolfchooten, M., Bolhoven L. & Rassin, E. (2025). How Dutch legal professionals assess statement credibility: Evidence from a survey and an analysis of 518 court rulings on sexual abuse. *Memory*, DOI: 10.1080/09658211.2025.2592947
- Verschuere, B., Lin, C.C., Huismann, S., Kleinberg, B., Willemse, M., Chong Jia Mei, E., Van Goor, T., Löwy, L. H. S., Appiah, O. K. & Meijer, E. (2023). The use-the-best heuristic facilitates deception detection. *Nature Human Behaviour*, 7, 718-728.
- Volbert, R. & Steller, M. (2014). Is this testimony truthful, fabricated, or based in false memory? Credibility assessment 25 years after Steller and Köhnken (1989). *European Psychologist*, 19, 207-220.
- Vrij, A. (2004). Why professionals fail to catch liars and how they can improve. *Legal and Criminological Psychology*, 9, 159-181.
- Vrij, A. (2005). Criteria-based content analysis: A Qualitative review of the first 37 studies. *Psychology, Public Policy, and Law*, 11, 3-41.
- Vrij, A. (2016). Baseline as a lie detection method. *Applied Cognitive Psychology*, 30, 1112-1119.

- Vrij, A., Fisher, R.P., & Blank, H. (2017). A cognitive approach to lie detection: A meta-analysis. *Legal and Criminological Psychology*, 22, 1-21.
- Vrij, A. & Mann, S. (2001). Telling and detecting lies in a high-stake situation: The case of a convicted murderer. *Applied Cognitive Psychology*, 15, 187-203.
- Vrij, A., Meissner, C.A., Fisher, R.P., Kassin, S.M., Morgan, C.A. & Kleinman, S.M. (2017). Psychological perspectives on interrogation. *Perspectives on Psychological Science*, 12, 927-955.
- Vul, E., Harris, C., Winkielman, P. & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Psychological Science*, 4, 274-290.
- Wagenaar, W.A. (1977). *De beste stuurlui dempen de put: Hoe mensen beslissen toegelicht aan de hand van spreekwoorden*. Baarn: Ambo.
- Wagenaar, W. A. (1988). The proper seat: A Bayesian discussion of the position of expert witnesses. *Law and Human Behavior*, 12(4), 499-510.
- Wagenaar, W.A. (1989). *Het herkennen van Iwan: De identificatie van de dader door ooggetuigen van een misdrijf*. Amsterdam: Swets & Zeitlinger.
- Wagenaar, W.A. (1998). De vraag aan de deskundige. In R.A.R. Bullens (red.). *Getuige deskundigen in zedenzaken: De positie van de gedragswetenschapper bij strafzaken rondom mogelijk seksueel misbruik van kinderen* (pp. 47-59). Leiden: DSWO press.
- Wagenaar, W.A. & Groeneweg, J. (1990). The memory of concentration camp survivors. *Applied Cognitive Psychology*, 4, 77-87.
- Wagenaar, W.A. & Van der Schrier, J.H. (1996). Face recognition as a function of distance and illumination: A practical tool for use in the courtroom. *Psychology, Crime & Law*, 2, 321-332.
- Wason, P.C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, 20, 273-281.

- Wegner, D.M., Schneider, D.J., Carter, S.R., & White, T.L. (1987). Paradoxical effects of thought suppression. *Journal of Personality and Social Psychology*, 53, 5-13.
- Wegner, D.M., Wenzlaff, R., Kerker, R.M., & Beattie, A.E. (1981). Incrimination through innuendo: Can media questions become public answers? *Journal of Personality and Social Psychology*, 40, 822-832.
- Weisberg, D.S., Keil, F.C., Goodstein, J., Rawson, E. & Gray, J.R. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience*, 20, 470-477.
- Wells, G.L., Bull Kovera, M., Douglass, A.B., Brewer, N., Meissner, C.A., & Wixted, J.T. (2020). Policy and procedure recommendations for the collection and preservation of eyewitness identification evidence. *Law and Human Behavior*, 44, 3-36.
- Wells, G.L. & Lindsay, R.C.L. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin*, 88, 776-784.
- Wells, G. L. & Loftus, C. A. E. (1990). Police lineups as experiments: Social Methodology as a framework for properly conducted lineups. *Personality and Social Psychology Bulletin*, 16(1), 106-117.
- Wells, G.L., & Luus, E. (1990). Police lineups as experiments: Social methodology as a framework for properly-conducted lineups. *Personality and Social Psychology Bulletin*, 16, 106-117.
- Wells, G.L. & Olson, E.A. (2002). Eyewitness identification: Information gain from incriminating and exonerating behaviours. *Journal of Experimental Psychology: Applied*, 8, 155-167.
- Wigmore, J.H. (1931). *The principles of judicial proof or the process of proof as given by logic, psychology, and general experience, and illustrated in judicial trials*. Boston: Little Brown.

- Willcock, E., Morgan, K. & Hayne, H. (2006). Body maps do not facilitate children's reports of touch. *Applied Cognitive Psychology*, 20, 607-615.
- Wilson, K. & French, C. C. (2006). The relationship between susceptibility to false memories, dissociativity, and paranormal belief and experience. *Personality and Individual Differences*, 41, 1493-1502.
- Winkler, J.D., Kanouse, D.E. & Ware, J.E. (1982). Controlling for acquiescence response set in scale development. *Journal of Applied Psychology*, 67, 555-561.
- Wise, R.A., & Safer, M.A. (2004). What US judges know and believe about eyewitness testimony. *Applied Cognitive Psychology*, 18, 427-443.
- Wiseman, R. (1995). The megalab truth test. *Nature*, 373, 391.
- Witztum, D., Rips, E. & Rosenberg, Y. (1994). Equidistant letter sequences in the book of Genesis. *Statistical Science*, 9, 429-438.
- Wixted, J.T. & Wells, G.L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest*, 18, 10-65.
- Zajac, R., Garry, M., London, K., Goodyear-Smith, F., Hayne, H. (2013). Misconceptions about childhood sexual abuse and child witnesses: Implications for psychological experts in the courtroom. *Memory*, 21, 608-617.

