Jan J.F. ter Laak
Faculteit Sociale Wetenschappen
Afdeling Ontwikkelingspsychologie
Universiteit Utrecht

Psychologisch diagnosticeren onder vuur

Wat hebben statistici, behandelaars, psychometrici en het publiek op de diagnose van uw vraag of probleem aan te merken en hebben ze een punt?

'Heer, hoe talrijk zijn mijn belagers, velen vallen mij aan, velen zeggen van mij: God zal hem niet redden' (psalm no. 3, vers 1).

'*Een troep stieren staat om mij heen, buffels van Basam omsingelen mij, roofzuchtige, brullende leeuwen sperren hun muil naar mij open'* (psalm no. 22 vers 5).

De Bijbel: Psalmen (2004, pp. 835-1023). Auteurscollectief. Querido Jongbloed.

Voor Marie-Louise, Mose en Ebba

Psychologisch diagnosticeren onder vuur

Wat hebben statistici, behandelaars, psychometrici en het publiek op de diagnose van uw vraag of probleem aan te merken en hebben ze een punt?

Inhoud

Voorwoord

Er is geen vak of wetenschappelijke discipline of er is wel iets op aan te merken. Het is zelfs een hoog in de hiërarchie staande onderwijsdoelstelling, dat je kritiek kunt leveren op een wetenschappelijke tekst. Bij de diagnose en het diagnosticeren is dat niet anders. De kritiek liegt er ook niet om. De statistici hebben het klinisch oordeel buitenspel willen zetten. De controverse is een verwarrende omdat ze op verschillende lagen gespeeld wordt. Om het met een metafoor te zeggen: voetballen ze wel op hetzelfde veld? Hebben ze het niet over verschillende zaken? De practicus wordt vooral geraadpleegd bij niet gewenste gedragingen. Zijn eerste reactie is: wat kan ik eraan doen. Hij is met andere woorden op behandelen uit. Heeft hij de diagnostiek dan nog wel nodig? Kies de beste behandeling en laat de diagnose achterwege! Heeft de behandelaar een punt? En wat te zeggen van de diagnose-behandel-combinaties DBCs)? Is dat de toekomst? Diagnostici gebruiken vaak tests en vragenlijsten. Die worden gescoord en de klassieke en moderne testtheorie gaan onder meer over het scoren van items en tests. Ze schrijven tevens de interpretatie van testresultaten voor. Is de diagnosticus daarmee gebaat of beknelt de testtheorie hem in zijn diagnose van zijn cliënt? Het (niet) weldenkende publiek heeft kritiek op psychologie en diagnostiek. Is dat terecht? Is de diagnosticus duidelijk genoeg over wat hij kan?

Ten geleide

Elk geschrift bevat vooringenomenheden, willekeurige onderscheidingen en onvolledig gearticuleerde epistemologische en theoretisch inhoudelijke standpunten. Het maakt gebruik van ander werk en van vakgenoten als inspiratiebronnen. Een auteur ontkomt niet aan vooroordelen en preoccupaties over wetenschap, haar toepassing en beoefenaren. Een auteur is niet objectief en niet neutraal. Als een onderwerp hem niet interesseert of niets oplevert, leest en schrijft hij er niet over. Er wordt geen poging gedaan om deze persoonlijke bagage of ballast te verbergen. Als u deze prolegomena, reflectie of ruminatie vooraf liever mijdt, kunt u beginnen bij **Leeswijzer**, p. 11.

Vooringenomenheden ten opzichte van diagnostiek en psychologie Wetenschappelijke arbeid is bedoeld om de werkelijkheid binnen en buiten de mens - het Ik en het Niet-Ik - te beschrijven, verklaren, controleren en voorspellen. Wat tot het Ik gerekend wordt wisselt in de psychologie: leer-, ontwikkelings- en informatie verwerkende mechanismen worden soms vervangen door biologische en neurologische. Hier wordt gekozen om gedrag psychologisch te beschrijven en te verklaren. Van gedrag zijn we voor een deel de eigenaar en dat zijn we niet direct van neurologische en biologische mechanismen.

Sociale wetenschap heeft met concurrenten van doen (Westbroek, 2013) in de vorm van systemen die beloven vragen te beantwoorden en behoeftes te vervullen: ideologieën, religies, heersende opvattingen, reclame, lobbyisten, enzovoort. De psychologie is vooralsnog geen geslaagd systeem. De resultaten van studies en de successen van behandelaars geven daar onvoldoende reden toe.

De opbrengst van de diagnostiek wordt geridiculiseerd door universitaire psychologen (werk van doctorandi) en hulpverleners (wat is uw meerwaarde; het is een mager halffabricaat). Beider houding belemmert het zicht op wat diagnostiek kan en is.
Diagnostiek gaat over het gedrag van de cliënt. Het is geen biologisch, medisch of natuurwetenschappelijk onderzoek. Daar heeft een diagnosticus geen verstand van en het zet hem op het verkeerde. Dat denken en onderzoek kan slechts metaforen bieden om gedrag te beschrijven.

Psychologie is zelf betrokken. Het betekent dat *psychologische diagnostiek tegelijkertijd diagnostiek van de psychologie* is: het kennisbestand van de psychologie en de activiteiten van diagnosticus en theoreticus/onderzoeker blijven niet buiten schot.
Het willen kennen en onderzoeken van het gedrag van de persoon vindt zijn oorsprong in de rationele en emotionele uitrusting van de mens. Rationaliteit leidt tot theoretische constructen en empirische toetsing. Emotionaliteit leidt tot verlangen naar inzicht, soms tot hang naar een veilig geloof, maar ook tot ongerustheid en onzekerheid over wat bereikt is of kan worden. Beide zijn nodig voor dynamiek en balans.

Wetenschappers en practici hebben nier de volledige toegang tot de fysisch en sociale werkelijkheid. Kennis verschuift de horizon, verlegt de grens, maar het einde, dat wil zeggen volledige kennis van objecten, verschijnselen en subjecten bestaat niet. Ze komt ook niet tot stand door integratiepogingen binnen en tussen wetenschappen.

Elke activiteit - dus ook kennen, leren en denken - is ingebed in het alledaagse leven. Onze geleefde ervaring is de basis van onze kennis met inbegrip van onszelf en mensen om ons heen: Merleau-Ponty (1945/1957) en zijn opvolger Piaget.

Wetenschap wordt door elites soms gebruikt om hun positie te behouden en privileges te beschermen. Het publiek wordt op afstand gezet terwijl het ziet dat wetenschappers het zelden eens zijn en er zich onder hen fraudeurs bevinden (zie Van Kolfschoten, 2012). Van de weeromstuit wordt in een 2015 initiatief elke weldenkende burger gevraagd thema's en vragen te noemen waar onderzoekers zich over gaan buigen en antwoorden op geven.

Discussie op grond van argumenten is een onvermijdelijk onderdeel van wetenschappelijk werk. Als de psychologie een wetenschap is - en daar ben ik niet zo zeker van - hoort ook daar een dergelijke discussie thuis. Wetenschappers kunnen ontsporen door zich in te stellen op het overwinnen van anderen en het behalen van succes door bijzaken-barok, dat wil zeggen door zich 'met overmaat aan geleerdheid en sluwheid te richten op kleine gebieden of maatschappelijke thema's, daarbij een vaardigheid ten toon spreidend, die weinigen hen zullen verbeteren en waar geldig en vals en goed en *tactische* begrippen zijn' (vrij naar Menno ter Braak, cursief, jtl).
Het 19 de -eeuwse romantisch beeld van de wetenschapper is dat van de uitzonderlijk begaafde of van de briljante psycholoog die door je heen kijkt. Het cliché van transpiratie (95%) en inspiratie (5%) komt dichter bij wat er gebeurt in onderzoek en praktijk: respectievelijk het routineus uitvoeren van een gefinancierd project (*output*) plus het opschrijven volgens richtlijnen van het *Publication Manual* van de APA en het rapporteren over de cliënt volgens een protocol.

Wetenschappers zijn conservatief. Wat ze weten is kostbaar en moeizaam verworven. Een succesparadigma wordt niet gauw opgegeven. De VU wiskundige Meester (2014) voegt er aan toe dat sommige arrogant zijn en hun kennis en weten overschatten. Ze zien verschijnselen vanuit een tunnel duiden en duiden afwijkingen van hun tunnelvisie als onwetenschappelijk. Het ging daarbij over onzekerheid in de evolutieleer.

Psychologen vertellen wat het publiek en tijdschriftredacties willen horen en vermijden te zeggen wat men niet wil horen. Gemakkelijk bij het publiek en redacties liggende thema's worden uitgewerkt, bijvoorbeeld man-vrouw verschillen, agressie, positieve psychologie, persoonlijke groei en ontwikkeling, genetische basis van gedrag, het lijden dat zich overal

voordoet en de rol van hersenen voor begaafdheid en pathologie. Als onderzoeksgegevens afwijken van wat *Dass Man* - van het publiek tot de bestuurlijke elite vindt - valt er iets uit te leggen. Succes is niet verzekerd.
Psychologen kunnen het publiek en de bestuurlijke elite ook een spiegel voorhouden en aan de hand van gegevens tonen dat ze niet het centrum van de wereld zijn, geen *tabula rasa*, niet rationeel of irrationeel, niet vrij of onvrij, niet gelijk of ongelijk, niet transparant voor elkaar, maar ook geen vreemden en niet onafhankelijk of afhankelijk.
Psychologists go, where the money goes Dit uit zich in aansluiting bij modieuze thematieken, goed in het gehoor liggende theorievorming en methodologie. Een *kitchen and sink* typologie van psychologen met mijn voorkeur voor driedelingen is: de Opportunist: steeds een nieuw thema kiezen dat het goed doet: van geleerde hulpeloosheid naar positieve psychologie, de Monopolist: één eenvoudig populair thema kiezen en dat monopoliseren: bijvoorbeeld de *Big Five*, *Attachment*, Adoptie en Echtscheiding en de Bekeerling: van psychometricus naar persoonlijkheidspsycholoog of vice versa. Een bekering heeft iets moedigs vooral als je er niet mee op de TV komt.

Het is het moeilijk uit te maken of meta-analyses en meta-analyses van meta-analyses iets nieuws brengen of een eind maken aan eenzelfde type onderzoek, c.q. een vraag afdoende beantwoorden of een probleem oplossen. Wellicht is het slechts een superieure vorm van boekhouden. Is dit het westers ultieme geloof in de Zelfopenbaring van het Zijn? Meta-analyses, *big data* en *data mining* zijn niettemin zinvol, al steunen ze op de empiristische veronderstelling: hoe meer onderzoek, hoe beter. Empirisch onderzoek ontnuchtert en matigt hoge verwachtingen en onderzoek kan niet zonder een gedisciplineerd geloof (Heideggers *Seinsglaube*). Dat wil zeggen, het kan niet buiten de empirie om bedachte theorieën, structuren en modellen, gecreëerde betekenissen in de hermeneutiek en analyse van vooronderstellingen over wat de wetenschap en de mens zijn en bewerkstelligen.

Er is een pikorde in de wetenschap. Deze zet de psychologie op afstand van de fysica, scheikunde, neurologie en biologie. Sociologie, antropologie, taalkunde en wijsbegeerte worden er onder geplaatst (Simonton, 2004, 2009). Binnen de psychologie is er eveneens een pikorde met functieleer aan de top en ontwikkelings- en klinische psychologie onderaan. Deze pikordes verhinderen om van een andere (sub)discipline te leren en te profiteren. Dat gebeurt dan ook zelden of nooit.

Onderscheidingen en standpunten Om de diagnostiek van de cliënt te structuren worden drie niet willekeurige bronnen van theorie onderscheiden: impliciete alledaagse, expliciete theorievorming in onze tekstboeken: de heersende paradigma's van Kuhn (1962) en alternatieven voor expliciete theorievorming. Deze worden geordend in drie oriëntaties: individuele verschillen, ontwikkeling en fysische en sociale context. De drie dekken bijna alle

theorieën in de tekstboeken. Twee bij twee gecombineerd leveren ze de bekende tweedelingen van Persoon-Situatie, Organismisch-Mechanistisch en Aanleg-Omgeving op.

Methoden, statistiek en protocollen worden benut met als oogmerk na te gaan wat ze de diagnosticus bieden om de vraag van de cliënt te analyseren. Wat leveren klassieke en moderne testtheorie, uitkomsten van studies over het klinisch oordeel en statistische modellen en regels op voor het diagnostisch handwerk, wat leren *heuristics* (heuristieken) en *biases* (vertekeningen), de *multi-attribute-utility-theory*, het hypothese toetsend model (HTM) en de Bayesiaanse regel de diagnosticus? Hun intrinsieke kwaliteit wordt niet betwist, wel de pretentie om het diagnosticeren van het gedrag van de cliënt afdoende te regelen.

Beschrijven, voorspellen, verklaren/controleren en beslissen worden als doelen van diagnostiek geoormerkt. Hun succes hangt af van toegankelijkheid en helderheid van conceptuele analyses en empirische bevindingen, onder meer resultaten van meta-analyses en van het gezonde verstand, want sommige analyses vragen naar de bekende weg en andere zijn pseudo-empirisch (Smedslund, 2009).
Beschrijven is naast categoriseren - bijvoorbeeld zoals Linnaeus dat met planten deed en de DSM – ook ontwerpen van een wiskundig geformuleerd model dat uit zijn aard intrinsiek klopt. Dat kan een doel op zich zijn. Deze voorkeur is terug te vinden bij psychometrici met hun modellen, functies en formules en bij structuralisten, waar Piaget een voorbeeld van is. Van vertegenwoordigers van beide werk- en denkwijzen wordt verwacht dat ze een brug naar observeerbaar gedrag slaan. Als dat niet zichtbaar gemaakt wordt is het elegant, maar zonder betekenis voor diagnostiek.

Verklaren gaat terug op de Griekse Pre-Socratici en Plato en Aristoteles. Het ging hen om een verklarend principe (oorsprong, *archè*), zodat alles opgenomen kon worden in een overkoepelend geheel of uit één bron oplichtte. Zo worden uiteenlopende zaken verklaard. Aristoteles wilde bijvoorbeeld met zijn principe van de gulden middenweg zowel evenwichtstoestanden bij natuurverschijnselen (harmonie der sferen) als politiek (zijn vorm van stadsstaat democratie) en ethiek (hoe het goede leven: *eudamonia* bereiken) verklaren. Bij diagnostiek als een doelgerichte activiteit komen beide als een *hybride* aan de orde: diagnostische kennis bestaat voor een deel uit categoriesystemen, structuren en formules én berust op empirische feiten en ervaringskennis (inter- en supervisie) én op diagnostische praktijken als observatie, experimentjes om gedrag te ontlokken aan de cliënt en het gebruik van een valide instrumentarium.

De grenzen van voorspellen, verklaren/controleren en beslissen van en over gedragingen tekenen zich af als onderzoek voortgaat op de wijze waarop het nu vormgegeven is en gepubliceerd wordt. Dit standpunt wordt gemotiveerd en geïllustreerd met resultaten van

meta-studies. En die bieden r- en d- waarden die niet verder komen dan significante samenhangen en effecten en dan verklaarde variantie van maximaal 50%. Ze bieden met andere woorden *halve waarheden*.
Gedragingen en sociale contexten zijn dynamisch en variabel. Bij het verkennen van grenzen van de impact van oorzaken en de hechtheid van samenhangen wordt nadruk gelegd op effectgroottes, betrouwbaarheidsintervallen, *power* van toetsen, *eye-balling* en gezond verstand.

De psychologie is rijk aan tweedelingen. Er wordt niet gekozen voor de een of de ander. Ze zijn gefabriceerd en uitgevonden, zoals lichaam versus geest (*mind-body*), theorie versus meten, persoon versus situatie, stabiliteit versus verandering, natuur versus cultuur, links versus rechts en genen versus omgeving.

Er is aarzeling om complexe protocollen, regels en statistische modellen te aanvaarden als veilige en vooral afdoende hulpen bij het diagnosticeren van de cliënt. Dit verwijst niet zozeer naar Piet Vroons slagzin: hoe meer regels hoe meer vlegels, als wel naar het feit dat ze hun doel voorbij kunnen schieten, gegeven de beperkte informatieverwerkingscapaciteit en de natuurlijke werkwijze van de diagnosticus.
Er is terughoudendheid ten opzichte van diagnose-behandeling-combinaties (DBCs). Diagnostiek stelt beperkt in staat tot valide, laat staan foutloos beschrijven, voorspellen en controleren. De practicus zegt regelmatig geen meerwaarde van diagnostiek te zien voor het inzetten van een behandeling. Effecten van zijn behandelingen zijn echter ook beperkt gegeven de r- en d-waarden. Voor kinderen en adolescenten met psychische klachten is er in Nederland al een honderdtal DBCs. Het is onwaarschijnlijk dat ze passen bij 100 verschillende vragen en problemen of bij 100 groepen cliënten of stoornissen.
Er is terughoudendheid om correlaties tussen stimuli en responsen uit experimentele en predictor-criterium studies te benutten als basis voor individuele voorspellingen, verklaringen, beslissingen en voor diagnose-behandeling-combinaties.

Deze tekst bevat geen recepten voor afnemen van interviews en tests, het uitvoeren van interventies en het schrijven van psychologische rapporten. Ze is bedoeld als hulp bij een *rijke en voldoende beschrijving van het probleem/de vraag van de cliënt*, zoals Simon dat voorstond en De Groot aanbevolen heeft. Daarnaast is de ecologische oriëntatie van Brunswik inspiratie en leidraad. Het gaat hier om de oude Simon. Hij verdedigde later dat intelligentie louter het manipuleren van abstracte symbolen is en zet de relatie met de biologische *wetware* en psychologische processen, de *software* tussen haakjes of denkt die niet nodig te hebben. De symbolen en structuren zijn bij hem niet metaforisch, maar letterlijk bedoeld.
Elke symbolische en formele structuur in de psychologie kan niet zonder semantische interpretatie (Searle, 1984).

Methodologie heeft een epistemische dimensie nodig om het object van onderzoek, het gedrag van de cliënt, recht te doen. Er wordt naar gestreefd om een relatie tussen model, formule en iets buiten het model of de formule te verduidelijken. Waarheid, geldigheid en objectiviteit gaan over een relatie en dat is niet alleen de consensus tussen experts of die van een forum van weldenkende collega's. Zij sluit een contrafactisch streven in, want die relatie, uitspraak, bewering over gedrag en het waargenomen gedrag is nooit volledig en sluitend gelegd. Als dat het geval was, dan zouden wetenschap en onderzoek al lang voltooid zijn.

Er wordt naar gestreefd om naast aandacht voor individuele verschillen, ontwikkeling en sociale context te bespreken. In diagnostisch onderzoek overheerst de individuele verschillen oriëntatie. Dat betekent dat daar het meeste onderzoek verricht is en gepubliceerd is.

Het onderwerp van diagnostiek is het gedrag van de cliënt. Menselijk gedrag is onderwerp van elke wetenschap, want zonder de mens is er geen wetenschap. De mens en zijn gedragingen zijn bijgevolg intrinsiek een *go-between* tussen alle wetenschappen. Het risico om daarbij buiten je boekje te gaan is groot en wordt voor lief genomen.

Fysica, scheikunde en biologie hebben een hogere status dan de psychologie. De identiteit van psychologen ligt minder vast dan die van fysici en medici. Meedeinend met de tijdgeest en zo nu en dan geconfronteerd met een identiteitscrisis, richt de psycholoog zich soms op andere takken van wetenschap, bijvoorbeeld neuropsychologie en hersenonderzoek: het geloof in het DNA-maals als substituut voor het hierna-maals. En, evenals bij religie gaat het daarbij niet zelden om belangen.
Verder doet zich in een poging gerespecteerd te worden een toewending voor naar beschrijving met behulp van formele en statistische modellen. Wiskunde is immers de exactheid en geldigheid zelf en staat hoog in de wetenschappelijke pikorde. Men zet zich af tegen losjes geformuleerde verklarende semantische theorieën en constructen. De modellen zijn waar, ze kloppen op de wijze van regels van het schaakspel. En, bij een confrontatie met psychologische onderzoeksvragen lopen ze door hun complexiteit soms voor op de realiteit van de beschrijving van concrete gedragingen.

In deze tekst ligt de nadruk op de vraag/het probleem van de cliënt, op psychologische mechanismen en op constructen die zijn gedrag beschrijven, verklaren en voorspellen. Daarbij zijn abstracte modellen welkom, maar hebben geen prioriteit. En, als er al gebruik van gemaakt wordt, blijken tot heden geen betere voorspellingen gedaan worden, denk aan validiteitscoëfficiënten van fysiologische variabelen en neurologische parameters.

Sociologie, culturele antropologie en taalkunde gaan over groepsgedrag en culturele praktijken. Deze liggen niet zover van de psychologie af en formuleren mechanismen die er toe doen om het gedrag van de cliënt te begrijpen en te verklaren.

Een *horizontal world view* wordt aanvaard (Van Dijk & Wilthagen, 2014). Deze steunt op Wittgenstein en James' werk waarin het hogerop zoeken door reductie, zoeken naar de kleinste eenheid en naar verborgen onder- en bovenliggende lagen secundair is aan de concrete beschrijving met behulp van voorbeelden. Wittgensteins opmerking: (1953, sec. 2 xiv): *The existence of experimental method makes us think we have the means of solving problems which trouble us; though problem and method pass one another by* heeft niets van zijn kracht verloren.

Om het gedrag van de cliënt in de concrete situatie voor ogen te houden wordt aangesloten bij Brunswiks representatieve ontwerpen voor onderzoek, bij Rorty's afwijzen van *foundationalism* en bij Smedslunds nadruk op culturele praktijken en de invloed van onmiddellijke situaties.

Onderzoekers en diagnostici mogen vrij analyseren, denken en theoretiseren. De vrijheid is er ook aan de methodologische zijde. Er rust niet bij voorbaat een taboe op intuïtieve, subjectieve of kwalitatieve benaderingen naast de experimentele en correlationele methoden. Dit heeft zin zolang de practicus niet verdwaalt in vertalingen in de vorm van reducties, verborgen mechanismen, producten van neurale activiteiten, herenscans, kleinste eenheden en veronderstelde hogere mentale bewustzijnstoestanden. Dit leidt tot een *Lost in translation*, zoals in de film waar de twee congresgangers, de acteurs Bill Murray en Scarlett Johansson, niets verstaan van de lezingen in het Japans en maar met elkaar gaan praten, zoals studenten bij onze hoorcolleges. De verticale metafoor van het dieper en hoger zoeken, bevat als iedere metafoor - ook de horizontale - het risico *of holding us captive* (Wittgenstein, 1953, Sectie 115).

Een open, eclectische houding past bij diagnostisch werk en omdat de diagnosticus zélf een rol in speelt in de diagnose is inter- en supervisie nodig.

Psychologie is hofleverancier van theorieën, methoden en geldigheidscriteria voor diagnostische uitspraken. Vooruitgang of stagnatie in diagnostiek hangen samen met ontwikkelingen in de psychologie. Is er na Watson, William James, Skinner, Freud, Piaget, Newell Simon, Tversky en Kahneman en in Nederland na Strasser, Linschoten, Kouwer, De Groot en vele anderen iets nieuws te verwachten? Komt er een doorbraak, een andere kijk, die ons dichter brengt bij begrijpen/verklaren van het gedrag van de cliënt of een nieuw en belangrijk facet van zijn gedrag aan het licht brengt? Kan er originaliteit verwacht worden op een gebied dat zo breed is en waar alle mogelijke standpunten, voorkeuren en

temperamenten al lang en vaak beschreven en vertoond zijn en iedere zogenoemde nieuwe gedachte meteen in een vertrouwde categorie ondergebracht kan worden? Of doen er zich subtiele, bijna onopgemerkte verschuivingen voor, zoals met de PC en het internet? En, als het zich voordoet, wordt het dan opgemerkt? Of gaat het ermee als met de klokkenluider: hij wordt niet gehoord en als hij wel gehoord wordt, belandt hij op dood spoor of wordt regelrecht buitengezet?

Diagnostiek heeft robuuste bevindingen en resultaten nodig om het gedrag van de cliënt te kunnen voorspellen en controleren. Meta-studies beloven zulke bevindingen te bieden, maar laten bescheiden tot gemiddelde waarden en grote variabiliteit zien: r capaciteiten - werkprestaties tussen .15 en .65, predictieve validiteit van tests: r tussen de .02 en .75 en sekseverschillen in schoolprestaties: d-waarden tussen -0,30 tot 0,17. Dat duidt op steekproef- en lokaal gebonden zijn van resultaten. De recente replicatiecrisis is er altijd al geweest, zie de varianties van d- en r- waarden

Geraadpleegde literatuur en vermelde studies zijn noodzakelijkerwijs een selectie. Ze zijn zo gekozen dat kennis nemen daarvan in staat stelt specifieke onderwerpen uit te diepen. Omdat tijdschriften elektronisch toegankelijk zijn is dat te verwezenlijken. Als begrippen en namen niet bekend klinken zijn ze te vinden via zoekmachines.
Opsomming van empirische onderzoeken is droge kost. Ze zijn genummerd en apart gezet. Naast klassieke studies wordt recent onderzoek vermeld met voorkeur voor meta-studies. Recent onderzoek toont gesofisticeerde analyses maar ze zijn bijna nooit aanleiding om oude conclusies te herzien. Na elke opsomming van empirisch onderzoek volgt een conclusie. Er is naar tegenvoorbeelden gezocht. Als ze er zijn, worden ze vermeld.
Elk boek is, zoals Plato (4de eeuw *Before Common Era*: BCE) zei een *farmacon*, dat wil zeggen zowel een medicijn tegen het vergeten als een gif voor het geheugen en het kennen. Elke lezer loopt het risico ontgift te moeten worden na lezing van welke tekst dan ook.

Je kunt een tekst op verschillende manieren lezen. Een is zoeken naar eigen thema's en vaststellen of die er niet of onjuist in staan: Ik word niet/verkeerd geciteerd. Een brede tekst wordt nauwelijks gelezen, gegeven de gerichtheid op precies te omschrijven onderwerpen en de uitwerking ervan in *least publishable units*. Een collega zei eens onderzoek doen is als een blind paard op je doel afgaan. Verder wordt de originaliteit wordt betwist: Dat heb ik al eerder verteld in mijn artikel van 10 jaar geleden. Waar heb ik dat eerder gehoord. Een tweede is nagaan of het waar of juist wat er staat. Dat is moeilijk in de psychologie. Daar geldt eerder hermeneutiek dan waar/juist versus onwaar/vals. Dat wil zeggen het betreft het ontdekken van betekenis en zin. Wat is de boodschap? Een derde is retorisch lezen: de argumentatiestructuur ontrafelen.

Een boekje schrijven is pretentieus. Dit boek bevat ook nog eens geen nieuwe inzichten en is niet oorspronkelijk. Het brengt iets bijeen van wat de psychologie te bieden heeft voor de diagnostiek van de cliënt. De pretentie is dat bestaande kennis en methoden bijeengezet, uitgelegd en geëvalueerd worden. Theorie en onderzoek worden geordend: drie elementen, die bronnen en drie oriëntaties. Er wordt een keuze gemaakt om diagnostiek los van behandeling te analyseren.

Er wordt geen poging gedaan om de overmaat aan onderzoek en theorie te verzoenen of te integreren. Ze krijgen afzonderlijk plaats en aandacht. Combinaties leiden tot onvruchtbare hybriden, *muddy dichotomies*.

Sommigen twijfelen of psychologie een wetenschap is. Ze is het in ieder geval niet als je de Newtoniaanse natuurkunde als maatstaf neemt. Ze zou slechts informatie bevatten over samenhangen tussen gedragingen onderling en over relaties met de sociale en fysische context. Wetenschappelijke tijdschriften zijn in dat geval dagbladen, *journals* en onderzoekers journalisten. Ze bevatten informatie, geen kennis of inzicht. Deze twijfel wordt toegelaten: misschien geldt dat voor de psychologie. Er wordt alleen geprobeerd een stap verder te komen. Er is een kennisbestand en dat zinvolle informatie, feiten over het gedrag van de cliënt op een rij zet.

Zonder als diagnosticus bescheiden hoeven te zijn over het vak - het is wat het is - leiden theorie en onderzoeksresultaten tot een bescheiden houding over wat beschreven, verklaard en voorspeld wordt over het gedrag van de cliënt.

De *persona* - een archetypisch beeld van wat een diagnosticus is, gegeven zijn karakter en drijfveren - ziet er hier zo uit: Hij biedt geen wetten, waarheden, dogma's en geen nieuwe apparaten of technieken. Hij toont relativeringsvermogen en een door het kennisbestand van de psychologie getemde verbeelding over het kennen van de persoon die de cliënt is. Hij is niet in eerste instantie degene die problemen oplost, onderzoeksgeld binnenhaalt of een onderzoeksbedrijf opzet en runt. Hij is probleem-, niet oplossingsgericht. In dit boek wordt in problemen en niet in oplossingen gedacht. Het is niet gericht op hulpverleners. Het wil de vorming van een *rich and sufficient description of the problem field* bevorderen met inbegrip van de ecologische structuur waarin een individuele vraag of probleem is ingebed. Het doel is het midden te houden tussen het Hegeliaanse *Dass Wahre ist das Ganze* en het reductionisme van het behaviorisme, neurologisme, empirisch-analytisch methodologisme, staticisme, monotheorisme en beperking tot één psychologische thema of één methodische invalshoek.

Het midden houden, maat houden is één van de vier Aristotelische deugden. Ze kwamen in de Middeleeuwen meestal aan de orde bij begrafenissen, nu bij afstudeerpraatjes, feestjes

en begrafenissen. Daar hebben we nu de *Big Five* voor. De vier passen op diagnostiek van de cliënt: matigheid, verstandigheid/voorzichtigheid, dapperheid/sterkte en rechtvaardigheid. Met het wijzen op deugdzaamheid is het uitkijken, want deugden wordt door niets zozeer onderuitgehaald als door de saaiheid van haar pleitbezorgers.

Psychologische diagnostiek is een bezigheid *sui generis.* Daar bedoel ik mee dat de diagnosticus zelf zijn weg moet zoeken. Dat wil niet zeggen dat ze autonoom is, want ze bestaat in relatie tot het kennisbestand van de psychologie, de cultuur en de samenleving. Haar rol en bijdrage worden steeds opnieuw uitgevonden. Recent staat *valorisatie* in de aandacht. Wat draagt ze bij aan analyse van vragen en problemen in de samenleving? Deze term neemt enigszins gas terug van het streven Engelstalige tijdschriften te vullen met korte artikelen. Dat kan immers een circuit zijn dat losgezongen is van de realiteit van alledag en dat zichzelf in stand houdt.

Dit boekje is niet om te volgen of te repliceren, maar om eigen te maken, kritisch te evalueren en te valoriseren om terminologisch bij de tijd te komen. Elke tekst, protocol, model, empirisch resultaat wordt rammelende ballast, dood gewicht als het niet wakker gelezen en getoetst wordt door een geïnteresseerde, gemotiveerde en onafhankelijke lezer.

Leeswijzer

Ordening van de tekst Elk onderwerp is gemarkeerd door twee nummers: het eerste duidt het hoofdstuk aan en het tweede de sectie. De secties bevatten sub-thema's die cursief gedrukt zijn. Elke sectie wordt afgesloten met een *Samenvatting en conclusie*. Voorbeelden van empirische studies zijn gemarkeerd en genummerd. Als er tegenvoorbeelden gevonden zijn worden ze vermeld. Bovendien wordt de variantie van de uitkomsten van vergelijkbare studies vermeld. Elk hoofdstuk eindigt met *Reflectie en evaluatie*. Reflectiviteit houdt in dat de ambigue, gefragmenteerde en betwistbare aard van psychologische kennis wordt erkend zonder voorbij te gaan aan feiten en aannemelijke interpretaties. Het houdt ook in dat er multiple perspectieven op gedrag zijn en er verschillende constructen geldig zijn om die te beschrijven. Er wordt gelet op de positie die een onderzoeker inneemt. Is hij lid van een groep met bepaalde praktijken? Dit beïnvloedt zijn waarneming en begrip van gedragsverschijnselen. Evaluatie verwijst naar een oordeel over de kwaliteit van de constructen en kennis: wat is bereikt, zijn pretenties waargemaakt, worden bevindingen realistisch weergegeven? Aan het eind van de hoofdstukken worden *Termen en Begrippen* vermeld die als subjectindex fungeren.

Geen glossary Psychologische constructen liggen niet vast. De Wittgenstein van *Philosophical Investigations* (1953) beweert dat definities in de psychologie niet veel verhelderen omdat ze betekenis krijgen in een context. De sinoloog Schipper (1988, p. 13) noemde het ontbreken van definities kenmerkend voor de open leer van het Chinese Taoïsme. Heideggers *Sein und Zeit* (1927/1962) kan gelezen worden als een kruistocht tegen de solide definities van het Zijn door de christelijke filosofische traditie. Het delen van deze standpunten houdt in dat ik er niet naar streef de diagnosticus definitieve constructen aan te praten. Aan het eind van een hoofdstuk worden thema's vermeld.

Boxen Er zijn boxen tussen de tekst gevoegd in een kleinere letter. Deze vermelden achtergronden van debatten en controversen en bevatten ook idiosyncratische gezichtspunten op een psychologisch of maatschappelijk verschijnsel.

Referenties De literatuur kan nooit recht gedaan worden. De referenties zijn bedoeld als vindplaats. Er wordt vanuit gegaan dat de lezer toegang heeft tot elektronische tijdschriften. Voor niet onmiddellijk inzichtelijke begrippen zijn er de zoekmachines.

Voorkennis Er wordt bij de lezer een actieve kennis van methodologie, psychologische theorievorming, psychometrie en statistiek op BA niveau verondersteld.

Doelgroep en Doel

Dit boek is bestemd voor studenten BA en MA psychologie en pedagogiek en HBO toegepaste psychologie. Het is ook bedoeld voor die practici en GZ psychologen in opleiding, die enerzijds het aan de leiband een model of protocol lopen waarderen als het verstrekken van duidelijkheid hoe het moet, maar anderzijds willen reflecteren op voorschriften, die ze soms ontoereikend achten voor hun beroepsuitoefening. De tekst is ook bestemd voor leken die psychologische diagnostiek kritisch willen volgen. In deze tekst wordt het 'leken'

oordeel over psychologische diagnosticeren serieus genomen en studies naar dat oordeel worden vermeld. Bovendien is hetgeen niet direct gearticuleerd is toegankelijk via het internet.
Het doel is het bevorderen van een zoektocht naar de grondslagen voor diagnostiek van ons gedrag en de activiteit van het diagnosticeren. Er wordt een reflectieve houding aangemoedigd: denk na en kijk uit, voor je springt. Er worden geen praktische handvatten aangereikt die voorschrijven wat een diagnosticus moet doen in de vorm van diagnose-behandel-combinaties (DBCs) of Handelings Gericht Werken (HGW). Daar zijn andere teksten voor van bijvoorbeeld Witteman et al. (2014) en De Bruyn et al. (2015).

Vooruitblik

Eerder heb ik aannemelijk gemaakt (ter Laak 2011, 2015) dat psychologische diagnostiek heeft geen eigen en/of specifiek gedragsdomein (materieel object) en geen eigen gezichtspunt op gedrag (formeel object) heeft. Daarin wijkt ze af van de sub-disciplines in de psychologie. Ze gaat in beginsel over elk gedrag en ontleent *ad hoc* aan het kennisbestand van de psychologie verschillende theorieën en methoden en empirische bevindingen. Ze is uit zichzelf onbepaald. Een gevolg daarvan is dat ze niet ingebed is in de structuur van een universitaire faculteit, afdeling of vakgroep. Niettemin wordt diagnostiek binnen vrijwel alle afdelingen van de psychologie en pedagogiek faculteiten bedreven en in de praktijk van de gezondheids-, onderwijs- en organisatiepsychologie en de pedagogiek.
Uit en door zichzelf bepaalde disciplines met eigen formele en/of materiële objecten en de toegepaste psychologie en pedagogiek stellen eisen aan de diagnostiek. Er is bemoeienis met diagnostiek vanuit statistische modellen, behandelaars die de meerwaarde van diagnostiek voor hun werk aangetoond willen zien, validiteitstheorie, testtheorie en van het kritische publiek. Ik werk de drie bronnen van de kritische wetenschappelijke bemoeienis met diagnostiek van de cliënt uit. Verbeteren zij de diagnose? Dat wil zeggen zorgen ze voor substantieel minder valse positieven en negatieven bij categorietoewijzing en dragen ze bij aan preciezere voorspellingen en effectievere controles? Draagt wetenschappelijke bemoeienis bij aan het verbeteren van het diagnostisch proces zoals de diagnosticus dat voltrekt?
In Nederlandse handboeken over diagnostiek wordt gesuggereerd dat de diagnose en het diagnostisch proces zo vormgegeven zijn dat zij de toets van de wetenschappelijke kritiek kunnen doorstaan. Witteman et al. (2014) beschrijven de diagnostische stappen die

'...een leidraad voor het uitvoeren van psychodiagnostisch onderzoek in de praktijk bieden; *'...inzicht geven in hoe verschillende stappen in het psychodiagnostisch onderzoek op een goede manier uitgevoerd kunnen worden'* (onderlijning jtl).

De Bruyn et al. (2015) maken een handleiding, een praktijkleer die laat zien,

... *'dat diagnostische besluitvorming in de praktijk op een wetenschappelijk-professioneel verantwoorde wijze kan worden doorlopen'* (p. 18).

Als *caveat* vermelden de auteurs weliswaar de beperkingen van het kennisbestand van de psychologie maar het volgen van het prescriptieve model van de diagnostische cyclus leidt volgens hen tot wetenschappelijk en professioneel verantwoord diagnosticeren.
In deze tekst doe ik *een stap terug* en werk uit wat de kritische bemoeienis van vertegenwoordigers van de wetenschap en het publiek vertelt over de diagnose en het proces. Daartoe heranalyseer ik in Hoofdstuk I de tegenstelling tussen klinische en statistische predictie. Statistici beweren het te winnen van de diagnosticus wat betreft het aantal valse positieven en negatieven bij categorietoewijzing en voorspellen van waarden op (on)gewenste gedragingen van de cliënt. Dit resultaat was onder meer de aanleiding voor onderzoek van het diagnostisch proces. Kan de diagnosticus geldige *cues* opsporen, vaststellen en ze combineren om tot een deugdelijke diagnose te komen? Professionals die op behandeling gericht zijn tref je vooral aan onder pedagogen, ontwikkelings-, bedrijfs- en klinisch psychologen en (kinder)psychiaters. Diagnosticeren is tijdrovend. Levert ze nu iets op om de goede, de naar verhouding effectiefste behandeling uit de voorraad te kiezen? Dit heeft vorm gekregen in de zoektocht naar slimme Diagnose-Behandel-Combinaties (DBCs). Ze zijn uitgewerkt in procedures voor handelingsgericht werken. Daar staat de oplossing van het probleem en het antwoord op de vraag centraal. Deze tekst is niet oplossings- maar probleemgericht, dat wil zeggen inzicht verkrijgen in het (on)gewenste gedrag van de cliënt. Dit standpunt wordt beargumenteerd en ik neem de DBC hausse onder de loep (Hoofdstuk II). Niet alleen de diagnose en het diagnostisch proces moeten deugen, ook behandelingen dienen bewezen effectief te zijn. Dat komt tot uitdrukking in de aandacht voor *evidence based treatments*. Zijn er voldoende DBCs en voldoen ze aan de verwachtingen? De centrale en principiële vraag bij diagnose en onderzoek is die van de geldigheid, validiteit van de diagnose en het diagnostisch proces en van het onderzoeksresultaat. In Hoofdstuk III wordt nagegaan wat de betrouwbaarheids- en validiteitstheorie te bieden hebben voor diagnose, behandeling en hun combinaties. Het instrumentarium van de diagnosticus bevat in beginsel elke actie die zinvolle informatie over het gedrag van de cliënt oplevert. Niettemin spelen toetsen, tests en vragenlijsten een prominente rol. In Hoofdstuk IV vraag ik me af wat de testtheorie te bieden heeft voor de diagnose en het diagnostisch proces. Diagnostiek ligt onder vuur van wetenschappers, professionals en leken. In hoofdstuk V ga ik in op interne kritiek van wetenschappers en onderzoekers op hun werk en die van het publiek op psychologisch onderzoek en diagnostiek. Daarbij kun je denken aan de recente replicatiecrisis en aan wetenschapsfraude. Deze wordt in de dagbladen uitvoerig besproken. Het publiek denkt bij diagnostiek of testen meestal het eerst aan toetsen en testen van cliënten. Deze staan onder het regime van testtheorie. Wat heeft deze de diagnosticus en het proces te bieden?

Ik stel me de vraag of kritiek vanuit modellen en de statistiek en DBCs, kennis over het valideren van instrumenten en procedures en de testtheorie hout snijdt en leidt tot betere diagnose van de individuele cliënt. Alle pretenderen voorschriften te bieden die het aantal fouten bij diagnoses vermindert. Bovendien zijn ze van invloed op onderzoek van het diagnostisch proces zelf voor zover dat tot fouten aanleiding geeft. Passen ze bij het detectivewerk dat een diagnose is of schieten ze hun doel voorbij? Ook de cliënten kunnen de zaak mooier voorstellen dat ze is. Ze kunnen *faken* op persoonlijkheidstest en in de anamnese niet de hele waarheid vertellen.
Op de kritiek van het publiek is het niet gemakkelijk een antwoord te geven. Zeggen dat ze geen experts zijn en hun mond moeten houden kon lang geleden nog, maar het sneed en snijdt geen hout. Er is bovendien geen afdoende antwoord. Het enige wat de diagnosticus kan doen is transparant zijn over zijn handelwijze en duidelijk zijn over wat een diagnose vermag. De cliënt heeft recht op een objectieve kijk op zijn gedrag. Dat wil zeggen dat zijn vraag/probleem voldoende en rijk beschreven wordt, zoals Simon en de Groot dat bedoelen (ter Laak, 2015). Het komt erop neer om realistische verwachtingen te wekken bij de cliënt. Hij is doorgaans een individuele persoon, maar het kan ook over grote eenheden gaan. Daarbij kun je bijvoorbeeld denken aan een school die een programma om pesten tegen te gaan inkoopt. Er zijn er al 100 in omloop in Nederland. De school mag verwachten dat de diagnosticus een effectieve aanwijst zonder aanziens des persoons. De school kan immers blij zijn met een programma dat zijn doel niet bereikt. Het bureau HALT kan een programma voor jonge alcoholdrinkers aanbieden maar het blijkt empirisch geen effect te hebben. De behandeling van jonge criminelen in een instelling blijkt geen effect te hebben op de recidive. Wat betekent deze kennis voor de diagnosticus en voor de behandelaar. Het is niet gemakkelijk een programma te stoppen. Er zijn immers belangen mee gemoeid.

Het eerste onderwerp (Hoofdstuk I) sluit aan bij een oud thema uit de diagnostische literatuur: klinische versus statistische predictie. Het is als specifieke tegenstelling gearticuleerd door Meehl (1954) en heeft veel onderzoek opgeleverd. Is er een bevredigend antwoord of suddert de tegenstelling door? Het tweede is onder de naam Diagnose-Behandel-Combinatie bekend (Hoofdstuk II). Daar sluit handelingsgericht werken (HGW) in de recente diagnostisch literatuur bij aan Het is een uitloper van aloude zoektocht naar en selectie van effectieve behandelingen en therapieën dat door Cronbach en Snow (1977) is ingezet onder de naam *Aptitude Treatment Interaction* (ATI). Hoofdstuk III betreft de waarheid, geldigheid van de uitspraken over het gedrag of het psychologisch functioneren van de cliënt. Psychometrici hebben dat op haar staart getrapt en ingekaderd met hun betrouwbaarheids- en validiteitstheorie. Past dat bij het n = 1 onderzoek dat de diagnosticus verricht? Verbetert het de diagnose en het proces? Het vierde thema (Hoofdstuk IV) verbindt diagnostiek met de nadrukkelijk aanwezige testtheorie in de diagnostiek. Hoewel zij een minitheorie is (een curve, functie, dimensie) over minigedrag (één of een aantal antwoorden op vragen, items) is haar impact op het instrumentarium groot. Helpt deze

atomistische aanpak van het gedrag van de cliënt de diagnosticus en het proces? Welke winst behaalt hij door het toepassen van klassieke en moderne testtheorie? Het vijfde thema (Hoofdstuk V) over fraude in de wetenschap is voor de hand liggend. Wetenschappers zijn mensen. Köbben heeft zo'n 40 jaar geleden al een poging gedaan het wetenschapsbedrijf antropologisch te analyseren. Dat werd hem niet in dank afgenomen. De wetenschapsjournalist van Kolfschoten heeft met Köbben samengewerkt en in 2012 een samenvatting gemaakt over fraude aan de Nederlandse universiteiten. Daarnaast vermelden kranten betrekkelijk selectief psychologisch onderzoek. Ze zijn er doorgaans als de kippen bij om fraude te melden en de replicatiecrisis in sociaalwetenschappelijk onderzoek onder ogen te brengen. Daarnaast is er vooral oog voor spectaculaire uitslagen en specifieke thema's zoals man-vrouw verschillen.
Hebben de statistici, behandelaars met hun behoefte aan goede BDCs, validiteits- en testtheoretici een punt in hun kritiek op de diagnose en het diagnostisch proces of maken ze er een stropop van die ze vervolgens verbranden? En, geven de woordvoerders van het publiek een fair beeld? Of is het *fun* gaten in het wetenschapsbedrijf te schieten?

I Statistisch versus klinisch voorspellen

De diagnosticus moet zich verantwoorden over zijn werkwijze tegenover de wetenschap. Een uitwerking daarvan is de controverse tussen klinische en statistische predictie. Tegenover behandelaars en hulpverleners moet hij zich evenzeer verdedigen. Wat is de meerwaarde van uw diagnostiek? Hij wordt in de verdediging gedrongen omdat zijn vak geen eigen theorie, thema, perspectief of methode heeft. Als gevolg daarvan ontbreekt een sociale structuur in de vorm van een vakgroep, afdeling, disciplinegroep of departement dat een eigen domein en werkwijze claimt. Zo kun je de diagnosticus altijd vragen: 'Wat bent u aan het doen; is het wel wetenschappelijk verantwoord wat u doet', 'wat is uw meerwaarde voor de behandeling, interventie of beslissing'? In de controverse komt de zelfbetrokkenheid van de diagnostiek tot uitdrukking.

Evenals de slagzin: 'Geneesheer (m/v) genees u zelf', kan gezegd worden: 'Diagnosticus diagnosticeer u zelf'. Dat komt hem op veel kritiek te staan gegeven een normatief model of statistische regel. Daarnaast leidt deze kritiek tot de analyse van diagnosticeren en klinisch oordelen zelf. De informatieverzameling bij een cliënt en het interpreteren van die informatie worden regelmatig gekwalificeerd als klinisch, subjectief of on- of pseudowetenschappelijk.

De klinisch-statistisch controverse begon in de jaren twintig van de 20ste eeuw. Het debat zwakt nu af. In plaats van de producten van klinische predictie af te doen als subjectief is er is aandacht voor het proces zelf. Voor de waarneming is een soortgelijk verloop beschreven door Brunswik in de jaren 50 van de vorige eeuw. Zijn ecologische, probabilistisch functionalistische aanpak van waarnemen en diagnosticeren is levend, bijvoorbeeld in het werk van Gigerenzer. De wending naar het diagnostisch proces levert een beeld van wat de diagnosticus werkendeweg doet. Hoe ziet klinisch oordelen en voorspellen eruit en hoe het statistische? Kan de tegenstelling conceptueel overbrugd en empirisch beslecht worden? Helpen voorschriften, regels en modellen het valse positieven en negatieven in klinische diagnoses te verminderen? Zijn voorschriften en modellen realistisch en gebruiksvriendelijk, zodat de practicus er zijn voordeel mee kan doen? Naast de zoektocht naar de feilbaarheid van de diagnosticus/clinicus wordt geëist dat zijn diagnose aansluit op een behandeling: Wat is de meerwaarde van diagnostiek voor de interventie? Wat is de achtergrond van de tegenstelling? Waarover gaat die? Zijn het twee partijen met verschillende werkwijzen of houden ze zich met verschillende zaken bezig?

1. Oorsprong van de controverse

Aan het *mind-body* vraagstuk ontkomt de psychologie en de diagnostiek niet. Hoe zijn kenmerken van objecten, gebeurtenissen, objecten en stimuli en gedrag gerelateerd aan- en gerepresenteerd in onze waarneming en daarna in de modellen? Wat is de oorzaak van individuele verschillen in de waarneming van objecten, gebeurtenissen en gedrag? Kunnen we ons gedrags-, biologische en fysische kenmerken accuraat waarnemen? Er is geen één op één correspondentie tussen stimuli en waarneming. Biofeedback en leugendetectors doen ons verbaasd staan over onze reacties. We weten niet hoe onze huidgeleiding gaat, hoe hoog onze bloeddruk is. We weten in beperkte mate hoe anderen onze gedragingen zien en interpreteren. Ons lichaam en denken zijn ons het meest nabij, maar tegelijkertijd vol verrassingen en zelfs vreemd.

In het begin van de 20ste eeuw was er een opleving van Kants denken (Honderich, 2005, pp. 466-472). Zijn epistemologie is een Duitse reconstructie van het Britse empirisme en Franse rationalisme. Mensen kunnen het object op zich (*Ding an sich*) niet kennen. Niettemin hebben ze met hun *praktisch Vernunft* weet van kenmerken door waarneming en vergelijking. De kenner brengt iets mee, los van alle waarneming: de oordelen van ruimte, tijd en causaliteit. Deze verlichten of zo u wil verduisteren alles maar dat weet je nooit precies. Onze begrippen hebben een waargenomen lading nodig om iets voor te stellen en om er iets mee te kunnen doen. Objecten en gedragsbegrippen hebben - al kennen we ze niet op zichzelf met het *reine Vernunft* - betekenis door ons *praktische Vernunft*. *In praktischer Hinsicht* bestaan volgens Kant de objecten en concepten, constructen wel degelijk. Hij erkende het bestaan van *correct theory* (Goodman et al., 2011) dat wil zeggen dat ze over iets in de werkelijkheid gaat en die beantwoordt er ook voor een deel aan. Dat is mogelijk door *the blessing character of abstraction*. Tegelijk confronteert abstractie ons met het feit er objecten en constructen bestaan waar we geen greep op krijgen. We kunnen handelen en denken *alsof* we dingen kennen en inzicht hebben in de natuur en gedragskenmerken. Dit *alsof* is door de Neo-Kantiaan Vaihinger (1911/1986) wijsgerig uitgewerkt.

In de 19de eeuw kwam '*science'* in een stroomversnelling. De psychologie moest een plek vinden naast of tussen de filosofie, wiskunde, chemie, biologie en sociologie (Simonton, 2004). Deze zoektocht is te herkennen in het nomothetisch-idiografisch debat dat verbonden is met de klinisch-statistisch tweedeling. Bij nomothetisch kun je denken aan Newtons beeld van universele wetten en bij idiografisch aan de fenomenologische, gesitueerde beschrijving van het gedrag van de persoon in zijn sociale en fysische context.

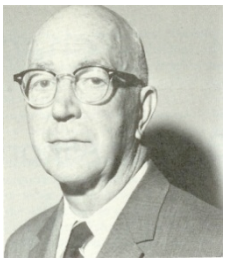
Nomothetisch of idiografisch Natuurkunde, scheikunde en biologie maar ook sociologie (Dürkheim) namen een hoge vlucht aan het eind van de 19de eeuw. De Neo-Kantiaan *Windelband* gaf als decaan in 1894 een lezing in de VS waarin hij onderscheid maakte tussen de nomothetische en idiografische studie van complexe voorwerpen, gebeurtenissen en verschijnselen (Hurlbert & Knapp, 2006). Nomothetisch verwijst naar het zoeken van

algemene wetten door middel van analytisch denken, theorievorming en empirisch toetsen. Idiografisch onderzoek gaat over het achterhalen van het unieke, niet-inwisselbare van objecten, gebeurtenissen en gedrag. Het bevat concrete en zo volledig mogelijke beschrijvingen. Beide werkwijzen zijn volgens Windelband op *alle* voorwerpen, gebeurtenissen en verschijnselen van toepassing.
Psychologen vroegen zich af welke werkwijze paste bij hun discipline. De organisatiepsycholoog Münsterberg (1863-1916) noemde de studie van de geschiedenis idiografisch maar die van de psychologie nomothetisch (1912). Zij lijkt volgens hem meer op natuur- dan op geesteswetenschappen. Hij verwachtte dat psychologen wetten zouden vinden die toegepast konden worden bij werknemers. Hij spande zich in om de basis te leggen voor een *linking science* die onderzoek en praktijk verbond.
In het begin van de 20ste eeuw gaven onderzoekers en praktiserende diagnostici concrete en zo volledig mogelijke beschrijvingen van individuele cliënten om oorzaken van stoornissen op te sporen en sollicitanten/studenten aan functies en opleidingen toe te wijzen. Dit is een idiografische benadering maar de overtuiging was dat er algemene wetten mee ontdekt en getoetst konden worden. Freud pretendeerde algemene wetten van hysterie op het spoor te komen door onderzoek bij één cliënt. Skinner karakteriseerde zijn n = gedragsmodificatiestudies als nomothetisch want hij wilde leer*wetten* formuleren. Hieruit blijkt dat 'concreet en compleet' bij één cliënt volgens deze auteurs tot universele wetten kan leiden.

Nomothetisch en idiografisch Stern (1911) voelde voor de idiografische benadering in persoonsleer en diagnostiek. Zijn biografische methode is gericht op de studie van de persoon als uniek geheel. Daarnaast onderscheidt hij de psychografische methode die in staat stelt de algemene structuur van de persoonlijkheid te ontdekken. Op deze algemene structuur zijn variaties mogelijk: individuele verschillen en een idiosyncratische ontwikkeling door de tijd heen. Uniciteit is idiografisch en structuur nomothetisch. Allport droeg er toe bij dat in de jaren dertig van de 20ste eeuw het onderscheid tussen nomothetisch en idiografisch in de Amerikaanse psychologie bekend en aanvaard werd. In tekstboeken wordt zijn opvatting over de persoon beschouwd als een verdediging van de idiografische opvatting en methode. Een bekend citaat gaat over hoe met 100% zekerheid voorspeld kan worden dat professor X naar de film gaat. Je moet daarvoor zijn attitudes en voorkeuren kennen, aangevuld met unieke informatie over professor X, bijvoorbeeld, dat hij zijn been niet gebroken heeft, zijn favoriete PhD studente geen liefdesverdriet heeft en troost behoeft, geen colleges hoeft te geven en zijn favoriete actrice niet speelt:

'*if seven in ten Americans go to the movie each week, it does not follow that I have seven in ten chances of attending. Only a knowledge of my attitudes, interests and environmental situation will tell you my chances, and bring your prediction from a 70% actuarial statement to a 100% certain individual prediction'* (1942, pp. 16-17).

Allport zegt over de benaderingen overigens: *a complete study will embrace both approaches* (p. 32).



De visie van Gordon W. Allport op de persoon en op methoden om hem te bestuderen was van invloed in de VS en Europa. Zijn opvatting paste in de tijdgeest van de ondernemende, onafhankelijke Amerikaanse burger vanaf midden jaren 30 van de 20ste eeuw, kort na de economische/ bankencrisis van 1929.

De *geschiedenis* laat zien dat er sprake is van een combinatie van idiografisch en nomothetisch en niet van een tegenstelling. Filosofen bijvoorbeeld Pierce (1935, zie Honderich, 2005) noemden de scheiding tussen nomothetisch en idiografisch onhoudbaar. Iedere wet is gecontextualiseerd door tijd, plaats en (sociale) condities. Zelfs de zwaartekrachtwet geldt onder specifieke condities. Hij onderscheidde drie manieren van redeneren: (1) abductie: het verzamelen van feiten die tot een theorie kunnen leiden (2) deductie, logisch redeneren vanuit premissen en (3) inductie, empirisch onderzoek: conclusies trekken na verzameling en toetsing van gegevens. Het onderscheid is niet zo strikt als handboeken suggereren. Niettemin speelt het in de diagnostiek een rol. Klinisch wordt daarbij verbonden met idiografisch en statistisch met nomothetisch. De tegenstelling is niet door de alleszins redelijke overwegingen van Stern en Pierce verdwenen. Zo beweert Cautin (2011) dat de kloof tussen wetenschappers en practici samenvalt met de verdeling van psychologen die geloven in het klinisch oordeel en zij die vertrouwen op resultaten uit empirisch onderzoek. De tegenstelling is levend.

De cliënt verstrekt idiografische informatie De diagnosticus heeft in zijn eerste gesprek te maken met idiografische informatie. Zijn kennis over het gedrag van de cliënt begint ermee. Het op waarde schatten van deze informatie erkent de dynamische aard van het gedrag van de cliënt. De cliënt beschouwt hij niet als inwisselbaar voor ieder ander met dezelfde vraag. Deze inductieve en individuele aanpak sluit generalisatie niet uit. Salvatore en Valsiner (2010) stellen verzoening voor. Dat is wat al te gemakkelijk zegt Molenaar (2004) en wakkert de competitie aan. In zijn essay kiest hij voor een idiografische aanpak met moderne technieken en data analyses. Hij wijst op een kwetsbaar punt van de psychologie van het individu: zij stelt interindividuele verschillen in steekproeven vast en verwaarloost tijd-afhankelijke ontwikkelingsvariatie binnen iedere persoon. Hij pleit voor n = 1 tijdserie studies. Dat is mogelijk door een relevant segment van het gedrag van een persoon *on line*

te volgen zonder dat het hinderlijk is voor hem. Nu we meer dan een decade verder zijn komt dat in zicht door het gemak waarmee personen alles wat ze meemaken delen op facebook en twitter.

De noodzaak om *time-series n = 1 studies* te doen verdedigt Molenaar door te wijzen op het achterblijven van de psychologie bij wetenschappelijke ontwikkelingen in andere disciplines. Hij doelt op de stochastische revolutie die Einstein heeft ingezet voor natuurkundige processen. Deze wordt halfhartig of helemaal niet verwerkt door psychologen. Ze volgen eerder Newton en schatten parameters in steekproeven door middel van correlationele en experimentele studies. Ze verwaarlozen individuele dynamiek, ontwikkeling, adaptatie en leren in veranderende contexten. De revolutie heeft nauwelijks effect gehad in de psychologie. In de natuurkunde is er de leer van de bewegingen van alle kleine en lichte deeltjes waarin kansverdelingen een centrale rol vervullen. In de biologie bestudeert men patroonvorming en evolutieprocessen en in de wiskunde is er de stochastische calculus. Intra- en interindividuele verschillen zijn gelijk onder beperkte en meestal niet realistische condities. Dat betekent volgens Molenaar dat de klassieke psychometrische en statistische modellen niet volstaan voor de studie van het gedrag van de individuele persoon. Hij beveelt tijdserie analyse aan voor alle domeinen: intelligentie, cognitie, klinische, sociale, ontwikkelings- en onderwijspsychologie. Hij ziet mogelijkheden om *on line* de zich ontvouwende geschiedenis van een persoon vast te leggen.

Ruim een decade na zijn Manifesto zijn er nog geen sprekende voorbeelden, hoewel het intensieve verzamelen van gegevens verzamelen al begonnen is: *data mining*. Dat is overigens gericht op consumentengedrag en niet op de psychologische persoon. Bovendien is ongeveer alles van een persoon is bekend bij bedrijven en overheden. Miljarden telefoongesprekken worden opgeslagen. Molenaars manifest is uitdagend maar heeft (nog) geen gevolgen gehad voor diagnostiek. Zijn aanpak levert een massa data op die niet eenvoudig te interpreteren is. Feitelijk worden tijdseriegegevens vaak eenvoudig geïnterpreteerd, zoals toename, afname, snelheidsverschil en afwisseling van lijnen en curven van een specifiek gedrag. Je kunt je afvragen of men voor zulk een eenvoudige interpretatie zulke complexe technieken en statistiek nodig heeft. Niettemin, zou zulke data voor een onderzoeker een *Fundgrube* zijn. Ze leiden mogelijk tot nieuwe constructen over het verloop van gedrag binnen een persoon door de tijd heen.

Samenvatting en conclusie

De controverse tussen statistisch en klinisch voorspellen van het gedrag van de cliënt is verbonden met het onderscheid tussen nomothetisch (universele wetten voor alle cliënten) en idiografisch (iedere cliënt is uniek). Dit onderscheid hield aanvankelijk geen tegenstelling in omdat objecten, gebeurtenissen en gedrag op beide wijzen onderzocht kunnen worden. Het is bovendien geen tegenstelling in de n = 1 studies van Freud en Skinner. Allport is ten onrechte tot voorstander van de idiografische methode verklaard. Dit neemt niet weg dat het onderscheid diagnostici verdeelt in clinici en statistici. Het voorstel om n = 1 tijdserie

studies te beginnen klinkt plausibel maar ze zijn moeilijk uit te voeren en de interpretatie van zo'n hoeveelheid gegevens is vaak eenvoudig: (curvi)lineaire toe- of afname.

2. Overeenkomsten en verschillen

In 1944 suggereerde Sarbin *overeenkomst* in het perspectief op integreren van informatie. Beide houden in dat er een lineaire regressieanalyse wordt verricht waarin een aantal voorspellers gecombineerd wordt om een criterium te voorspellen. De statisticus gebruikt daarvoor een formule opgesteld op basis van empirisch onderzoek. De clinicus integreert de informatie op basis van zijn ervaring met soortgelijke gevallen. Dat betekent dat er voor de clinicus bij een feitelijk lineaire combinatie van predictoren niet meer dan een gelijkspel in zit. De formule minimaliseert immers die fouten: een rechte, de regressielijn die de som van de gekwadrateerde afwijkingen minimaliseert. De verhouding tussen de formule en klinisch diagnostische voorspelling en tussen de psychometricus en de practicus heeft een analogon in de natuurkunde: wetten van de thermodynamica, zie bijvoorbeeld Baker (2010, 3de druk).

Box De practicus/clinicus wint het nooit van de formule

Als een formule geldig is kan een diagnosticus alleen gelijkspelen met zijn oordeel of voorspelling. De formule veronderstelt een statisch verschijnsel en dat is zelden juist in gedrag. De formule is een ideaalvorm. Een gelijkspel zit er niet in want de formule is een gesloten systeem waar niets bij of af moet en kan. Maar de werkelijkheid is geen gesloten systeem. Onze eigen toekomst kennen we: de dood. Of, zoals de dichter Bloem (*Verzamelde Gedichten,* 1982, p. 223, Amsterdam: Athenaeum-Polak & Van Gennep) het zegt: *Denkend aan de dood kan ik niet slapen, En niet slapend denk ik aan de dood, En het leven vliedt gelijk het vlood, En elk zijn is tot niet zijn geschapen'.* Het spel houdt nooit op tot subatomair vervallen toe want er is geen stilstand. De dingen liggen niet verpletterd op zichzelf. Je krijgt de atomen in materiaal nooit tot stilstand want het absolute nulpunt is onbereikbaar. Dat is een ideale toestand uitgedrukt in een formule.

Het is onbegonnen werk om de psychometrische formule - de psychometricus - van zijn ongelijk te overtuigen: gelijk of ongelijk is een zelfs een categoriefout: zijn gelijk is besloten in de regels van het systeem. Maar er is naast het formele *kritische Vernunft* een geleefde *praktische Vernunft*. Een voorbeeld: de psychometricus heeft een zoon. Hij komt bij u om zijn psychologisch rapport te bespreken. Hij is er als psychometricus van overtuigd dat wat u doet klinisch oordelen is en wat u zegt opmerkingen met een vluchtig bestaan zijn: *in the eye of the beholder*. De practicus weet immers niet wat meten is, begrijpt IRT niet en interpreteert er lustig op los. Toch vraagt hij u: 'Wat betekenen die WISC en CBCL scores voor mijn kind'? U legt dat met omzichtigheid uit. Uw psychometricus aanhoort uw CBCL score interpretatie en zegt naar aanleiding daarvan dat hij hoopt dat zijn kind goed zal reageren op Ritalin. Uw navraag of de performale en verbale scores op de WISC meer dan een standaarddeviatie uiteenlopen bij het totaal IQ van 107 antwoordt hij bevestigend. U licht dat verschijnsel toe en uw psychometricus begrijpt dat want hij ziet de voorkeur van zijn kind voor bepaalde spelletjes. Hij bedankt u, gaat verder met zijn onderzoek en onderwijs met als boodschap dat het klinisch oordeel onbetrouwbaar is en criteria matig voorspelt. Voor hem bestaat de facto het onderscheid tussen *praktische* en *reine Vernunft*. De empirie laat bij voorspellen

zien dat de formule voor groepen, steekproeven het meestal wint van het particuliere klinische oordeel. Dat is niet altijd het geval, soms is er een gelijkspel. De formule en het gedrag van het kind van de psychometricus passen niet één op één.

De *verschillen* tussen diagnostiek en psychometrie en diagnosticus en psychometricus worden vaker benadrukt dan overeenkomsten. Onderscheidingen maken en verschillen van mening zijn (terecht?) een moeilijk te stuiten bezigheid van wetenschappers.

Box De statisticus tegen de clinicus

In het dagelijks leven maken we ons niet druk over tegenstellingen zoals klinisch-statistisch of wetenschappelijk-pseudowetenschappelijk. Er zijn bijvoorbeeld evenveel alternatieve genezers als huisartsen. De eersten beloven je te helpen waar de reguliere geneeskunde ophoudt. Sommigen willen de reguliere zelfs vervangen door de alternatieve. Zolang er zich geen incidenten voordoen kunnen we ermee leven. Soms gaat het mis, bijvoorbeeld als een filmster of echtgenoot van een politicus hulp zoekt in het alternatieve circuit en dat uit de hand loopt. Meestal gaat dat weer uit als een nachtkaars. Alternatieve genezers worden zelden vervolgd, zelfs als er iets mis gaat. Hetzelfde geldt overigens voor de reguliere geneeskunde. Professionele en niet-professionele genezers lopen elkaar kennelijk niet voor de voeten. Een groep reguliere genezers die tegen kwakzalverij is wordt weliswaar nu en dan boos maar krijgt weinig voor elkaar: het alternatief genezen vermindert er niet door. Dat geldt niet bij *Clinicia versus Psychometrica.* Statistici beschuldigen clinici van onwetenschappelijkheid en onethisch gedrag (Dawes et al., 1989). Er zijn ingrediënten voor een controverse: er zijn twee partijen, de controverse is uitgelegd en bevestigd door wetenschappers en professionals, één van de twee noemt zich superieur, de ander betwist dat en zegt dat de vergelijking niet eerlijk is. Slechts 5% van werk van psychiaters en psychologen zou een predictie bevatten. Verder worden voorspellingen van onervaren clinici vergeleken met statistische resultaten en de statistische predictie is niet altijd en mogelijk dus noodzakelijk superieur.
Hoe hiermee om te gaan? Feitelijk kunnen er verschillende predicties gedaan worden. Als ze op hetzelfde neer zouden komen zou de beslissing met minste kosten de voorspelling moeten leveren. Statistische predictie claimt de overwinning en het imago van de invoelende, wijze en helpende clinicus loopt schade op (Dawes et al. , 1989 in het prestigieuze tijdschrift *Science*). Mediatie tussen sterkten en zwakten van beide helpt niet. De controverse blijft in stand als de nieuwkomers tot het veld meteen ingedeeld worden bij één van de kampen en het aantal overlopers of bekeerlingen gering is. De groepen worden gestereotypeerd: statistici zijn bang voor mensen en clinici voor formules.

Verschillen blijken uit de houding van clinici en statistici en uit de wijze van informatie verzamelen, integreren en uit het toetsen van veronderstellingen. In de klinische werkwijze is het object een subject, een persoon met unieke kenmerken. Hij leeft in zijn specifieke sociale context en maakt zijn unieke ontwikkeling door. Een vraag of probleem van de persoon wordt geanalyseerd door te kijken naar de combinatie van zijn specifieke kenmerken, ervaringen en sociale omstandigheden. Er wordt een theorie ontworpen om relevante kenmerken van zijn gedrag te selecteren, te beschrijven en te verklaren. Dit levert

een concrete beschrijving van die cliënt op en mondt uit in een advies of voorspelling en soms inclusief een voorstel voor een behandeling, een voorspelling van de reactie op de behandeling, bijvoorbeeld de kans op recidive, of succes op school en in het beroep.
De psychiater Holt (1970) beweerde dat predictie zelden voorkomt in de hulpverlening. Er wordt eerder een beeld van de cliënt geschetst op basis van informatie uit verschillende bronnen. De clinicus/diagnosticus kan naar eigen inzicht bronnen raadplegen, zoals het levensverhaal en informatie van familie, leeftijdsgenoten en collega's. Hij gebruikt zijn ervaringen met vergelijkbare gevallen en zijn theorie over (on) gewenste gedragingen van de cliënt. De clinicus gaat een dialoog aan met de cliënt. Er zijn geen standaardvragen en het verloop is dynamisch. Er is geen protocol, hooguit enkele onderwerpen. Hij hanteert geen formule om de informatie uit allerlei bronnen, inclusief test- en vragenlijstscores te integreren. De clinicus probeert de cliënt te begrijpen en zijn gedrag te verklaren: hermeneutisch en empirisch analytisch met behulp van gegevens uit tests en vragenlijsten. Hij zou zo doende streven naar een 100% of perfecte voorspelling, zoals eerder Allport zei.
In de statistische werkwijze wordt de cliënt beschouwd als element van een populatie. De empirische kennis over hoe het gedrag van de steekproef eruit ziet wordt benut om de individuele cliënt te karakteriseren. Het advies heeft de vorm van een categorisering, voorspelling of beslissing, gebaseerd op empirisch onderzoek van de steekproef waarvan de cliënt deel uitmaakt. De statisticus toont voorkeur voor objectief te scoren tests en vragenlijsten en voor feitelijke informatie over de cliënt. Testgegevens kunnen geïnterpreteerd worden aan de hand van Normen die opgesteld zijn voor de specifieke populatie waarvan de cliënt een element is. Scores leiden tot een profiel dat vergeleken kan worden met het profiel van de steekproef. Hij vormt geen specifieke theorie voor een cliënt. De voorspelling wordt gedaan in de vorm van een kansuitspraak. De correlatie tussen predictoren en criteria is immers nooit 1.00.

Statistici en clinici nemen elkaar graag de maat Einhorn (1986) is voorstander van de statistische werkwijze en beschrijft clinici als deterministen. Ze beschouwen gedrag als uitdrukking van onderliggende causale processen. Zij reconstrueren het verleden van de cliënt op zo'n manier dat het huidige gedrag uit dat verleden moet volgen, bepaald als het is door de persoon in combinatie met eerdere en nu aanwezige condities. Het gedrag kan dus 100% voorspeld worden en een beslissing kan met zekerheid worden genomen. Ze beweren dat deze aanpak het probleem van de cliënt het best oplost of zijn vraag beantwoordt. Deze beschrijving lijkt op de manier waarop artsen te werk gaan: zij proberen een specifieke en noodzakelijke oorzaak te vinden voor een bepaalde ziekte. Clinici noemen hun werkwijze dynamisch, globaal, zinvol, flexibel en holistisch. Statistici omschrijven die als primitief, slordig, voorwetenschappelijk, ongedisciplineerd en ongecontroleerd.
Statistici noemen zich probabilisten. Zij denken in waarschijnlijkheden en accepteren fouten want variabelen en criteriummetingen bevatten meetfouten. Ze stellen vast dat alle psychologische kennis gefragmenteerd is en elke theorie een deelbenadering is van gedrag.

Het accepteren van fouten leidt per saldo tot minder fouten over een reeks voorspellingen. Ze gaan ervan uit dat bevordering van de wetenschap de koninklijke weg is om cliënten te helpen. Statistici beschrijven zichzelf als rigoureus, wetenschappelijk, precies en zorgvuldig. Clinici zeggen: ze zijn pedant, gefragmenteerd, laag bij de grond, triviaal en gekunsteld. De tegenstellingen zijn:

gegevens van individuele cliënt tegenover groepsgegevens
dialoog versus scores van tests in een profiel
voor- of pseudowetenschappelijk tegenover wetenschappelijk
'loser' - 'winner'
determinist - probabilist
onmiddellijke hulp tegenover kiezen voor lange termijn winst op basis van wetenschappelijk onderzoek.

Samenvatting en conclusie

Hoewel statistisch en klinisch voorspellen aanvankelijk als aanvullend beschouwd werden, ontstond een controverse door verschillen te benadrukken over het object: een unieke persoon versus een element uit een populatie, over het hoe: dialoog versus objectieve tests en standaardmethoden, over wie: de persoon van de clinicus als een romantische intuïtionist versus een neutrale statistische wetenschapper en over wat de beste hulp is: communicatie met de cliënt versus het beoefenen van wetenschap. Als men elkaar karakteriseert als een verliezer of winnaar, is er sprake van een controverse.

3. Inhoud van de controverse

Meehl (1920-2003) heeft een centrale rol gespeeld in de bepaling van de inhoud van de controverse. De oorsprong is de publicatie van *my disturbing little book* (1954): *Clinical versus statistical prediction: A theoretical analysis and review of evidence.*


Paul Meehl: filosoof, psycholoog, methodoloog en therapeut

Meehl beweert dat Allport het idee van statistische predictie zou accepteren omdat hij 99% of 80% ook voldoende gevonden zou hebben. De 100% is immers een uitzondering. En, het is terecht dat men formules (actuarische tabellen) niet gebruikt onder afwijkende omstandigheden, bijvoorbeeld als de professor ziek is of de Nobelprijs ontvangt. De aard van de tegenstelling spitst hij toe op de manieren van combineren van gegevens. De één kiest voor een empirische combinatie met behulp van een op onderzoek berustende

formule. De ander opteert voor een intuïtieve combinatie van gegevens. Hij voegt eraan toe dat het verschil er toe doet want het kan tot verschillende conclusies leiden. Meehl geeft het voorbeeld van een formule die voorspelt dat een crimineel 70% kans heeft om binnen drie jaar weer de fout in te gaan. Een empathisch sociaalwerker zegt echter dat er nauwelijks kans is op recidive. Wat moet een rechter met deze tegenstrijdige informatie?
Eén hoofdstuk van zijn boek gaat over de vergelijking tussen de werkwijzen. Daarvoor gebruikte hij twintig studies waarvan de resultaten konden worden vergeleken. Zijn conclusie luidde*:*

'... in all but one of which the predictions made actuarially were either approximately equal or superior to those made by a clinician' (1954, p. 119).

Meehls omschrijving van de tegenstelling en de vergelijking in zijn meta-studie *avant la lettre* hebben het debat richting gegeven: *formulas versus heads.* Ze hebben de competitie tussen *Clinicia* en *Psychometrika* op gang gebracht en uitgewerkt. Grove en Meehl (1996) vatten bijvoorbeeld 136 studies samen die *informal, subjective, romantic and impressionistic prediction* vergelijken met *formal, mechanical and algorithmic prediction'.* Het resultaat en de conclusie zijn dezelfde als die van Meehl (1954). Dit artikel wordt al op deze plaats vermeld omdat het een opsomming bevat van de argumenten waarmee clinici hun werkwijze verdedigen. De auteurs uit het statistische kamp spreken alle tegen.

Clinici benutten statistische formules én klinische informatie en dat geeft het beste resultaat.
Objectieve instrumenten zijn niet meer valide dan klinische procedures.
Voor de groep waartoe mijn cliënt behoort is geen regressieformule bepaald.
De beschikbare formule past niet bij mijn cliënt.
Predictiestudies zijn duur en worden inadequaat uitgevoerd, bijvoorbeeld bij specifieke steekproeven.
Ons doel is niet voorspelling: we helpen de cliënt zijn gedrag te accepteren of het te veranderen.
We werken idiografisch en niet nomothetisch.
Op kwalitatieve klinische gegevens kan geen regressieformule bepaald worden.
We gaan een dialoog aan met de cliënt.
Aan die vergelijkende studies hebben onervaren clinici, studenten deelgenomen.
Ons doel is begrijpen en helpen, niet voorspellen.
Actuariële methoden (predictieformules) leiden tot kansuitspraken, onze cliënten begrijpen dat niet ze willen een richtlijn.
Regressieformules verouderen snel, evenals resultaten op tests en vragenlijsten: samenhangen zijn niet stabiel door de tijd heen en variëren per steekproef.

Meehls boek en de erop volgende competitie tussen de werkwijzen beheersen de discussie. Zijn interpretatie van de zwakte van het klinisch oordeel/predictie is dat men verkeerde gewichten aan variabelen toekent. Clinici zijn zelfs inconsistent in het toekennen van gewichten die ze zelf afgesproken hebben. De strijd lijkt op het gevecht van man tegen

machine in het begin van de industriële revolutie: de man (een zekere Henry, een arbeider in de VS) won even in het begin maar verloor uiteindelijk met grote cijfers. Eender verging het de klassieke boekhouder die het even van de reken- en boekhoudmachine won.

Samenvatting en conclusie

De inhoud van de tegenstelling tussen statistisch en klinisch voorspellen is door Meehl geformuleerd en door anderen overgenomen. Zij bestaat uit verschillende wijzen van informatie integratie: in het hoofd of met een formule c.q. aan de hand van een actuarische tabel. De vergelijking valt in het voordeel van statistische predictie uit. Sommige auteurs willen de werkwijzen combineren, anderen beschouwen ze als categorisch verschillend. Er zijn drie zaken te onderscheiden: algemene wetten, gecontextualiseerde wetten en persoonlijke, unieke betekenisverleningen. De verschillen zijn breed uitgemeten en de controverse leeft nog. Er wordt beweerd dat wetenschap niet over individuen gaat: *Scientia non est individuorum*. Ze gaat over schatten van parameters in een populatie. Weer anderen zeggen dat tijdserie analyses van gedrag van één cliënt de informatie bieden die de diagnosticus nodig heeft om de vraag te beantwoorden.

4. Combineren van statistische en klinische strategieën

In 1961 nam de Amsterdamse methodoloog en onderwijspsycholoog De Groot (1914-2006) deel aan het debat. Hij stelde een *improvement design* voor dat benut kan worden bij een voorspellingsvraagstuk, bijvoorbeeld van schoolsucces op basis van schoolprestaties, IQ en prestatiemotivatie. Er is een predictieformule bepaald op basis van valideringsonderzoek. Clinici krijgen deze informatie en zij mogen alles toevoegen wat ze maar willen, bijvoorbeeld projectieve tests, interviewgegevens en informatie van significante anderen. Ze beschikken zo over meer informatie dan in de formule verwerkt is. Het resultaat van de predictieformule kan met oordelen van de clinici worden vergeleken. Het is geen eerlijke vergelijking maar eerlijkheid is geen methodologisch concept merkt De Groot op*: If fairness were a value, we would never play off a hypothesis against such an ignorant thing as the null hypothesis*. Daarmee gaf hij te kennen dat het toetsen tegen een nulhypothese weinig opleverde. Zijn *improvement design* werkt gunstig als 'formule' tegenover 'formule + klinische informatie' verschil maakt. Er werd geen verschil gevonden. De Groots voorstel was overigens niet bedoeld om aan de strijd deel te nemen. Hij wilde een manier vinden om voorspellingen van schoolsucces te verbeteren.

Vijftig jaar na Meehls boek hebben Westen en Weinberger (2004) de controverse onder de loep genomen. Klinische activiteiten hebben twee betekenissen: de eerste is de uitleg van Meehl: de informele, subjectieve integratie van informatie die onwetenschappelijk is en ook niet verbeterd wordt door training. Als tweede noemen ze de mentale processen en producten van het klinisch oordeel op basis van praktijkervaring. Deze processen en ervaring leiden niet zonder meer tot fouten. Onder gunstige voorwaarden zijn klinische oordelen valide. Dat is het geval bij het gebruik van goede tests, vermijden van te algemene

en abstracte uitspraken, oordelen over het gebied van eigen expertise en het organiseren en benutten van *feedback.* Ze pleiten er voor deze valide klinische kennis in de predictieformules op te nemen. De auteurs stellen een combinatie voor die het beste van beide benaderingen behoudt en in één formule onderbrengt. De categorische afwijzing van het klinisch oordeel in 1954 wordt in 2004 weliswaar genuanceerd maar de inbreng van de clinicus moet wel in de statistische formule opgenomen worden.

Samenvatting en conclusie

De benaderingen zijn gecombineerd in een *improvement design* (De Groot). Clinici krijgen alle informatie en mogen alles er zelf bij halen wat ze maar willen. Men zou verwachten dat de voorspelling verbetert in de lijn van de gedachte: hoe meer, hoe beter. Dit bleek niet uit de gegevens. In een vijftig jaar later opgemaakte balans wordt erkend dat onder bepaalde condities klinisch voorspellen valide is maar die moeten vervolgens ondergebracht worden in een formule.

5. Onderzoek om het geschil te beslechten

Volgens Meehl en Grove & Meehl (1986) is de casus rond. De clinicus is de verliezer en daar moet hij het mee doen. Dit nam niet weg dat er studies volgden die de predicties van de twee vergeleken volgens regels van Meehl: twee groepen met dezelfde middelen en materialen onderzoeken, oude rotten en nieuwkomers (studenten) nemen deel en er is een goed meetbaar en betrouwbaar criterium. Sawyer (1965) maakte 75 vergelijkingen, waarvoor hij 45 studies bijeensprokkelde. Hij onderscheidde enkele manieren van gegevensverzameling: observaties, impressies, gesprekken met de cliënt tegenover objectieve tests en vragenlijsten. De beste predictie werd verkregen met objectieve tests gecombineerd met empirisch bepaalde formules. Deze aanpak was beter in 75% van de 75 casus. Dit resultaat is duidelijk. Niettemin beweerde Korman (1968) het tegenovergestelde: klinische predictie is superieur bij het voorspellen van prestaties van functies in het bedrijfsleven. Ofschoon hij veel literatuur vermeldt, is er geen overlap met het materiaal van Sawyer. Is dit een voorbeeld *cherry picking* of is er serieus verschil tussen de domeinen? Is de managementprestatie zo onderscheiden van andere prestaties, dat die alleen in een klinisch oordeel valide gevat kan worden?

Clinici verzetten zich tegen de conclusies van Sawyer. De psychiater Holt (1970) verweet Sawyer dat kruisvalidering ontbrak. Deze procedure kan helpen om te ontdekken of een predictieresultaat overeind blijft in nieuwe steekproeven. Verder merkte hij op dat slechts in 12 van de 45 studies een realistisch criterium werd voorspeld. In 23 van de 45 namen onervaren clinici deel en/of de steekproeven waren gering van omvang. Grove en Meehl hebben deze kanttekeningen alle als onjuist bestempeld.

Het argument dat resultaten niet gevoelig zijn voor verschil in domeinen blijkt uit een samenvatting van Dawes et al. (1994). Statistische predictie is beter bij schoolsucces, fraude in bedrijven, levensverwachting, succes van militaire training, hartinfarct,

neuropsychologische aandoeningen, op borgtocht vrijgelaten worden, behalen van diploma van politieofficier, geweld en psychiatrische diagnoses. Kormans gebied: het succes in het bedrijfsleven, staat er niet bij.

Box Een populair thema: voorspellen van geweld en misdaad

Meehl vermeldt een lineair model met drie variabelen: soort misdrijf, aantal veroordelingen en het overtreden van gevangenisregels. Het voorspelde recidive beter (r = .22) dan het oordeel van experts (r = .06). De ernst van het misdrijf correleerde .27 met het oordeel van de experts en .47 met wel of niet heroïnegebruik. Een ander voorbeeld is het voorspellen van het risico op geweld bij psychiatrisch gestoorde misdadigers. Harris et al. (2002) gebruikten een vragenlijst als voorspeller (VRAG: *Violent Risk Appraisal Guide*) in een groep van 467 gestoorde mannelijke misdadigers. De statistische voorspelling was superieur aan het klinisch oordeel van psychiaters en psychologen. Drie variabelen werden benut: betrouwbare achtergrondvariabelen: de 'moeilijke jeugd', volwassen aanpassing, IQ en persoonlijkheidsvariabelen; behoeften van de misdadigers, bijvoorbeeld aan therapie. Gewelddadig recidivisme werd gedefinieerd als elke aanval die tot lichamelijk letsel leidde, kidnapping, gewapende overvallen en seksuele misdrijven. Patiënten die opnieuw in de fout gingen hadden hogere VRAG scores dan minder gewelddadige patiënten. Er was overigens geen relatie tussen VRAG scores en recidive bij *misdadigsters.* Vrouwen kregen lage scores op de VRAG. Weinig variantie vermindert de kans op covariantie en dus op correlatie. Dat kan het geval zijn bij de criminele vrouwen.

Twee meta-studies Na Sawyer, Korman en Grove & Meehl zijn meerdere studies verricht. Hier worden twee exemplarische meta-studies vermeld. Een meta-studie is een procedure die in staat stelt om tot een besluit te komen over een aantal vergelijkbare onderzoeken. Elke studie is een element in een steekproef. De eerste is de eerdergenoemde van Grove et al. en hun conclusie verschilt niet van die van Meehl. Vaak wordt verondersteld dat *fresh research* beter is: het nieuwe is altijd beter. Is elke vondst zo aan de tand des tijds onderhevig dat alles steeds opnieuw onderzocht moet worden?

Box Wat is beter: statistische/mechanische of klinische predictie?

Deze meta-studie bestrijkt meer gebieden dan voorafgaande en de auteurs werken de informatie genuanceerder uit*. Grove en collega's* (2000) verzamelden eerst studies over het thema via Psylit (nu Psychinfo) en via Medline vanaf 1966 (Sawyers studie ging tot 1965) tot 1998. Ze selecteerden 163 studies waarvan er 136 voldeden aan de opname criteria. Een criterium was bijvoorbeeld dat de clinicus en de formule van dezelfde voorspellers en criteria gebruikmaakten. Er werden 617 vergelijkingen gemaakt. Er werden codes gebruikt om de studies te kunnen vergelijken, zoals jaar van publicatie, type: boek, artikel in erkend tijdschrift, promotiestudie; type voorspeller: objectief, subjectief, expertoordeel, resultaten op persoonlijkheidsvragenlijsten. Verder werden kenmerken van clinici (opleiding, ervaring, duur van training) gecodeerd. Dit kwam tegemoet aan bezwaren van Holt (1970) tegen de studie van Sawyer (1965). Codes voor de voorspellers waren categorie, schaal

en soorten test, zoals DSM en *SATs* (*School Achievement Tests*). Succes werd bepaald door het aantal correcte voorspellingen (*hits*) en door de hoogte van de correlatiecoëfficiënten.
De codes werden door twee onafhankelijke beoordelaars toegekend. Er was voldoende overeenstemming. Een statistisch criterium werd gedefinieerd om te bepalen of de voorspellingen beter/slechter of gelijk waren. Verschillende gedragingen, stoornissen en ziektebeelden werden onderscheiden: recidive, schoolsucces, leiderschap, fraude, jeugddelinquentie, psychiatrische diagnose, geweld, vrijlaten op borgtocht, huwelijksgeluk, productiviteit, succes als manager (heeft Korman dan toch ongelijk?), diagnose van homoseksualiteit, failliet gaan van bedrijven, beroepstevredenheid, adoptie, succes als ondernemer en therapiesucces. Zevenenveertig procent van de voorspellingen was ten gunste van de statistische, mechanische formule of actuarische methode, 47% was gelijk en in 6% kwam de klinische werkwijze als winnaar uit de bus.

Aegisdottir et al. (2006) komen niet uit de stal van Meehl. Ze vatten 56 jaar onderzoek naar voorspellen van succes van interventies in de geestelijke gezondheid samen. Ze berekenden effectgroottes (ESs). Cohen (1988) beval deze aan als vervangers voor significantietoetsen. Een gewogen effectgrootte - de d-waarde - wordt gebruikt om het verschil tussen behandelingen te kwantificeren. Dat is het gemiddelde verschil op de afhankelijke variabelen tussen twee steekproeven uitgedrukt in standaarddeviatie eenheden, gecorrigeerd voor de steekproefomvang. Een grote steekproef draagt meer bij dan een kleine en een representatieve meer dan een gelegenheidssteekproef. De d-waarde is het verschil tussen de gemiddelde accuraatheid van het oordeel van de ene conditie afgetrokken van de gemiddelde accuraatheid in de andere conditie, gedeeld door het geometrisch gemiddelde van de twee standaarddeviaties.

Box Statistische en klinische predictie in de GGZ

Aegisdóttir et al. (2006) deden een meta-studie over 56 jaar onderzoek van professionals in de geestelijke gezondheidszorg. Ze verzamelden 67 onderzoeken die 92 vergelijkingen mogelijk maakten. Zij gebruikten *Effect Sizes* (ESs) om de grootte van de verschillen te kwantificeren. Cohen (1988) heeft deze voor vergelijkingen tussen gemiddelden aanbevolen. ESs zijn een alternatief voor het toetsen van nulhypothesen. Een gemiddeld gewogen effectgrootte (d) wordt gebruikt om het verschil tussen klinisch en statistisch voorspellen uit te drukken. De d-waarde is het verschil tussen de gemiddelde accuraatheid van het oordeel van de ene werkwijze verminderd met de gemiddelde accuraatheid van de andere werkwijze, gedeeld door de gemiddelde standaardafwijking (geometrisch gemiddelde) van de twee, gecorrigeerd voor de steekproefgrootte.
De studie is op dezelfde manier vormgegeven als die van Grove et al. (2000). De gebieden van onderzoek zijn: hersenletsel, persoonlijkheidsstoornissen, duur van verblijf in het ziekenhuis, psychiatrische diagnose, aangepast zijn, prognoses van ziekteverloop, geweld, IQ, schoolprestaties, MMPI-profielen, zelfmoordpogingen en homoseksualiteit. Een geselecteerd deel van de steekproef van 36 studies toonde effectgrootten van d = -0,57 en d = 0,73. Gemiddeld genomen waren de d-waarden in het voordeel van de statistische predictie. Er werd een grotere nauwkeurigheid (p. 341) van de statistische methoden geconstateerd. De studie bevestigt de conclusies van Meehl (1954) en Grove et al. (2000).

Beide studies bevestigen Meehls conclusie van 1954 op basis van de wet van de grote getallen. Ze leveren extra informatie. Grove et al. vonden bijvoorbeeld geen relatie tussen succes van voorspellen en jaar van publicatie: ouder betekent niet slechter en het soort publicatie: tijdschriftpublicaties worden hoger gewaardeerd, maar er was geen verschil in kwaliteit tussen tijdschriftartikelen en boeken en boekhoofdstukken. In de medische setting was de predictie beter dan in de psychologische maar artsen voorspelden niet beter dan psychologen op vergelijkbare taken. Ervaring van de diagnostici hing *niet* samen met de kwaliteit van hun predicties. Aegisdóttir en collega's vonden een bescheiden effect van het soort voorspelling. Aanpassing en prognose van de stoornis, geweld en schoolprestaties werden naar verhouding beter voorspeld (d-waarden van 0,14 tot 0,17) Cohen (1992) noemde waarden tussen 0,20 en 0,49 een *small to medium effect*. Opmerkelijk was dat clinici afkomstig uit een andere setting als waar de gegevens verzameld waren iets beter voorspelden. Ze waren accurater bij stoornissen waarmee ze *niet* dagelijks te maken hadden. Lineaire regressieformules leverden hogere ESs op dan logische regels ontleend aan protocollen, opgesteld op basis van werkwijzen van ervaren clinici. De hoeveelheid informatie hielp de clinici tot op zekere hoogte, maar meer informatie verslechterde de voorspelling. Dit verschijnsel deed zich ook voor in de studie van De Groot (1961). Beginners, novicen verloren het van de formules maar expert-clinici kwamen aardig in de buurt. Vrijwel dezelfde conclusies zijn gerapporteerd in Garb (1998) die meer dan 1000 artikelen over dit thema samenvatte. Een overzichtsstudie van Wood et al. (2002) meldt hetzelfde resultaat.

De onontkoombare conclusie is dat - voor een steekproef - statistische voorspelling de voorkeur verdient. En, als vaak *the winner takes it all* (zie Dawes, et al., 1989, 1994, 2000, 2005). Toch is de overwinning niet zo afgetekend als deze auteurs suggereren. Er is een *gelijkspel* in ongeveer de helft van de voorspellingen. Dit feit rechtvaardigt het onvoldoende om klinisch diagnosticeren te kwalificeren als 'een kloppend verhaaltje vertellen', 'uitdrukking van een verlangen en zucht naar irrationaliteit', 'te groot vertrouwen in je niet pluis gevoel' en 'geloof in eigen kunnen'. Maken statistici er niet een stropop van die ze vervolgens verbranden?

Samenvatting en conclusie

Het boek van Meehl uit 1954 heeft de controverse gemunt als statistisch versus klinisch voorspellen en inhoud en vorm ervan bepaald. Vergelijkende studies tonen de statistische voorspelling als winnaar op veel thema's: van schoolsucces tot vrijlating op borgtocht. Twee meta-studies (Grove et al., 2000; Aegisdóttir et al., 2006) bevestigen dat beeld: in de helft van de studies is statistische voorspelling beter, in 45% gaan de twee gelijk op en in 5% is de klinische predictie de winnaar. De meta-studies laten zien dat de kwaliteit van de voorspelling niet afhangt van de bron van publicatie, dat medische criteria iets beter worden voorspeld dan (gezondheids) psychologische, dat hoeveelheid ervaring van clinici

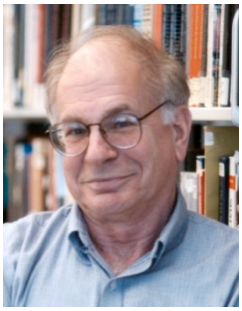
niet met predictief succes samenhangt en dat lineaire regressieformules minder fouten maken dan regels ontleend aan protocollen, opgesteld na ondervraging van experts. Het pleit is gewonnen en de tegenstelling lijkt empirisch beslist. Niettemin is er in bijna de helft van de studies een gelijkspel. Dit rechtvaardigt het niet om klinisch voorspellen weg te zetten als irrationeel of een mooi verhaal. Het opent de weg naar analyse van het klinisch oordelen als een verschijnsel *sui generis.* Zo komen intuïtie en beslissen in natuurlijke en riskante situaties in beeld. Daarvoor zijn geen formules beschikbaar.

6. Analyse van klinisch oordelen en voorspellen

Deugdelijk categoriseren, voorspellen van waarden op (afhankelijke) variabelen en verklaren/controleren door middel van de onafhankelijke variabele, bijvoorbeeld een interventie of therapie zijn doelen van diagnostiek. De klinisch-statistisch controverse heeft ondanks het gelijkspel de statistische werkwijze als winnaar aangewezen. De clinicus maakt met andere woorden fouten. Hoe komt dat? Deze vraag is heeft geleid tot het entameren van studies naar de aard van het proces van klinisch voorspellen. Zo kom je erachter waar de fouten gemaakt worden. Je kunt het proces verklaren op basis van het denken, de *beliefs* en het gedrag van de diagnosticus. En, als je daar zicht op hebt kun je het proces disciplineren.

Het klinisch oordeelsproces wordt kritisch bejegend. Daar doen zich een aantal verschijnselen voor die teruggedrongen moeten worden volgens de auteurs. Lilienfield et al. (2003) waarschuwden bijvoorbeeld voor te veel *vertrouwen* in eigen oordeelsbekwaamheid. Men voelt zich zekerder als meer en nieuwe informatie wordt toegevoegd ook als dat meer van hetzelfde is en geen extra bewijs bevat. Meer heeft zo het karakter van een mantra. Daarnaast wijzen ze op het zoeken naar bevestiging van wat een keer als idee heeft postgevat, op het zien van correlaties die er empirisch niet zijn en op het wantrouwen van clinici in correlationeel en experimenteel onderzoek. Het gaat over steekproeven en er wordt zelden naar individueel gedrag gekeken. Moore en Healy (2008) waarschuwen voor *overconfidence:* overschatting van de eigen prestaties en het beter achten van eigen prestaties in vergelijking met die van anderen. Onderzoek laat zien dat onderschatting zich voordoet bij verschillende soorten taken. Ik pas dit toe op clinici en kan daar waarschijnlijk ook economen, rechters en dokters aan toevoegen. Bij moeilijke taken overschatten zij hun werkelijke prestaties maar nemen ten onrechte aan dat hun prestaties slechter zijn dan die van anderen. Diagnostiek is een complexe taak waarbij veel factoren en predictoren in interactie betrokken zijn. Bij moeilijke taken roepen ze gemakkelijk de hulp van deskundigen in. Bij gemakkelijke taken onderschatten ze hun prestaties maar nemen ten onrechte aan dat ze beter zijn dan anderen. Bij omschreven, heldere en concrete taken op een voor hen bekend domein overschatten ze zichzelf minder en gaan ze er niet vanuit dat anderen hun oordeel kunnen verbeteren. De aard van de taak speelt een rol.

Heuristieken en Vertekeningen Klinische predictie is een complexe taak met risico's van over- en onderschatten van eigen kunnen in het vaststellen van diagnostisch waardevolle informatie. Een speciale plaats in de discussie over het klinisch oordeel neemt de heuristieken en vertekeningen traditie (*Heuristics & Biases*: HB) in. Deze is gestart door wijlen Amos Tversky en Nobelprijswinnaar Daniel Kahneman. De twee vulden elkaar inhoudelijk en qua werkstijl aan. De eerste bedacht 's avonds de experimenten die de tweede de volgende dag uitvoerde. Ze bedachten in hoog tempo ingenieuze experimenten. De laatste is opgeleid als psycholoog en bekend van zijn kritische analyse van het menselijk oordeel, vooral bij het voorspellen van winsten op de beurs.


Daniel Kahneman kreeg de Nobelprijswinnaar voor de economie. Hij is psycholoog en vormde met Amos Tversky een ideaal onderzoekersteam.

Kahneman las Meehls verontrustende boekje van 1954 in 1955. Hij was betrokken bij de selectie van kandidaten voor de officiersopleiding in Israël. Hij stuitte op subjectieve overtuigingen van gesettelde officieren om kandidaten af te wijzen of toe te laten. Zij namen statistische informatie over de prestaties op de officiersschool niet serieus. Ze vonden bovendien dat deze experts evenals methodologen en statistici onjuiste antwoorden gaven op de vraag welke omvang een steekproef moest hebben om een conclusie te kunnen trekken. Deze maakt eerst kans stabiel te zijn als die berust op een representatieve streekproef van voldoende omvang. Deze overtuigingen zijn onjuist en dat leidde tot een sceptische houding tegenover expertoordelen (Kahneman, 2003; Kahneman & Klein, 2009). Dit werd de start van HB onderzoek naar het gebruik van simpele vuistregels en vertekeningen.
Een zo'n gemakzuchtige vuistregel is de *availability heuristic*. Dit is het verschijnsel dat we gebeurtenissen waarschijnlijker vinden als we die gemakkelijk uit ons geheugen kunnen ophalen. Als een diagnosticus veel delinquenten in zijn praktijk heeft zal hij gemakkelijk delinquentie waarnemen in gedrag van een cliënt, ook als die geen delinquent is. Verder zijn leken, clinici en andere experts gebrekkige informatieverwerkers. Ze zijn selectief, nemen waar in de richting van hun eigen standpunt, negeren zinvolle beschikbare informatie en houden geen rekening met het ontbreken van belangrijke informatie. Ze verdisconteren in het oog springende informatie meer dan empirische gegevens die over langere termijn verzameld zijn. Eén recent geval van kindermishandeling in de krant krijgt zo meer gewicht dan statistieken over jaren. Ongeveer 3% van de kinderen komt in aanraking met misbruik,

maar als we na een documentaire over misbruik op de TV een schatting moeten maken is die hoger dan 3%. We veronachtzamen de mate van voorkomen van verschijnselen (incidentie, *base rate*). Nieuwe en interessante maar irrelevante informatie krijgt een beslissend gewicht: een goed uitziende, grotere sollicitant krijgt eerder een baan dan een minder goed uitziende, kleine maar gelijkwaardige kandidaat.

Als niet aan statistische en logische voorschriften niet gehoorzaamd wordt, leidt dat tot verkeerde uitspraken over frequentie en samenhang van verschijnselen of worden onjuiste beweringen gedaan. Clinici zondigen tegen statistische kansregels. Ze schatten de frequentie van verschijnselen niet juist in. Frequente verschijnselen doen zich - als de naam zegt - vaak voor. Als men echter vraagt wat gevaarlijker is, het weerlicht, de bliksem of het ontbreken van een trapleuning wordt het weerlicht regelmatig gevaarlijker geacht. Ze onder- of overschatten verandering en groei. Exponentiële groei wordt systematisch onderschat. Ze baseren oordelen op infrequente, extreme waarden. Het verschijnsel van regressie naar het gemiddelde is voor leken en ook voor experts tegen-intuïtief. Het leidt er ook toe dat we oorzaken toeschrijven aan gebeurtenissen die er niet zijn. Stel, de toestand 'je goed of slecht voelen' fluctueert. Als je je niet goed voelt, ga je naar de dokter. Hij (m/v) geeft je een medicijn of placebo. Omdat je toestand fluctueert is er een kans dat de toestand van 'slecht voelen' overgaat in 'goed of neutraal voelen'. De fluctuatie en niet het medicijn/placebo is dan de 'oorzaak' van de verandering.

Verder veronderstellen clinici soms relaties tussen verschijnselen die er niet zijn: illusoire correlaties. Een voorbeeld is het beoordelen van homoseksualiteit op basis van menstekeningen. Men veronderstelde dat er een verband is tussen het tekenen van gespierde mannen en homoseksualiteit. Er is echter geen empirisch onderzoek dat het bevestigt. Dit een van de redenen dat interpretatie van tekeningen door leken en clinici door onderzoekers gewantrouwd wordt.

Kans houdt in dat er onzekerheid is over het zich voordoen van een gebeurtenis. Dit is ongemakkelijk voor leken en experts. We denken soms als deterministen: als iets gebeurt, moet het ook zo gebeuren. We doen daarom soms alsof we controle hebben: *illusion of control.* Denk aan de begrotingen van ministeries, de vijfjarenplannen van de Sovjeteconomie, de kosten van de Noord/Zuidlijn in Amsterdam, het tijdstip van gereedkomen van de hogesnelheids- en Betuwelijn, enzovoort. Clinici en lekenoordelaars zijn gevoelig voor volgorde-effecten. Als gegevens in een logische volgorde worden gepresenteerd, stellen ze niet gauw kritische vragen. Mogelijk is dat een van de redenen voor het een tijdlang uitblijven van een kritische analyse van het hypothese toetsend model. Leken en clinici zijn gevoelig voor het kader waarin gegevens gepresenteerd worden. Dit heet *framing*: Wat is meer: de winst van de bank dit jaar is een kwart miljard of 250 miljoen? Wie ging harder: hij knalde met een vaart van 15 km tegen een auto', of 'hij reed met een snelheid van 15 km tegen een auto'? Als effect van begeleiding van criminele jongeren in inrichtingen wordt vaak vermeld dat er winst is geboekt, omdat bij gelijkblijvende criminaliteit het geweld bij de misdaden afneemt (ex- directeur Rentray: H.

Lodewijks op Radio 1, 2009). Heeft begeleiding tijdens het verblijf ertoe geleid dat de criminelen hetzelfde bereiken, maar met minder moeite, met minder geweld?
Hilbert, (2012, p. 214) vat de vertekeningen bij informatieverwerking en beslissen samen en stelt een theoretisch kader voor: het *noisy memory channel framework*. We kunnen ons niet onttrekken aan ruis in ons geheugen, waardoor we waarnemings- en denkfouten maken. Het is uitgewerkt als in de signaal-detectie theorie. Het kader is even abstract en aspecifiek als in de benaming *defective information processing*.

Box Vertekeningen bij diagnosticeren, oordelen en beslissen

Een *bias* is volgens Hilbert een geobserveerde afwijking van een verwachting die berust op empirisch vastgestelde normen, een logische regel of wiskundig model. Hij vat een vertekening op als iets, waar we ons van moeten ontdoen en niet als een verschijnsel *sui generis* van onze alledaagse informatieverwerking.
Conservatisme houdt in dat we kansen schatten in zonder rekening te houden met extremen. We zien zo veranderingen niet aankomen, bijvoorbeeld de omvang van een brand.
Illusory correlation gaat over stereotyperen. Er worden relaties gelegd tussen twee verdelingen die empirisch niet samenhangen. We doen dat op basis van een waargenomen, veronderstelde of instant bedachte gelijkenis. Dit is een gevaar van hermeneutisch verbinden zoals dat zou kunnen plaatsvinden bij diagnose-behandel-combinaties (DBCs).
Placement verwijst er naar dat ik denk beter in te kunnen schatten hoe ik zelf ben en wat ik doe dan dat ik dat over anderen kan. Empirisch blijkt soms het tegendeel het geval.
Subadditivity is het verschijnsel dat ik twee elkaar uitsluitende componenten een grotere kans van voorkomen toeken dan één alleen.
Exaggerated expectation duidt erop dat we extremere waarden verwachten dan er zich feitelijk voordoen. We maken een probleem groter dan het is.
Confidence zegt dat we te veel vertrouwen hebben in ons oordeel.
Hard-easy wijst er op dat we de moeilijkheid van een taak of ernst van een stoornis matig kunnen inschatten. We denken bijvoorbeeld dat we bij een moeilijke taak onderpresteren en bij een gemakkelijke overpresteren.

De Ierse informatici Costello en Watts (2014) pakken het weer iets anders aan. Ze gaan er vanuit dat mensen kansverschijnselen begrijpen en een *probability theory* hanteren. We zijn dus rationeel (en niet dom). Het probleem is het omgaan met ruis en toevalsvariatie in het redeneerproces. Ze komen op deze wijze tot vier systematische vertekeningen in het redeneren over kans. Voor een deel overlappen ze met Hilberts analyse. Mensen blijven uit de buurt van extremen in hun oordelen over een kans van optreden van een gedrag of verschijnsel (*Conservatism).* Gemiddeld is de som van de schattingen dat gebeurtenissen 1...n zich voordoen groter dan de kans dat gebeurtenis 1 zich voordoet (*Subadditivity)*. Bij twee geordende gebeurtenissen waarbij de p (A) ≤ p (B) is de kans op samen optreden (conjunctie) van A en B ≤ dan de kans op zich voordoen van A. Dit is de *Conjunction fallacy*. Dit is het bekende voorbeeld van Linda die 31 jaar, vrijgezel, een 'personality' en slim is. Ze

heeft PhD behaald op een goede universiteit. Als student was ze zeer begaan met discriminatie en sociale rechtvaardigheid en deed mee aan ban-de-bom-demonstraties. Aan proefpersonen werd gevraagd hoe groot de kans is dat Linda een bankemployee (A) is of een bankemployee en actief in de vrouwenbeweging (A + B). De *Disjunction fallacy* is het spiegelbeeld. A en B doen zich voor als B zich voordoet, dus is de kans op A en B ≥ dan de kans op optreden van B.
We nemen genoegen met onze gebrekkige manier van informatieverwerken. We houden van eenvoudige oplossingen die werken. We stellen geen vragen waarop we moeilijke, complexe en met voorwaarden belaste antwoorden krijgen. We vermijden verwarring en onzekerheid en leven bij voorkeur in een harmonieuze, voorspelbare wereld. We nemen genoegen met de illusie van controle, denk aan de eerder genoemde vijfjaren plannen van de Sovjet planeconomie en de infrastructurele werken in Nederland. Planeconomie is door het werk van Kahneman een *contradictio in terminis* geworden. Het had beter casino-economie genoemd kunnen worden. Er werken zoveel krachten dat het resultaat een verrassing is. Het zijn probabilistische verschijnselen. Wie heeft de bankencrisis van 2008 zien aankomen en maatregelen voorgesteld? Er waren er zeker enkelen maar leken en experts luisteren eerder naar wat hun mening bevestigt dan ontkent. Onder druk worden we oppervlakkige informatieverwerkers en passen ons aan wat belangrijke anderen zeggen. Dit is in experimenten aangetoond maar toch is verlangen naar eenvoud soms zinvol. Het werkt af en toe en het is toereikend voor het dagelijks (over)leven.
Diagnostische en oordeelsfouten worden vastgesteld door ze te vergelijken met normatieve logische regels en statistische modellen. Gebruik daarvan zou de kans op bepaalde soorten fouten verkleinen, bijvoorbeeld onterechte conclusies uit premissen en onjuiste beweringen over relaties en effecten. We kunnen het niet van de regels winnen en slechts gelijkspelen: je kunt immers toepasselijke normatieve regels alleen navolgen, al het andere is onjuist. Afwijking van regels of modellen wordt opgevat als een diagnostische of oordeelsfout. We categoriseren bijvoorbeeld door op één in het oog springend kenmerk te letten, we zijn ongevoelig voor de incidentie van een verschijnsel of gedrag, we kunnen geen lineaire verbanden waarnemen in een scatter plot.

Een sceptische houding Die kan het gevolg zijn van HB onderzoek bij studenten en practici ontmoedigen. Er is opgemerkt dat de HB onderzoekers er op uit waren fouten te vinden en ze aan de kaak te stellen. Laten zien dat clinici ernaast zitten is *fun* zegt Van Dam in haar monografie *Fixatie op fouten* (1991). Dit soort onderzoek past goed in de Amerikaanse attitude dat zowel mislukking als succes in ieders leven thuishoren. James (1902, Nederlandse vertaling, 2010, p. 106) haalt de schrijver Robert Louis Stevenson aan. Hij is bekend door boeken als *Treasure Island* en *The strange case of Dr. Jekyll and Mr. Hyde*. Hij zegt het zo: *'...waartoe we ook bestemd zijn, we zijn niet bestemd tot slagen. Falen is het ons toebedeelde lot'*. Hij voegt er met optimisme aan toe: *'...onze opdracht is door te gaan met falen in een goed humeur'*. Stevenson was oorspronkelijk een calvinistische Schot maar

huwde een Amerikaanse en maakte zich mogelijk zo het Amerikaanse optimisme eigen. Wat platter: het lijkt ook op *bloopers* die sommigen graag in TV programma's bekijken.
Clinici, diagnostici en ook rechters en dokters en lekenbeoordelaars komen er niet best af. Ze hebben te veel/weinig vertrouwen, hanteren gemakzuchtige vuistregels, houden vast aan onjuiste vooringenomen standpunten, wegen informatie niet goed en houden zich niet aan van logische en statistische regels (Lilienfeld et al., 2003; Gilovitsch et al., 2002; Hogarth, 1987).

Samenvatting en conclusie
Meehls boek en meta-studies daarna maakten duidelijk dat statistische predictie superieur is aan klinische. Het werd ook aanleiding voor de analyse van klinisch voorspellen en oordelen. Dit leidde tot het *heuristics en biases* onderzoek door Tversky & Kahneman et al. Dit toonde dat leken en diagnostici beperkte informatieverwerkers zijn en simpele regels gebruiken om gedrag te voorspellen. Een lineaire combinatie van een aantal predictoren verslaat de clinicus meestal. Het gebruik van eenvoudige en soms onjuiste heuristieken, het vertekenen en selecteren van informatie kunnen een sceptische houding bij studenten en professionals veroorzaken.

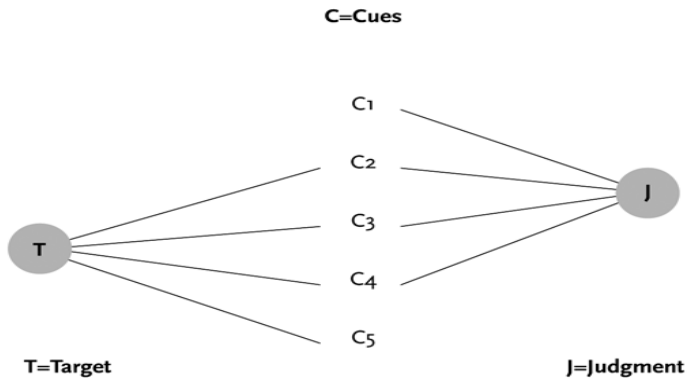
7. Reacties op de kritiek
Zoveel verwijten aan het adres van alledaags oordelen en klinisch voorspellen lokken reacties uit (Van Dam, 1991). De clinicus vraagt zich af of de kritiek terecht is en denkt dat hij er niet altijd naast kan zitten. Zoveel commentaar krijg ik immers niet. Was dat het geval dan had ik geen werk meer. En, bijna ieder redt zich toch in onze complexe wereld. Dat is een positieve reactie. Je neemt het oordeelsproces serieus als adaptatie aan de complexe wereld en vat klinisch oordelen niet op afwijking van logische regels of statistische modellen. Ik vermeld als voorbeelden het lensmodel van Brunswik (1952, 1958), de studie van intuïtie door Hogarth (2001, 2011), het werk van Klein (1998, 2003) over *Natural Decision Making* (NMD) en *fast* en *frugal heuristics* (Gigerenzer, 2000).

Brunswiks lensmodel (1952, 1958) is oorspronkelijk bedoeld om de waarneming te bestuderen aan de hand van realistische opdrachten. Hij ging het laboratorium uit en wees de studies daar af als ecologisch niet valide.


Egon Brunswik is de vader van het lensmodel voor de waarneming. Hij pleitte voor het gebruik van representatieve onderzoeksontwerpen. Al voor Meehl in 1954

het klinisch oordeel kritiseerde ontwikkelde hij zijn adaptieve waarnemingsmodel. Hij komt uit de Duitse centraal Europese Gestaltpsychologische traditie van de beschrijving en verklaring van de waarneming. Hij was een leerling van Karl Bühler, die weer beïnvloed werd door de fenomenoloog Husserl en door Von Brentano (Radler, 2015).

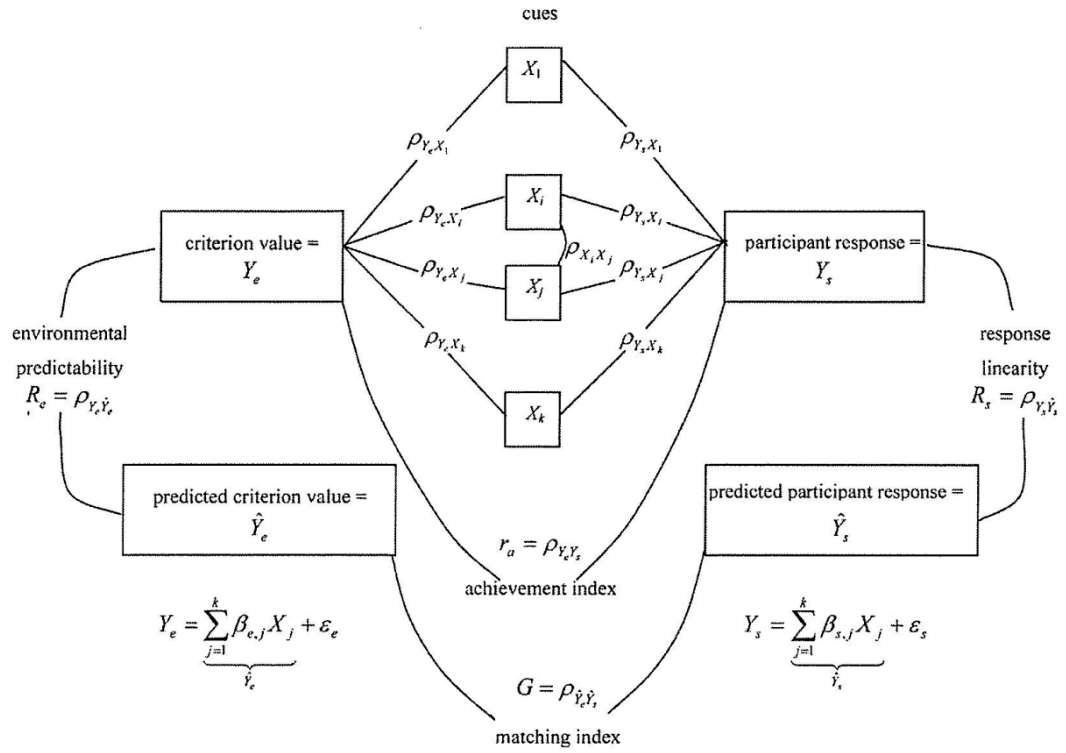
Het lensmodel beschrijft wat er gebeurt bij het waarnemingsproces van een object, gebeurtenis, persoon. Het is ook gebruikt bij het voorspellen van criteria (Figuur 1).



Figuur 1 is een eenvoudige weergave van het lensmodel. Het kan behalve voor de waarneming ook gebruikt worden om het klinisch oordeels- en het diagnostisch proces af te beelden. T is het doel, criterium (*target*), bijvoorbeeld succes op de middelbare school; de C 1 tot 5 zijn de *cues*, bijvoorbeeld IQ, motivatie, steun van de ouders, kwaliteit lerarenkorps, leeftijdgenoten. Sommige dragen niet bij, bijvoorbeeld (hypothetisch!) steun en aandringen van ouders maar die van leeftijdgenoten wel. J is het oordeel, de waarneming *(judgment*), de voorspelling, bijvoorbeeld de kans dat Jan binnen zes jaar het einddiploma gymnasium behaalt.

Een proefpersoon neemt een gebeurtenis of object waar en stelt vast wat het is door een aantal *cues* (aanwijzingen) te wegen. In het voorspellingsjargon zijn dat de predictoren. Sommige *cues* zijn niet beslissend, andere wel: ze verschillen met andere woorden in hun diagnostische bijdrage. Dat leert de waarnemer in de loop van de tijd want de werkelijkheid en andere waarnemers corrigeren hem af en toe als hij fouten maakt. Hij is een adaptief wezen dat van zijn fouten leert. Brunswik gaat ervan uit dat waarneming geen illusie is, waarmee we alles naar onze hand zetten. We zijn echter ook niet in staat om de werkelijkheid onmiddellijk en foutloos waar te nemen. Zijn opvatting wordt probabilistisch functionalisme genoemd. De nadruk ligt niet op de leek/professional als brokkenmaker, gebrekkige waarnemer, gemankeerde statisticus, slordige informatieverwerker of naïef realistische kentheoreticus. Hij wijst op hun vermogen zich aan te passen aan de grillige sociale en fysische werkelijkheid. Dat lukt niet altijd, want er zijn *wicked environments* en ook grillige onvoorspelbare cliënten. Bovendien heeft hij soms met onervaren waarnemers en raadgevers van doen. Figuur 2 bevat een uitgewerkte variant van het lensmodel met latente trekken. Bij zo'n eenvoudige weergave als in figuur 1 laat men het niet. In het model

zijn schatters van fouten opgenomen en als zodanig is het een uitwerking van het probabilistische karakter van onze waarneming.



Figuur 2 is een uitgewerkt, deftig lensmodel. Uit: Determinants of linear judgment door Karalaia, N. & Hogarth, R.M. (2008). *Psychological Bulletin, 134, 3,* 404-426 (p. 405). Het diagram is een uitwerking van het model van figuur 4.1 met latente variabelen. Het oordeel Ys is gemodelleerd als een lineaire functie van een reeks cues Xj, j = 1...k: Ys = ∑ ß.jX + εs, waarbij ßs.j de gewichten weergeven, die de diagnosticus aan de cues geeft en ε de foutenterm van Y is op de Xj. Het omgevingscriterium wordt gemodelleerd door s door e te vervangen: Ye = ∑ße.jX + εe.

De aanwijzingen met diagnostische waarde convergeren - de lens - naar een valide waarneming, oordeel, diagnose, bijvoorbeeld '... daar is een rel; daar komt een gevangenisopstand aan of dat is een persoon met wie ik langere tijd bevriend kan zijn'. De waarneming of diagnose is gemodelleerd als een lineaire combinatie van *cues*. Dhami en Harris (2001) pasten het toe om te achterhalen hoe artsen diagnosticeren. Ze gingen er niet van uit dat ze de *cues* slecht integreerden en foute vuistregels gebruikten, zoals in de HB traditie van Tversky & Kahneman. Ze observeerden en stelden vast dat artsen een snelle procedure hanteren met een stopregel. Ze zijn doelgericht: is het een longontsteking, verkoudheidje, aanstellerij, of nog iets anders en zoeken naar de belangrijkste *cue* op grond van hun kennis van ziektebeelden en ervaring. Als ze die *cue* hebben en voldoende denken te weten en zich zeker voelen stoppen ze. Het omgekeerde doet zich ook voor: 'er is niets aan de hand', zegt een cliënt maar de professional voelt weerstand dit te accepteren en gaat zoeken. Dit lijkt op de *fast* en *frugal heuristics* (Gigerenzer, zie verderop). Dat zijn snelle

en zuinige vuistregels die niet meteen tot foute waarnemingen, diagnoses en beslissingen hoeven te leiden.
Naast de rechttoe rechtaan optelling van *cues* is ook *een interactieve versie* denkbaar, waarbij rekening wordt gehouden met de interacties (correlaties) tussen de *cues*. Het lensmodel maakt onderzoek mogelijk naar vertekeningen (onnodige, irrelevante, onjuiste *cues*) en naar accuraatheid (verkeerde gewichten: de ßs). Het model heeft ruimte voor leren van feedback.

Box Achteraf weten we het: wetenschappelijk discutabel, maar waarom mag dat niet?

Brunswiks model kan gebruikt worden om na te gaan of een oordeel over een persoon verandert als informatie wordt toegevoegd. Zo lijkt het model op de Bayesiaanse aanpak. Achteraf kennis, 'met de kennis van nu' *(hindsight*) wordt een gemakzuchtige *heuristic* genoemd. Het is de tendens om achteraf te herstellen wat niet goed voorspeld is. In de wetenschap is pre-dictie verplicht en post-dictie gemakzucht. Het is een voorbeeld van zich niet aan het onderzoeksprotocol houden. Men kan echter *hindsight* (een foutieve regel in de Kahneman/Dawes benadering) ook opvatten als een bijproduct van het *updaten* van kennis. Dit is precies wat Nestler et al. (2012) deden. Ze onderzochten drie processen die *hindsight* effecten bij het oordelen over personen kunnen verklaren. De deelnemers moesten de Big Five scores voorspellen met en zonder *hindsight*. Ze vonden dat feedback (het toestaan van *hindsight*) ertoe leidde dat de deelnemers (1) cues gebruikten die valider waren en (2) meer cues gebruikten die met het criterium samenhingen. Deze studie laat zien dat feedback een diagnosticus kan helpen om dichterbij de juiste *cues* en met de juiste gewichten te komen.

Het denken van Brunswik sluit aan bij de Weber-Fechner traditie van het vinden van een samenhang tussen objectieve kenmerken van objecten en personen en onze waarneming van die objecten en personen. Het model vooronderstelt de mogelijkheid van waarheidsgetrouwe waarneming. Het is een lineair compensatoir model. Er wordt niet meteen aan interacties gedacht. Het kernidee is dat waarneming en oordeel adaptief zijn. De nadruk ligt niet op fouten: heuristieken en vertekeningen van de waarnemer en diagnosticus.
Karelaia en Hogarth (2008) deden een meta-studie op basis van 86 artikelen naar het lensmodel. Zij bevestigen eerdere bevindingen dat

'...linear rules can provide higher levels representations of both human judgment and task environments' (p. 420)

Met andere woorden als er interactietermen zijn, dragen ze niet veel bij tot het voorspellen van criteria. Als de *cues* uit de omgeving niet opgeteld kunnen worden dan is de voorspelling matig. Oordelaars, ook professionals hebben vereenvoudigde verwachtingen over de omgeving en geven de voorkeur aan enkele duidelijke *cues* met gelijk gewicht. Bij

een teveel aan *cues* worden ze onzeker en de verschillen in gewichten zorgen voor het uit balans raken van hun habituele strategieën.
Het lensmodel en de HB traditie laten zien dat we in een transparante, eenvoudige wereld willen leven. Het lensmodel laat de voorwaarden zien waaronder fouten - ook in onderzoek - kunnen optreden: een groot aantal onderzoeken met wisselende uitslagen, een groot aantal *cues* en teveel onderzoekers, oordelaars en clinici die ver van het objectieve criterium af staan, ecologisch niet valide onderzoeksopzetten en *inter-cue* redundantie: niet weten dat *cues* hetzelfde meten, maar ze wel elk een gewicht geven. Het is een creatieve prestatie om bij de vraag van een cliënt diagnostisch relevante *cues* te identificeren en de juiste gewichten te geven.

Intuïtie Dit onderwerp wordt vermeden door HB onderzoekers en voorstanders van statistische predictie. Toch vroeg Hogarth (2001), een ex-HB onderzoeker zich af wat intuïtie is, hoe het functioneert en of het onjuiste voorspellingen oplevert. Hij is er zich van bewust dat het geen populair thema is want wetenschap heeft rationaliteit, rigoureuze methodologie en statistische analyses hoog in het vaandel. Hij vond in de literatuur meer referenties bij 'illusies' en 'inzicht' dan bij 'intuïtie'.
Intuïtief handelen is niet precies te definiëren. Het verwijst naar moeiteloos begrip van bijvoorbeeld het gedrag van iemand. Het sluit de overtuiging in dat er een vermogen is om kennis te verwerven zonder stapsgewijs rationeel te denken en langs logische weg conclusies te trekken. Mensen vellen oordelen en trekken conclusies zonder bewuste, gearticuleerde en verdedigbare stapsgewijze processen. Ze meanderen en voelen zich comfortabel bij een intuïtieve conclusie. Het proces verloopt snel, zonder overleg en wordt niet achteraf ge(re)construeerd. Het lijkt op Brunswiks opvatting over waarnemen: een organisme verwerft een waarneming van een object of persoon: 'dit is Jan, zij is ziek, hij is kwaad' door spontaan *cues* te integreren die met het waargenomene te maken hebben: bijvoorbeeld, een lange magere man met baard; ze ziet bleek en beweegt langzaam, hij praat snel en met flikkerende ogen en heftige gebaren. Het conclusie uit de waarneming wordt niet gerechtvaardigd. Het is automatisch, doelgericht, snel en bevat informele weging van een paar *cues.* Hampton (2012) legt uit dat intuïtie kan afwijken van conclusies op grond van logische regels maar ze is plausibel, adaptief en creatief. Ze kan bijvoorbeeld concepten representeren met prototypische eigenschappen en zonder logische definities. Ze leidt soms tot nieuwe conceptuele combinaties: geleid door intuïtie denkt men gemakkelijker *out of the box* dan met rigoureuze methodologie (het SMG: Streng Methodologisch Geweten).
Intuïties komen niet uit de lucht vallen. Er is een voorraad aan kennis over een onderwerp in het spel ook al kan de persoon dat niet precies articuleren. Een goede voetballer kan niet verklaren waarom hij het doet zoals hij doet, denk aan een schitterende steekpass, ook al heeft hij uitentreuren van de trainer gehoord dat hij 'vaste patronen' moet oefenen. Ook bij wiskunde speelt intuïtie een rol (Lakoff & Núñez, 2000). In elke wetenschap moet men een hypothese kiezen met een kans op succes. Dat gebeurt vaak niet door een deductief logisch

geformuleerd aantal stappen af te wikkelen. Er is ruimte voor intuïtie omdat een persoon weliswaar een geheel is maar ook beschikt over verschillende en complexe informatie verwerkende systemen die los van elkaar kunnen opereren en niet rationeel en theoriegeleid zijn.

Box Intuïtie

Intuïtie heeft meer dan een gezicht. Het is niet alleen een slag in de lucht die geen tegenspraak duldt. Filosofen van Plato tot Merleau-Ponty en Heidegger hebben kenmerken van gedrag en kennen gedefinieerd zonder theorie en toetsing: het zijn hun grondintuïties, hun interpretaties van voorgaande filosofen en van de geschiedenis. Darwin is begonnen met een kerngedachte, een intuïtie en die bracht hem er toe de wereldzeeën te bevaren met de Beagle. En zelfs Higgs heeft zijn deeltjes bedacht terwijl ze niet 'bestonden'. Dit toont dat intuïtie niet iets is waar we voor uit moeten kijken of bang voor moeten zijn. Niettemin heeft de Cartesiaanse rationalistische Westerse traditie en de opkomst van de empirische wetenschappen gezorgd voor een logisch positivistische en empirisch-analytische aanpak. Die gaat gepaard met aversie tegen intuïtie. In tekstboeken komt men haar niet tegen terwijl vrijwel elke theorie begint met een intuïtief idee over een relatie, een mechanisme, een factor, een effect. De Groot en Gigerenzer zijn uitzonderingen. De eerste gebruikte de denkpsychologie van Selz en de kenleer van Popper om het vormen van theorie en hypothesen te beschrijven en wegen te vinden om te toetsen. De tweede staat dichtbij Brunswiks probabilistisch functionalisme en erkent de beperkingen van het experimenteren en theoretiseren volgens logisch positivistische voorschriften.
De houding ten opzichte van intuïtie verandert. Gore en Sadler-Smith (2011) onderscheiden *intuiting* van *intuition*. Het eerste bestaat uit een reeks domein-overstijgende en -specifieke cognitieve en affectieve mechanismen en processen van *intuiting*. Heideggers concept *Gestimmtheit* is een voorbeeld want het sluit zowel stemmingen als cognitieve referentiekaders in. Domein-overstijgende algemene mechanismen verwijzen naar fundamentele processen van leren, redeneren en oordelen. Ze worden automatisch in stelling gebracht door taakkenmerken, bijvoorbeeld, complexiteit, risicovolheid en onzekerheid van taken. Deze lokken spontane processen uit die resulteren in (a) complexe domein-relevante schema's op basis van impliciet en expliciet leren (b) zicht hebben op affectieve gegevens die lichamelijk zichtbaar kunnen zijn en (c) snelle oordelen, soms inclusief de fouten volgens de HB traditie. De specifieke processen worden eveneens automatisch in stelling gebracht door (a) steeds terugkerende kenmerken van het specifieke domein en (b) door geleerde patronen en prototypes. Het tweede: *intuition* is een uitkomst, een resultaat. De auteurs onderscheiden primaire typen die teruggaan op informatie verwerkende mechanismen in domeinen van probleemoplossen, creativiteit en moreel en sociaal oordelen. De secundaire typen zijn samengestelde vormen die tot uitdrukking komen in specifieke beroepsdomeinen en *settingen*, bijvoorbeeld ondernemers-, bankiers- , dokters- en psychologenintuïtie en gevoel voor de markt, ambtelijke verhoudingen, sfeer op de werkvloer en machtsverhoudingen.

De onderscheidingen en uitwerkingen klinken aannemelijk. Er kan een volgende stap gezet worden het operationaliseren: het werkendeweg ergens achter komen door empirisch te specificeren en daarna te meten. De auteurs articuleren het probleem maar werken die stap niet uit. Intuïtie empirisch specificeren/realiseren is niet eenvoudig omdat het niet zo maar even experimenteel opgeroepen en gemanipuleerd kan worden. Er zijn zelfrapportages en documenten nodig die min of meer spontaan tot stand gekomen zijn. Het vastleggen zelf is een lineair, dat wil zeggen stapsgewijs proces en dat is vreemd bij intuïtie. Daar gaat het gaat om onwillekeurige, springerige automatische processen. Gore en Sadler-Smith honoreren de tijdgeest door de basis te zoeken in evolutionaire processen en neurologische correlaten van *intuiting en intuition*. Ze hebben een nuttig onderscheid

gemaakt dat al bekend was uit andere onderzoeksgebieden, bijvoorbeeld van het proces-product onderscheid, het schatten van diagnostisch waardevolle kenmerken van een object of verschijnsel: één of een aantal beslissende cues en algemene en specifieke mechanismen bekend uit intelligentieonderzoek.

Het organisme heeft er miljarden jaren over gedaan om al die systemen op basis van toeval, door adaptatie en schade en schande te verwerven. Ze zijn een gevolg van de evolutie en zijn ontogenetische ontwikkeling. Ze zijn overgeleerd, verlopen automatisch en zijn functioneel autonoom. We leren door associaties en het toevallig samen voorkomen van stimuli en gedrag. We leren het beste in omgevingen met goede leerstructuren en daar blijven onderwijskundigen naar zoeken. Terugkoppeling van de omgeving is nodig om waardevolle intuïties op te bouwen. Deze omgevingen kan men opzoeken en proberen te ontwerpen. Hogarth zegt dat intuïties domeinspecifiek zijn als gevolg van evolutie en leren. Ze vergen dat je goed om je heen kijkt, vrijuit speculeert en generaliseert maar ook toetst aan feiten. De auteur beweegt in de richting van de ecologische benadering van waarnemen. Fouten zijn niet alleen te wijten aan de oordelaar en clinicus. Ze hebben ook te maken met *wicked environments* en dito personen. Immers, niet iedere persoon is even voorspelbaar. Kahneman en Klein (2009) erkennen dat en maken onderscheid tussen *predictable environments* (ze noemen *medicine* en *fire-fighting*) en *low validity environments* (bijvoorbeeld *individual stocks* en *long-term forecasts of political events*).

De diagnosticus aarzelt of geeft liever niet toe door zijn opleiding en training dat zijn intuïtie een bron van kennis is over de cliënt. De HB traditie heeft de zwakten van het spontane oordeel blootgelegd. Niettemin is afwijken van de rationalistische weg aanvaardbaar. Larkin heeft bijvoorbeeld in samenwerking met Simon (1980) laten zien dat ervaren fysici hun intuïtie gebruiken om snel een interne representatie van een probleem te maken waarbij ze vrijuit putten uit hun fysische kennis en ervaring. Dit liet ook verschil tussen beginners en experts zien. Beginners zaten meer aan de logische structuur vast en kwamen minder vaak tot een werkbare oplossing. Klein (2003) heeft de rol van intuïtie laten zien op een afdeling neonatologie. Ervaren verpleegkundigen pikten subtiele *cues* op bij pasgeborenen, bijvoorbeeld huidskleur als een teken van metabolische complicaties. Als een diagnosticus zijn intuïtie laat spreken dan kan dat in een team met als doel een rijke, vernieuwende probleempresentatie te vormen waarin de kenbronnen (alledaagse impliciete, wetenschappelijk theoretische en alternatieve vernieuwende) en oriëntaties (individuele verschillen, ontwikkeling en sociale context) en het empirische kennisbestand van de psychologie een rol spelen.
Hogarths lezing van intuïtie is gebaseerd op de informatieverwerking. Dat is het psychologisch niveau op de schaal van Simonton (2004). Chassy en Gobet (2011) stellen voor haar te verbinden met het biologisch niveau. Ze gebruiken het concept cel clusters *(cell assemblies*) dat verwijst naar adaptieve processen in de hersenen bij het parallel en

opeenvolgend verwerken van informatie. De neurale netwerken die verantwoordelijk zijn voor cognitie reorganiseren zich voortdurend. Daardoor is het mogelijk om meer objecten en verschijnselen te herkennen. Object-betrokken informatie is vaak onmiddellijk en in één stap verwerkt. Dit lijkt op Gigerenzers *Take the First and the Best* strategie bij beslissen (zie verderop). Chassy en Gobet vatten emoties op als versnellers van het coderen van informatie. Emotionele betrokkenheid maakt niet altijd blind. Het brengt een herkenningsproces in stelling en bewaakt onze belangen (Frijda, 1986).

Natural Decision Making Naast de HB traditie en statistische modellen is er het natuurlijk, alledaags beslissen door experts: *Natural Decision Making* (NDM). Dit zagen we al in het werk van De Groot (1946/1978) en van Chase en Simon (1973) over schakers. Experts beslissen onder tijdsdruk, maken gebruik van veel en hoog-niveau kennis en van goed getrainde vaardigheden. Men kan denken aan brandweerofficieren, schakers, personeel op boortorens en soldaten/officieren in gevecht. Men ging er rationalistisch van uit dat ze een aantal opties formuleerden en vergeleken. Het bleek dat ze *tacit knowledge* - niet gearticuleerde kennis - gebruikten, bijvoorbeeld hoe de vlammen zich gaan verspreiden en tekenen van het ineenstorten van een gebouw. Ze begonnen *niet* met een aantal opties, maar met één. Dit is de *Recognition-Primed-Decision* (RPD) strategie genoemd. Ze pasten de strategie meteen aan als die onjuist bleek (Klein, 1998, 2003). Je kunt je voorstellen dat diagnostici in de crisisopvang zo te werk gaan. Ze hebben de tijd niet om een uitvoerig onderzoek te doen of uitvoerig de documentatie over de cliënt te raadplegen.
Experts moeten vaag gedefinieerde doelen behalen onder tijdsdruk met beperkingen vanuit hun organisatie onder zich wijzigende omstandigheden en met wisselde hoeveelheden feedback. Soms doen ze het desondanks goed. Onderzoekers willen achterhalen hoe hen dat lukt. Zij gebruiken de *availability heuristic* soms met succes. Het is niet bij voorbaat onjuist om zulke *short cuts* te gebruiken. Er gaat ook wel eens iets mis, bijvoorbeeld bij het aanbieden van financiële producten, het laten landen van vliegtuigen en het beginnen van oorlogen tegen terreur. NDM studies stellen in staat protocollen te maken, maar experts lijken toch vooral op snel op hun doel af te gaan zonder uitvoerige overwegingen vooraf. Als ze achteraf een protocol moeten maken is dat geen intuïtief maar een lineair verhaal: een stapsgewijze achteraf reconstructie.

Fast en *frugal heuristics* Gigerenzer sluit aan bij Brunswiks ecologische benadering. Hij laat zien dat wat in de HB traditie een zonde is onder bepaalde omstandigheden het beste is om te doen. Onder sommige condities verslaan snelle en spaarzame, zuinige *(fast* en *frugal*) heuristieken van experts en beginnelingen de lineaire modellen. Een eerste voorbeeld is herkenning. Als men Amerikaanse studenten vraagt welke stad meer inwoners heeft San Diego of San Antonio beantwoordt twee derde dat correct. Duitse studenten gaven tegen de 100% goede antwoorden. De vuistregel lijkt: *'...if one of two objects is recognized more or easier than the other, then infer that the recognized object has higher values with respect to*

the criterion' (Goldstein & Gigerenzer, 2002, p. 76). Zo'n regel zal niet altijd gelden maar is soms effectief vanwege zijn ecologische rationaliteit. Dat wil zeggen er is informatie uit de natuurlijke omgeving die personen zonder nadenken gebruiken. Een tweede voorbeeld is het *Take The Best* heuristiek (TTB). Hierbij komen opeenvolgende *cues* beschikbaar en de eerste onderscheid makende aanwijzing wordt gekozen. De rest laat men links liggen. Als iemand moet beoordelen welk van twee steden de meeste inwoners heeft dan kijkt hij naar een meest voor de hand liggende cue, bijvoorbeeld welke stad heeft een topuniversiteit of een voetbalkampioen, Utrecht en Amsterdam dus. Mensen gebruiken deze heuristieken omdat ze adaptief en snel zijn. Ze zijn niet altijd fout en irrationeel (Katsikopoulos et al., 2008; Katsikopoulos, 2009).

De HB traditie, de normatieve-regel-benadering en de optelling van voorspellers tonen de diagnosticus als een brokkenmaker. In de benadering van Brunswik wordt nadruk gelegd op wat de waarnemer en oordelaar *doen* en niet of zij dat goed of fout doen. Hij kan leren van informatie uit de omgeving als die een leerbare en herkenbare structuur heeft en van anderen als die van wanten weten: de oordelaar en clinicus zijn adaptief.

Samenvatting en conclusie

Brunswik beschouwde de waarnemer en oordelaar niet als brokkenmaker. Hij onderzocht waarnemen onder ecologisch realistische condities en ging na welke *cues* worden gebruikt en hoe ze samengevoegd worden om tot een oordeel of voorspelling te komen. Hij formuleerde het lensmodel voor de waarneming van gebeurtenissen en verschijnselen. Dat is op het klinisch oordelen en prediceren toegepast. De omgeving wordt opgevat als drager van een structuur waar de waarnemer steeds dichterbij kan komen door leren, onderwijs en door behulpzame en ervaren waarnemers. Zijn prestatie hangt af van de *cues* (het aantal, de relaties, hun onderscheidend vermogen, hun ambiguïteit) en van de waarnemer die diagnostisch informatieve *cues* ziet*,* ze consistent integreert en expertise op een domein verworven heeft. De ene omgeving is de andere niet: er zijn *wicked* en *predictable environments*. De beurs is zo'n *wicked environment*; daar wordt gegokt en nauwelijks rationeel gehandeld volgens Kahneman.

Hogarth begon als HB onderzoeker maar bekeerde zich tot onderzoek naar intuïtie. Hij argumenteerde dat intuïtie bestaat al is ze lastig grijpbaar, kan worden geoefend en domeinspecifiek is. Intuïtie werkt vooral op een gebied waar je veel vanaf weet en veel op geoefend hebt. Ons informatieverwerking systeem is biologisch oud en bevat meer dan rationele calculus, rationele modellen en statistische en logische regels. Ze zijn het topje van de ijsberg. Ja, dat beweerde Freud altijd al.

Los van HB onderzoekers en Brunswik werden expertoordelen en handelingen onder moeilijke omstandigheden bestudeerd bij brandweerlieden, soldaten in gevecht, spoedeisende hulp artsen, intensive care verpleegkundigen en schaakgrootmeesters (*Natural Decison Making*). De experts gebruiken eenvoudige maar effectieve strategieën, bijvoorbeeld de *Recognition-Primed-Decision*. Ze ruilen de eerste strategie onmiddellijk in voor een andere als die niet werkt. Gigerenzer richt zich tegen de HB traditie en laat zien dat

onder bepaalde condities andere bepaalde heuristieken tot zinvolle oordelen leiden, zoals *Take The Best*: pak de eerste de beste aanwijzing want die is soms doorslaggevend en de herkenningsheuristiek. Als je tussen twee moeten kiezen, neem de bekendste, waar je het meest van weet of die je het vaakst gehoord hebt.
De diagnosticus ervaart de ecologische benadering als natuurlijker dan de HB traditie, maar deze maakt hem behoedzaam voor voetangels en klemmen in zijn waarnemen, oordelen en diagnosticeren. Dit laat onverlet dat hij op basis van veel kennis een rijke probleembeschrijving mag maken en daarbij intuïtie toelaat.

8. Combineren diagnostici informatie niet-lineair?
Bij lineaire modellen tellen we variabelen op. Nogal wat clinici verzetten zich daar tegen, want niet elke predictor is even belangrijk voor het probleem. Bovendien kunnen voorspellers interacteren dat wil zeggen dat het gewicht van de ene predictor afhangt van de waarde op een andere. Meehl suggereerde - hij was ook clinicus en therapeut - dat informatie in zijn en andere hoofden niet-lineair gecombineerd werd. De regressieformule is compensatoir: lage of hoge waarden op de ene predictor kunnen gecompenseerd worden door hoge of lage waarden op een andere. Een laag IQ kan gecompenseerd worden door een hoge motivatie. Clinici gaan uit van relevante combinaties van informatie, bijvoorbeeld van elkaar beïnvloedende kenmerken van cliënt en sociale omgeving. Het gewicht van het ene, bijvoorbeeld een matige intelligentie, kan gecompenseerd worden door een rijk en veeleisend 'Havo is geen optie' gezin.
Expertprotocollen worden gebruikt om na te gaan of diagnostici niet-lineair voorspellen. Ze zeggen bijvoorbeeld een conjuncte regel te volgen. Dat wil zeggen de bijdrage van een voorspeller bijvoorbeeld IQ aan een criterium afhangt van een bepaalde waarde op een motivatieschaal. Het is mogelijk dat diagnostici de compensatoire en conjuncte combinaties aanpassen aan de afzonderlijke cliënt. Er is weinig navraag gedaan bij diagnostici om na te gaan welk proces ze spontaan doorlopen en welke combinatieregels ze hanteren. Dat is tijdrovend kwalitatief onderzoek en schiet er dus gauw bij in. Er is wel nagegaan welke combinaties, lineair of niet-lineair, het diagnostisch resultaat *gemiddeld* het best voorspellen. Het gaat daarbij niet om de beschrijving van het *proces* bij één of de gemiddelde diagnosticus.

Welke combinatie voorspelt het product het best? Er zijn verschillende combinaties van predictoren om het product - de diagnose - te voorspellen. Dat zijn het behoren tot een categorie, het voorspellen van een waarde van het criteriumgedrag of de lucratiefste beslissing gegeven een aantal opties. Een voorbeeld is een klassieke studie van Hoffman en Wiggins (1968). Ze gebruikten gegevens van Meehl om de positie van cliënten te voorspellen op een elf-puntschaal: 1 = zeer neurotisch en 11 = zeer psychotisch. Negenentwintig oordelaars (13 experts en 16 beginners) moesten 861 cliënten onderbrengen in een geforceerde normaalverdeling met gemiddelde 5.5 en SD 2.5. De

informatie om de waarden toe te kennen was het MMPI-profiel op 11 schalen van de 861 deelnemers. Drie manieren van combineren werden gebruikt: lineair zonder/met interactie, kwadratisch en een *sign* systeem van diagnostische tekens en aanwijzingen. De lineaire regel zonder interactie kent alleen hoofdeffecten in termen van de variantieanalyse. Y moet voorspeld worden met $X_1...X_n$ als predictoren en b's slaan op de gewichten:

De lineaire regel (zonder interactie): $\hat{Y} = b_1X_1 + b_2X_2 + \ldots bnXn$
De lineaire + interactie regel: $\hat{Y} = b_1X_1 + b_2X_2 + b_3rX_1X_2$
De kwadratische regel (zonder interactie): $\hat{Y} = b_1X^2_1 + b_2X^2_2 + \ldots bnX^2n$

De vergelijking kan aangepast worden voor *moderated regression.* Daarbij is de relatie wel lineair maar verschillend voor twee groepen. Dat kan de helling *(slope*) betreffen: de relatie is in de ene groep sterker dan in de andere, of het intercept: het gemiddelde van de ene groep is hoger dan dat van de andere. Als blijkt dat de correlatie tussen inkomen en opleiding hoger is bij mannen dan bij vrouwen dan is er bij mannen een hoger regressiegewicht. Hun opleiding voorspelt beter wat ze verdienen dan die van vrouwen. Sekse is in dat geval een moderator variabele, een variabele die de correlatie drukt.
Het lijkt intuïtief aannemelijk dat optellen niet juist is. Stel men moet de ernst van een stoornis van twee cliënten bepalen op basis van twee scores. Beiden hebben hetzelfde gemiddelde waarbij de ene cliënt ongeveer gelijke scores behaalt en de ander een extreem hoge en een extreem lage. Volgens de lineaire regel zonder interactie zijn beiden even gestoord. Dat het kan anders uitpakken bij een niet-lineaire, bijvoorbeeld, conjuncte regel: de een kan dan gestoord zijn en de ander niet. Ten slotte werd een *sign* regel gebruikt met 70 klinische aanwijzingen. Ze werden door clinici geformuleerd om neurotici van psychotici te onderscheiden met behulp van MMPI schaalscores. Een voorbeeld van een *sign* is het verschil tussen Pt score (psycho-astenia) en de Sc (schizophrenics) score.
Uit de steekproef van 861 werden subgroepen getrokken om de drie waarden uit te rekenen. Deze werden in een nieuwe steekproef gebruikt. Er was geen empirische steun dat de interactie, *sign* en kwadratische regels de diagnoses beter voorspelden dan de lineaire. Deze voldeed beter dan de kwadratische bij 23 van de 29 oordelaars en beter dan de *sign* regel bij 17 van de 29 oordelaars. Een interactieve combinatie ligt intuïtief wel voor de hand en de kwadratische lijkt soms aantrekkelijk want die vergroot de verschillen tussen de voorspellers. Het product, het criteriumgedrag werd er gemiddeld *niet* beter door voorspeld. Het idee is overigens niet opgegeven. Ganzach (1995) gebruikte dezelfde data om te zien of er toch sprake was van niet-lineaire combinaties. De verschillen waren ook in zijn studie klein. Als de predictoren negatief correleerden was een niet-lineair model iets beter. Garb (1995) gebruikte computersimulaties en ook die lieten zien dat een niet-lineair model iets beter is als de voorspellers negatief gecorreleerd zijn.
Dougerty en Thomas beweren in 2012 nog eens dat het niet realistisch in om aan te nemen dat concrete gedragingen voorspeld kunnen met eenvoudige lineaire *input-output* functies.

Gedrag is immers een emergente uitkomst van complexe neurologische, cognitieve, sociale en omgevingsstimuli. Om emergente (dynamisch, grillig opkomende en organisch groeiende) gedragskenmerken te modelleren kozen ze voor een minimalistische benadering. Ze stellen een algemeen monotoon model voor dat toestaat om relaties tussen predictoren en criteria (tussen gedragingen onderling en gedrag-omgeving eenheden) af te beelden zonder restricties. Monotoon betekent dat er geen specifieke vorm van de functie wordt voorgesteld. Het gaat slechts om een monotone toe- of afname. Dit model kan wellicht het grillige gedrag en de *wicked*, stuurloze, alle kanten opgaande sociale omgeving wat beter afbeelden en als er een lijn in zit, wordt die ontdekt. Zij claimen meer succes voor hun model dan voor de overige en dat is te verwachten want de eisen zijn minder streng.

Het lineaire model blijft Het model is misschien wel realistisch maar zal het lineaire model niet verdringen want dat is eenvoudig, *parsimonious* en voldoet aardig. Al gebruiken individuele clinici naar hun zeggen in hun hoofd vele en verschillende combinatieregels het product: hun diagnoses, het voorspelde criterium van de *gemiddelde clinicus* kan afgebeeld worden als een lineaire combinatie van predictoren vooral als deze positief correleren. Het is vanwege de oordeelsgewoonte om te kijken naar positief samenhangende predictoren niet te verwachten dat diagnostici gebruik zullen maken van predictoren die een negatieve correlatie met een criterium vertonen, ook al zijn ze op zich even informatief als de positieve.

Protocollen gemaakt door klinische experts Gebruikmaken van de ervaring van professionals en op grond daarvan protocollen schrijven lijkt een zinvolle manier om de diagnosticus te helpen. Een gemiddelde ervaringsregel van een aantal experts kan bijvoorbeeld luiden: als meer dan vier schalen van de MMPI een waarde hebben van meer dan 70 dan moet er ingegrepen worden. Einhorn et al. (1979) vonden een correlatie van r = .46 tussen de regel op basis van een protocol met wat clinici zelf deden in vergelijkbare omstandigheden. Het komt erop neer dat de regel van het protocol niet consistent gevolgd wordt door clinici en diagnostici. De geprotocolleerde regel correleerde lager met het criterium dan de lineaire regressieformule: r = .46 tegenover r = .79.
Het lijkt gerechtvaardigd om niet af te wijken van de optelling van een beperkt aantal predictoren om een criteriumgedrag in een steekproef te voorspellen. Als er evident een afwijking is bij een cliënt, houdt de diagnosticus daar rekening mee, denk aan Allports professor die niet naar de film ging omdat hij de Nobelprijs kreeg. Dit heeft ook als gevolg dat de diagnosticus zich terughoudend kan opstellen als hij complexe modellen krijgt voorgeschoteld want ze zijn gemiddeld genomen niet beter dan de lineaire.

Samenvatting en conclusie

Meehl vermeldde al zijn indruk dat de individuele clinici diagnostische informatie of voorspellers niet-lineair combineren. De ene predictor is de andere niet en mogelijk hangen

ze samen. Voorspellers dragen verschillend bij, ze interacteren en een variabele kan criteriumgedrag voorspellen als die een drempelwaarde overschrijdt. Bijvoorbeeld, alvorens creatief te kunnen zijn is een IQ van > 110 nodig. Dit lijkt intuïtief geen rare gedachte. Je moet immers wel van wanten weten over een domein. Uitgaande van het product, de diagnose, hebben onderzoekers lineaire en niet-lineaire modellen gepast. Er was minder passing voor de kwadratische regel en voor een interactief model dan voor het lineaire. Een model met *signs* die experts aanbevolen hadden, leverde ook geen betere voorspelling van het diagnostische resultaat. Het gaat hierbij om het gemiddeld resultaat van een aantal diagnostici. Dat is tot stand gekomen doordat de een beter en de ander slechter diagnosticeerde met niet-lineaire modellen. Gemiddeld genomen kan men zeggen dat een lineaire combinatie van een beperkt aantal voorspellers het best werkt om de predictie van de gemiddelde diagnosticus te beschrijven. Het is herhaald gevonden, het model is dus robuust. Na onderzoek bleek wel dat een niet-lineair model iets beter werkte als de voorspellers negatief correleerden maar dat past weer niet in onze natuurlijke oordeelsgewoonte. Leken en professionals zoeken naar voorspellers die positief samenhangen met het criterium.

9. Reflectie en evaluatie

Het onderscheid tussen nomothetisch en idiografisch is de oorsprong van de tegenstelling tussen statistisch en klinisch voorspellen. Volgens Windelband en Allport was het geen tegenstelling. Het is aan Meehls 'verontrustend boekje' te danken dat het er een geworden is. De tegenstelling is toegespitst op het organiseren van diagnostische informatie in het hoofd versus met behulp van een lineaire regressie formule. De statistici waren de winnaars. En *the winner takes it all* ook al sputterden clinici tegen dat een formule nooit definitief is en de structuur van de problemen complexer is dan de formule afbeeldt. Holt (1986, p. 386) neemt '... *in non clinical psychologists a patronizing attitude to the clinical assessment enterprise…*' waar. Dawes gaat zo ver dat hij het gebruik van klinisch voorspellen *onethisch* noemt. Deze uitspraak bevat het risico van de *naturalistic fallacy*: van wat *is*, besluiten tot wat *moet.* Hij ziet dat gevaar maar (Dawes, 2005, p. 1253, zijn onderlijning) zegt desondanks*: '... we ought to use the relevant statistical prediction if one is available'.*

De bewering dat we logisch niet de stap van 'hoe iets is' naar 'hoe iets moet' kunnen maken is van David Hume, de Engelse sceptische empirist. De filosoof Moore (1903) noemde dit de *naturalistic fallacy* en die is verboden in de filosofie. Of dat overgenomen moet worden in de psychologie wordt betwijfeld. Brinkmann (2009) wijst erop dat er

(a) plaats is voor waarden in de wereld van de feiten,
(b) veel van ons sociaal gedrag en denken afhangt van volgen van normatieve regels die verdedigd worden met waarden,
(c) waardeoordelen over menselijk functioneren feitelijk kunnen zijn en

(d) verbinding tussen feit en waarde nodig is om veel gedrag te begrijpen. Mensen nemen bijvoorbeeld eenvoudigweg verplichtingen waar ten opzichte van sociale en institutionele contexten. Dit zegt dat iets is (een feit) maar anders moet (vanwege een aangehangen waarde).

We vermengen ook in hard empirisch onderzoek gemakkelijk neutrale en evaluatieve karakteriseringen van gedrag. Hoe vat u bijvoorbeeld 'hij behoort tot de 5% meest extraverte Nederlanders', 'de gemiddelde psycholoog (hypothetisch!) heeft een IQ van 108' op? De vermenging doet zich (gemakkelijk) voor in onderzoek naar het verbeteren van gedrag. Het is echter te verdedigen om als diagnosticus het onderscheid voor ogen te houden. Hij is geen moraalridder maar is zich ervan bewust dat inzicht in eigen en algemeen aanvaarde morele opvattingen nodig is om te begrijpen waarom het gedrag van een cliënt (niet) wenselijk is of gevonden wordt.
Het HB onderzoek van Tversky en Kahneman heeft de broosheid van ons oordelen en voorspellen zichtbaar gemaakt. We gebruiken vertekeningen en gemakzuchtige heuristieken zoals *availability* en representativiteit. We hebben geen zicht op de aanvangskansen van verschijnselen en gedragingen. Deze studies gaan over oordeelsfouten in het licht van normatieve, voorschrijvende modellen. Het HB onderzoek biedt steun aan statistische predictie want het hoofd maakt vaker oordeels- en voorspellingsfouten dan de formule. Het onderzoek gaat er niet over hoe het klinisch oordeel in elkaar steekt maar over hoe onhandig clinici en leken te werk gaan. Meta-studies laten zien dat de statistische voorspelling gemiddeld genomen weliswaar tot minder foute diagnoses leidt maar in ongeveer de helft van de gevallen is het resultaat gelijk.
Het kritische HB onderzoek lokte reacties uit. Waarom redden we ons in het dagelijks leven redelijk als we zulke brokkenmakers zijn? Brunswik lanceerde zijn lensmodel om te beschrijven wat er gebeurt bij het waarnemen van objecten, verschijnselen en gedrag. Hij vat de diagnosticus op als een adaptieve waarnemer die *cues* selecteert uit de (sociale) werkelijkheid rondom hem en integreert als in een lens. Zo komt hij tot een zo accuraat mogelijke waarneming of diagnose. De waarnemer kan de verkeerde *cues* gebruiken en ze een onjuist gewicht geven maar hij leert door de corrigerende werking van de werkelijkheid en door andere waarnemers. Hij komt zo al doende en lerende dichterbij het object, verschijnsel of gedrag: probabilistisch functionalisme. De werkelijkheid geeft soms gemakkelijk en soms moeizaam haar structuur prijs. Onder sommige condities, bijvoorbeeld *wicked environments* is een accurate waarneming, voorspelling, diagnose onwaarschijnlijk (de beurs, de kans op onderzoekssubsidie). In andere is dat beter te doen, bijvoorbeeld voorspellen of Jan het op het MBO zal redden.
Daarnaast wijzen auteurs erop dat een simpele strategie soms de beste is. Soms is één aanwijzing genoeg. De eerste de beste aanwijzing is raak. Als bijvoorbeeld de validiteit van de opeenvolgende *cues* exponentieel afneemt dan is de eerste *cue* inderdaad de klap die een daalder waard is: ½, ¼. ⅛.... Gigerenzer pakte de HB research direct aan. Hij zocht naar voorbeelden waarbij de 'foute HBs' tot een beter oordeel leiden dan de statistische, fouten

minimaliserende, modellen. Hij toonde het bestaan van *fast* en *frugal heuristics* aan. Nu gelden die onder specifieke condities maar ze laten zien dat HB onderzoek niet alles is. Gigerenzer noemt HB onderzoek zelfs een *fallacy in theorizing*: het is een *one word explanation* van een verschijnsel. Het beeld is dat de HB onderzoekers fouten opsporen en zoeken. Het is *fun* om zwakten van anderen te laten zien. Duitse (Brunswik, Gigerenzer), Angelsaksische (bijvoorbeeld Costello & Watts), Nederlandse onderzoekers (De Groot) en de filosoof en ontwikkelingspsycholoog Merleau-Ponty verdedigen waarnemen en oordelen als feilbare vermogens die onder geschikte condities tot geldige uitspraken leiden.

Onderzoekers van *Natural Decision Making* (NDM) kunnen niet anders als bij experts te rade gaan die in moeilijke omstandigheden snel moeten beslissen. Men kan niet voor onderzoek experimenteel een echte brand stichten, een auto ongeluk laten plaatsvinden of een oorlog beginnen. Dat onderzoek laat zien dat de experts vaak op één aanwijzing afgaan. Dat doen ze niet zomaar. Er is ervaring op een gebied voor nodig. Als die ene *cue* niet werkt, gaan ze onmiddellijk over op de volgende. Kortom, rationele stapsgewijze processen zijn niet natuurlijk bij alle beslissingen. Ze moeten dan ook protocollair afgedwongen worden. De aanpak van meteen afgaan op wat het meest kans biedt op een oplossing wordt ook beschreven in de SDT (signaal detectie theorie, zie Luan et al., 2011). Er wordt niet lang gezocht zoals bij de snelle en beperkte heuristieken van Gigerenzer.

De nadruk op normatieve statistiek doet vergeten dat het klinisch-diagnostisch proces een verschijnsel is met een eigen aard. Het studie verdient naar wat het is, evenals intuïtie en onmiddellijk beslissen in riskante situaties. Recent is daar weer aandacht voor na een tijd in diskrediet geweest te zijn. Een nummer van *Psychological Inquiry* van 2010 en vooral de bijdragen van Hogarth getuigen daarvan. Intuïtie wordt overigens schoorvoetend toegelaten en is beperkt tot enkele onderzoekthema's, bijvoorbeeld de analyse van het denken van schakers dat al in de jaren 40 door De Groot is ingezet.

Het HB onderzoek zegt dat rationele regels en protocollen moeten volgen anders gaat het mis met de diagnose. Het lijkt op de bureaucratische bestuurlijke aanpak om ontsporingen te voorkomen. Het vergroot inzicht in het diagnostisch proces niet, evenmin als de bureaucratie ons de samenleving beter doet begrijpen.

Het ecologisch onderzoek is Duits van oorsprong. Daarin wordt vooronderstelt dat een diagnosticus wel fouten maakt maar ook dichter bij de werkelijkheid kan komen en sommige zaken zijn niet te voorspellen. Andere wel omdat de structuur van de omgeving en gedragingen van de cliënt de diagnose of waarneming kunnen corrigeren. De oordelaar/clinicus kan leren en past zich aan.

De claim van clinici dat ze informatie niet-lineair combineren en rekening houden met soort en waarde van diagnostische aanwijzingen is niet rechtstreeks onderzocht. Wel is nagegaan welke modellen hun producten - de diagnoses - gemiddeld genomen het best voorspellen. Dat blijkt het lineair model te zijn: tel een beperkt aantal *cues* op en dan kun je aardig voorspellen. De fouten in de voorspelling leiden ertoe om voorschrijvende modellen te accepteren maar omdat mensen daar steeds van afwijken zijn ze gebruikt om weer nieuwe

te maken, zodat ze professioneel diagnosticeren weer wat beter zouden beschrijven (Weber & Johnson, 2009). Een model is een rationeel ideaalbeeld dat met de werkelijkheid vergeleken wordt. Die vergelijking leert ook hoe het alledaagse oordelen ervan afwijkt. Als we houding en bevindingen van Gigerenzer serieus nemen dan moeten we aanvaarden dat leken en professionals onder specifieke omstandigheden met succes simpele strategieën volgen bij oordelen en diagnosticeren.

De ecologische aanpak van Brunswik en Gigerenzer heeft ertoe bijgedragen dat modellen meer op de structuur van de werkelijkheid en de gewoonten van de diagnosticus afgestemd worden. Leken en diagnostici gebruiken geen ingewikkelde formules. Ze zijn geen rekenmachines en ze doen hun werk niet van achter een bureau. Psychologie heeft ook de taak wat zij *doen* uiteen te rafelen en ze niet alleen modellen voor te schrijven die ze vervolgens niet kunnen toepassen. De ecologische benadering van diagnosticeren laat de adaptieve aard van het oordeel zien. We kunnen leren van fouten. Zo kan een diagnosticus weerstand bieden aan een impressionistische strategie en kennis verwerven over relaties tussen *cues*, voorspellers en criteriumgedrag. Daarbij is hij zich ervan bewust dat de structuur van de buitenwereld, in dit geval het probleem, de vraag van de cliënt en zijn context, niet zo doorzichtig zijn als bij het blussen van een brand.

Realistische modellen voor diagnostiek moeten gedifferentieerd en complex zijn want er komt veel bij kijken: de gewoonten van de diagnostici, de aard, het aantal, kwaliteit en relaties van de *cues*, de kenmerken van de taak en de omgeving, de sociale context en de relevante kenmerken van de cliënt. HB onderzoek wijst op de verkeerde gewoonten. Waar is studie naar de goede gewoonten en zijn die aan te leren? Meta-studies leren over de *cues* die relevant zijn voor het ontstaan en in stand blijven van het probleem van de cliënt. De diagnosticus heeft ook realistische modellen nodig om de cliënt te kunnen uitleggen hoe hij tot zijn diagnose komt.

Niettemin blijven de bezwaren van Brunswik, Gigerenzer, Klein et al. en ook de Bayes regel berusten op rekenmodellen met de relevante dimensies en *cues.* Vooronderstelt wordt daarbij dat het onderliggend causaal model juist is en als zodanig aanvaard en toegepast wordt door de diagnosticus. Nu kan zijn intuïtie precies dat in de war brengen: wat als hij niet 'gelooft' in die onderliggende structuur met die observeerbare waarden op dimensies, *cues* en hun relaties? De diagnosticus is er in dat geval niet zeker van dat het geïnduceerde model past. Zijn diagnose volgt in dat geval de modellen en regels niet en zijn oordeel wordt van daaruit als suboptimaal gekarakteriseerd. Het gaat er dan om expliciet te maken of er een beter onderliggend causaal model te maken is met niet geobserveerde en gemeten kenmerken. Deze gedachte werken Meder et al. (2014) uit. Ze laten zien dat een oordelaar *beyond the information given* (de specifieke *cues* en waarden op dimensies) gaat en inferenties doet op basis van andere *cues* die niet zichtbaar zijn. Of zijn 'belief' tot betere diagnoses en predicities leidt is een empirisch vraagstuk. De auteurs laten zien dat de diagnosticus buiten de observeerbare en voorgeschreven cues gaat als hij gelooft dat er een alternatieve causale structuur.

Zowel de diagnosticus als de cliënt hebben niet veel aan de boodschap dat ze slechte informatieverwerkers zijn die irrationeel te werk gaan. Skinner wist al dat straffen zelden tot gedragsverandering leidt. Dit betekent ook dat diagnostisch teams alleen als leeromgeving kunnen functioneren als ze soepel met complexe voorschriften omgaan.
De klinisch-statistisch controverse is niet alleen *gefundenes Fressen* voor psychologen. Zij is ook aanwezig in de psychiatrie, maatschappelijk werk, cardiologie, huisartsgeneeskunde, radiologie, interne geneeskunde, enzovoort. De predicties zijn in die domeinen ongeveer even moeilijk of gemakkelijk (Meyer et al., 2001, 2002). Als dat zo is, zo vragen sommige Amerikaanse klinische psychologen zich af, waarom mogen wij dan geen medicijnen voorschrijven? McGrath (2010) werpt deze vraag op en meldt een doel van de Amerikaanse psychologen vereniging (APA) is om dit recht te verkrijgen. Het is slechts bereikt voor het leger, onderdelen van de *Public Health Services* en de Indiaanse Gezondheidsdienst. Als psycholoog moet je ergens beginnen en volgens het Ministerie van Gezondheid is dat onderaan.
De controverse lijkt een wedstrijd uit het verleden. Het aantal hits bij de *entry* over de twee voorspelstijlen neemt af. In medische tijdschriften is die er nog wel, vooral bij cardiologie en urologie. De strijdbijl is begraven en de aandacht verschoven. Diagnostisch voorspellen is een activiteit die het verdient als een verschijnsel *sui generis* bestudeerd te worden. Er is voldoende onderzoek dat op de zwakten in het spontaan diagnosticeren wijst. Niet alle zijn te voorkomen en de grenzen aan voorspelbaarheid zijn zichtbaar. Sommige criteria en verschijnselen zijn niet te voorspellen en er zijn *wicked environments* waar je maar beter je mond kunt houden. Het sluitstuk van een diagnose is steeds een probabilistisch antwoord op de vraag van de cliënt. Zekerheid kan een diagnosticus niet bieden.
De controverse is toch weer even op scherp gesteld door Dawes et al. Afgaan op intuïtie in plaats van op modellen en statistiek is onethisch gedrag. Gigerenzer (2000) laat zien dat eenvoudige intuïtieve regels onder specifieke omstandigheden beter werken dan modellen. Kruglanski en Gigerenzer (2011) verwerpen dan ook de *dual process theory* dat het óf om regel geleid, bewust, met inspanning tot stand gebracht en rationeel diagnosticeren gaat óf om intuïtief, associatief, onbewust, inspanningsloos oordelen. Dat is een *muddy dichotomy* die niets verheldert van wat diagnosticeren is en wat diagnostici doen. Roskes et al. (2012) bestrijden ook zo'n tweedeling en wel bij creativiteit. Ze stellen een *dual pathway* voor: intuïtief, flexibel, moeiteloos naast volhardend, rationeel, stap voor stap. Beide kunnen tot een creatief resultaat leiden. Ze concluderen dat rationele vermijdingsgemotiveerde deelnemers evengoed tot een creatieve oplossing kunnen komen dan creatieve, associatieve op hun doel afgaande deelnemers. De strikte scheiding tussen willekeurig, rationeel en intuïtief is niet houdbaar. Moxley et al. (2012) bestudeerden het schaakspel dat eerder gebruikt is om het analytisch proces van experts te analyseren (De Groot, 1946). Zij benadrukken het samengaan van *take the first the best* heuristiek en intuïtie met overleg, aarzeling en de tijd nemen om na te denken. Zelfs in de gecontroleerde en doorzichtige omgeving van het schaakspel spelen beide een rol. Het zijn geen strikt gescheiden

processen. Toch blijft de nadruk op rationaliteit: de wereldkampioen schaken die het van de computer verliest. Cone en Gilovitsch (2010) nemen afstand van hun eerdere afwijzing van intuïtie als een bron van valide kennen en diagnosticeren. In 2002 was Gilovitsch een van de redacteuren van een boek dat de snelle, gemakzuchtige, eenvoudige en foute heuristieken en vertekeningen van clinici beschreef. Het lijkt erop dat er gepleit wordt voor een plaats voor beide en zelfs voor een balans tussen protocolleren en intuïtie: het combineren van 'gaven van hoofd en hart'. Cone en Gilovitsch pleiten voor rationeel protocollair beslissen waar dat kan en voor intuïtie waar dat niet hoeft of niet kan. De ongewogen en statistisch lineair gewogen diagnose is niet meer overal en altijd superieur.

Wat betekent dit voor de diagnosticus? De beschuldiging dat hij simpele heuristieken gebruikt en zijn diagnose vertekend is, heeft in zoverre zin dat hij er een gewaarschuwd mens m/v door geworden is. De ecologische benadering moedigt aan de structuur van de werkelijkheid - vooral de sociale context van de cliënt en de wijze waarop hij zijn vraag formuleert en waarneemt - te betrekken in de diagnose. Deze kunnen ertoe leiden dat hij moet vaststellen dat die context soms een ondoorzichtige structuur heeft of de cliënt een probleem ziet dat er niet is of anders geduid kan worden. In sommige gevallen is er geen grote kans op een geldig advies of voorspelling. Het niet meer onvoorwaardelijk verdedigen van regels, protocollen, statistische modellen bij diagnosticeren heeft er toe geleid dat *case formulation* weer een aanvaarde diagnostische activiteit wordt (Gigerenzer & Gassmaier, 2011). Net als schakers en detectives gebruiken diagnostici intuïtie, snelle heuristieken, rationeel stapsgewijs doordenken en patroonherkenning van de combinatie kenmerken cliënt x context die covariëren met-, of de oorzaak zijn van zijn vraag of probleem (Moxley et al., 2012).

De statistici en modelmakers zoals Dawes et al. hebben van klinische predictie voor een deel een stropop gemaakt en die vervolgens aangestoken. Statistische predictie berust op steekproeven en de individuele cliënt kan gegeven de niet perfecte correlaties tussen zijn kenmerken en criteria van de regressielijn afwijken. Diagnostiek is n = 1 onderzoek en de modellen zijn gemaakt voor homogene populaties. De kans dat de diagnosticus een cliënt in zijn spreekkamer krijgt die in een populatie excentrisch is lijkt niet denkbeeldig. De diagnosticus moet wel kennis hebben van predictor-criterium correlaties maar tegelijkertijd zelf steeds uitmaken of de relatie in zijn geval geldt.

II Diagnose en behandeling

Naast de kritiek van de statisticus, modelbouwer en formulespecialist heeft de diagnosticus te maken met de economische, marxistische clichévraag naar zijn meerwaarde voor de behandeling. Vooral pedagogen, verdwaalde ontwikkelingspsychologen, therapeuten en (kinder) psychiaters willen meerwaarde zien. Waarom zou je anders diagnostiek niet overslaan en meteen met de beste behandeling beginnen? Of kan diagnostiek helpen bij het kiezen van de - door middel van experimenteel onderzoek gevonden - effectiefste behandeling? Of, is er per cliënt maatwerk mogelijk? Kan uit de vele treatments er één gekozen worden die past bij cliënt X die gekenmerkt is door zijn profiel op typerende en cognitieve gedragingen en specifieke sociale context? Het clichéwoord meerwaarde wordt vaak in behandelaarskringen gebruikt. Ze willen af van de tijdrovende diagnose. De diagnosticus moet hen het perspectief kunnen geven en de belofte gestand doen dat er combinaties van zijn diagnose en hun behandelingen zijn die het bevorderen of verminderen van (on)gewenst gedrag van de cliënt dichterbij brengen. Daarbij wil de behandelaar dat uit zijn arsenaal aan interventies er één of enkele gekozen worden op basis van de diagnose van de cliënt. De behandelaar legt het vraagstuk op het bord van de diagnosticus. Hij vlucht als het ware vooruit want de deugdelijkheid van interventie, behandeling, treatment wordt beperkt bevraagd, al dient dat evengoed aangetoond te worden. De roep om evidence-based behandelingen is voor behandelaars voor een deel een last. Mijn behandeling werkt al jaren, moet die nu weer passen in een gecontroleerd programma? De houding van de behandelaar is dat je iets moet doen en vooral niet niets als zich een probleem voordoet.

De diagnosticus is daarentegen probleem- en niet meteen oplossingsgericht. Al maakt hij zich er niet populair mee, hij denkt in problemen en niet in oplossingen. Daar kan de behandelaar niet mee uit de voeten. De afwachtende, uitzoekende diagnosticus heeft het tij niet mee.

1. Diagnose-Behandel-Combinaties: DBCs

Deze (oude) nieuwgevormde samenstelling kent twee elementen en hun relatie. De analyse gaat dan ook eerst over de afzonderlijke kwaliteit van de Ds en van de Bn. Als diagnose wordt opgevat als opstap naar een behandeling dan ligt het uitzoeken van effectieve DBCs voor de hand.

Verhouding van D en B en plausibiliteit van de DBC Als er twee dingen zijn die door onderscheiden groepen worden uitgevoerd is er naar westers gebruik onenigheid over wie het voor het zeggen heeft. Wat is D (de diagnose) waard? Draagt ze bij tot het kiezen van de beste behandeling (B) voor de individuele cliënt? Is er voldoende voorraad aan Bn en deugt wat op *stock* is? Een keten is immers zo sterk als de zwakste schakel. Zijn er al voldoende specifieke en effectieve DBCs? Ze liggen voor de hand ook buiten de hulpverlening. Menig universiteitsbestuurder maant bijvoorbeeld aan te sluiten bij de actualiteit, de praktijk, het bedrijfsleven en kennis en kunde te *valoriseren.* De ivoren toren uit en de media in met goede en verantwoorde berichten, informatie, wetenschappelijke vondsten. Bestuurders vragen om maatwerk, bijvoorbeeld in de Zorg. Persoons-Gebonden-Budgetten (PGBn) worden gewijzigd en vertraagd uitbetaald vanwege angst voor fraude en omdat uitkeringsinstanties de administratie niet aankunnen. Ze waren bedoeld om individuele patiënten en cliënten beter te bedienen. Niet meer algemene regels uit Den Haag want die leggen patiënten op een Procrustesbed. Je moet afstemmen. De zorgtoekenning, de organisatie en het uitbetalen moet je niet aan cliënten en mantelzorgers overlaten vanwege belangenverstrengeling. Gemeenten kunnen zorgen voor afstemming en betere controle want ze staan dicht bij de burger.

We zitten te wachten op DBCs Het is begrijpelijk dat er in de medische en nu ook in de psychologische hulpverlening naar DBCs uitgekeken wordt. Het is een open deur dat er niet over- en onder gediagnosticeerd en over- en onderbehandeld moet worden. Alles met mate. De twee leden van de afkorting: Diagnose (D) en Behandeling (B) moeten daarvoor afzonderlijk deugen. De diagnose moet valide zijn en de behandeling moet effect sorteren. Daarna kan nagegaan worden of de combinatie tot synergie leidt. Vormen ze een koppel met een goede relatie? De behandelaar vraagt van diagnosticus zijn meerwaarde te bewijzen. De diagnosticus stelt voor niet meteen te beginnen met een behandeling en eerst te kijken wat er aan de hand is, en veel problemen gaan vanzelf over immers. Geen relatie zonder wederzijdse verwachtingen en eisen. En, diagnostici en behandelaars nemen elkaar graag de maat.

Tweedelingen, we houden ervan Ik-niet-Ik, *res extensa - res cogitans, ying - yang, mind-body*, goed-kwaad, god-duivel, man-vrouw, managers-de werkvloer, verleden-heden, theorie-praktijk, romantiek-rationalisme, ideaal-werkelijkheid, empirisme-rationalisme, context van ontdekken-context van rechtvaardigen, orde-chaos, klinische-statistische

predictie, *nature-nurture,* links-rechts, *multitrait-multimethod* en ga zo maar door. En deze verdeelactie (soms een verdeel-en-heers actie) vertonen we gemakkelijk bij Diagnose en Behandeling. Dit wordt in de kaart gespeeld door het feit dat er groepen te onderscheiden zijn die zich vooral met het een of met het ander bezighouden. In werkelijkheid is bijna elk onderscheid niet zo duidelijk. Zelfs levend en dood materiaal vormen geen absolute tweedeling want zuurstof en fotosynthese zijn nodig voor leven, maar ook dood materiaal bevat zuurstof die nodig is voor leven en zuurstof kan vernietigend zijn. Voor de duidelijkheid houden we ze nog even uit elkaar. De schuld van Descartes. Zijn *idées claires et distinctes* werken immers gemakkelijk. En, er is meestal ook wel een reden voor een onderscheid. Twee dingen zijn zelden identiek of onlosmakelijk gekoppeld.

De relatie tussen Dn en Bn is niet één op één Misschien enkele in de medische zorg maar zelden in de psychologische hulpverlening. Zoals Fodor zei: het doel - verbetering van het gedrag van de cliënt dit geval - is vrijwel altijd multipel realiseerbaar. Niettemin zien we specifieke DBCs wel zitten. Het is helder, je helpt cliënten en het is goedkoper. Minder foute diagnoses, minder ineffectieve behandelingen. Dat is het verhaal op papier. In methodologisch jargon: Haagse maatregelen en voorschriften vormen experimentele programma's die bestaan uit complexe ketens van acties: de onafhankelijke variabelen. Het wordt ongedifferentieerd op iedereen toegepast die er - met of zonder diagnose - voor in aanmerking komt. Doel is maximalisatie van kwaliteit van zorg voor een zo groot mogelijk aantal cliënten: het utilitaristische principe. Gemeenten moeten maatwerk bieden en dat kan niet volgens een utilitaristisch principe: bekijk elk geval. In het jargon: ze moeten rekening houden met moderatorvariabelen en interacties. De variabelen/covariaten zijn kenmerken van de cliënt en zijn sociale context. Kwade tongen beweren dat afschaffen van het PGB voortkomt uit wantrouwen ten opzichte van de cliënt en de commerciële bureaus die de PGBs binnenhalen voor de zorgbehoeftigen. Anderen noemen het een bezuinigingsmaatregel. Er zijn altijd veel interpretaties en redenen voor het doen van het goede en het streven naar het betere.

Samenvatting en conclusie

Bij DBCs gaat het om twee elementen. Zodra er een tweedeling te maken is zijn we er blij mee want die schept orde. Tegelijkertijd is er een streven naar kennis van de eigen waarde en betekenis van de elementen. Als er ook nog verschillende groepen mee bezig zijn met ieder hun eigen expertise leggen ze al gauw hun eisen op het bord van de ander. Bij de DBCs is dat het bord van de diagnosticus. De behandelaars zouden evenzeer moeten kunnen laten zien dat hun behandelingen effectief zijn voor een doelgroep. Door de vraag naar *evidence-based treatments* komt dit er ook aan. De diagnosticus kan de bal terugspelen. Bij de afstemming van Ds en Bn kun je ervan uitgaan dat er nauwelijks één op één verbanden zijn. Vrijwel elk doel is immers multipel realiseerbaar.

2. Diagnose

D staat voor diagnose en het diagnostisch proces. Deugen ze, zijn ze valide? Bestuurders steunen op beoordeling van tests volgens richtlijnen van de Cotan. Studietoetsen worden door het Cito gemaakt volgens de moderne testtheorie (IRT). Daar valt weinig op aan te merken. Kwaliteit is geborgd. Voor het diagnostisch proces is het hypothese toetsend model (HTM) aanvaard als valide procedure. Ik herhaal Witteman et al.'s opmerking, het is

'...een leidraad voor het uitvoeren van psychodiagnostisch onderzoek in de praktijk' en het *'...geeft inzicht in hoe verschillende stappen in het psychodiagnostisch onderzoek op een goede manier uitgevoerd kunnen worden'* (onderlijning: jtl).

De Bruyn et al. (2015, p. 18) garanderen dat je het diagnostisch besluit wetenschappelijk verantwoord neemt als je hun procedures volgt. Alles geregeld met de diagnose! Behandelaars, zoals psychiaters, pedagogen en klinisch en ontwikkelingspsychologen zouden geen probleem van diagnostiek sec hoeven te maken. Soms zeggen ze dat er geen test is voor een gedrags- of leerprobleem waar ze tegen aan lopen. Als de vraag groot genoeg is, komt die test er: marktwerking in testzaken. Er is vrede of misschien beter gezegd onverschilligheid ten aanzien van aanbod en kwaliteit van het D-lid van de DBCs. Behandelaars maken zich niet zo druk om de D. Ze is bijna niet nodig, maar als je het zo graag doet dan moet het maar. Je moet immers tests, studietoetsen bij plaatsen van leerlingen/werknemers en instrumenten voor de bepaling van persoonlijkheidsstoornissen gebruiken. Elk jaar is er even protest tegen de Citotoets voor groep 8 maar dat gaat meestal snel over. Bestuurders houden er aan vast. Er is niets beters en het Cotan-oordeel is een expertoordeel. Het Cito zit een groot gebouw met vaklieden en heeft nauwelijks concurrentie.

Als je echter kijkt naar wat de diagnosticus op basis van het kennisbestand van de psychologie kan is bescheidenheid op zijn plaats. De d-waarden van experimenten en interventies zijn bescheiden evenals de r- waarden van predictief onderzoek. Meta-studies ondersteunen deze bewering en het zit er niet in dat de resultaten op korte termijn verbeteren. De waarden in steekproeven zijn bescheiden tot gemiddeld en er is ook nog aanzienlijke variantie in vergelijkbare studies. En, de DBC is er voor de individuele cliënt, maar hij is geen steekproef.

Onderzoekers hebben uit de aard van hun missie redenen te over om stil te staan bij de kwaliteit van de elementen van diagnostiek: theorievorming, operationalisatie & meten, het instrumentarium en het diagnostisch proces. In de praktijk wordt de houding ten opzichte van de diagnose vooral bepaald door de kwaliteit van de instrumenten en nauwelijks door de wetenschappelijk staat van de twee overige elementen. Diagnoses leiden niet tot sombere gezichten bij DBC uitvoerders. Ze vinden het niet zo nodig en vooral tijdrovend;

pak meteen de vraag/het probleem aan. Hulpverlening is het centrale doel: kom maar op met je diagnostische kennis.

Is Diagnose al niet Behandeling? Behandelen domineert zozeer dat een enkele auteur de diagnose al als *treatment* beschouwt. Het onderscheid tussen D en B is klein zeggen Poston en Hanson (2010). Ze zien de diagnose als een effectieve ingrediënt in de hulpverlening. Ze moet in dat geval gepersonaliseerd en in samenspraak met de cliënt uitgevoerd worden, denk aan Hermans' subject als mede-onderzoeker in de *Zelf Konfrontatie Methode*. Poston en Hanson vonden 17 studies waarin een dergelijke procedure gevolgd werd. Ze kwamen tot de slotsom dat de werkwijze effecten had die vergelijkbaar waren met klassiek uitgevoerde therapieën. Dit resultaat ondersteunt therapeuten die de tijdrovende diagnose met behulp van tests en het diagnostisch oordeelsproces willen overslaan. Dat mag en kun je doen want de relatie tussen diagnose en therapie is als gezegd niet zo hecht. Bij verschillende diagnoses wordt eenzelfde therapie voorgesteld en bij eenzelfde diagnose worden meer therapieën aanbevolen.
Er is ook kritiek. Lilienfeld et al. (2011) beweren dat Poston en Hanson effecten van 'diagnose-plus' overschatten. Ze zouden studies vermelden die diagnoses met behandelingscomponenten combineren. Dat zijn geen pure diagnoses. Bovendien sluiten ze studies uit die geen significante resultaten tonen. Er zijn preciezer opgezette studies nodig om te laten zien dat een diagnose al een effectieve ingrediënt van de therapie kan zijn. Hermans en Poston & Hanson zien een glijdende schaal van diagnose naar therapie, terwijl Lilienfeld et al. het voorlopig als twee categorieën beschouwen.

Samenvatting en conclusie
Auteurs van boeken over het diagnostisch proces suggereren dat het met de diagnose wetenschappelijk wel goed zit, als je hun procesvoorschriften volgt. Behandelaars maken zich niet druk over de diagnose. Hun probleem is dat diagnoses tijd verloren laten gaan voor de behandeling en niet perfect kunnen voorsorteren voor de beste behandeling. Ze slaan de uitvoerige diagnose over en laten het aan de diagnosticus om te tonen dat ze er iets aan hebben. Ze zouden kunnen wijzen op valse positieven en negatieven in diagnoses. Empirische (meta) studies tonen dat de predictieve validiteit van klinische oordelen en tests beperkt is. Een enkele auteur wijst erop dat diagnose en behandeling niet strikt gescheiden zijn. Dat is het geval als de diagnose gepersonaliseerd is en in samenspraak met de cliënt vorm krijgt. Het lijkt te gemakkelijk om te zeggen dat de diagnoses grotendeels slagen als je het protocol volgt.

3. Behandeling
Bij Ds merkte ik op dat het doel: weinig valse positieven en negatieven in bescheiden mate behaald wordt**.** Nu de B, die staat voor behandeling, interventie, therapie, ingreep, modificatie of algemeen voor een *treatment*. Meehl was behalve methodoloog en filosoof

ook therapeut. Hij heeft bijna 10.000 uur cliënten behandeld. Toch heeft hij niet gepubliceerd over *techniques of psychological treatments* en hun effecten. Op de vraag 'waarom' antwoordde hij: '*I didn't know, how*'. De huidige *evidence-based* brigade zou een lastige aan hem gehad hebben. Het is immers niet eenvoudig om de waarde en het effect van iets samengestelds als een behandeling te bepalen voor groepen en individuele cliënten. Niettemin willen we dat een behandeling effectief is.
Het experiment is de procedure om te bewijzen dat de onafhankelijke variabele - de behandeling - de oorzaak van de verandering in de afhankelijke variabele is - het (on) gewenste gedrag van een individu of groep. Experimentele programma's dienen om ongewenst gedrag door therapie om te buigen, cognitie te stimuleren en beperkingen op te heffen of te verminderen. Het experimenteel ontwerp kan overtuigen dat de behandeling de oorzaak van de gewenste gedragsverandering is. Dit ontwerp is geliefd omdat het ons op het spoor brengt van oorzaken: *causae efficientes*. Je kunt die mechanistisch opvatten en dat snap je meteen: ik geef een klap en iets gaat stuk; ik doe iets concreets, iets doseerbaars en er verandert iets. Kern is dat er oorzaken zijn die we kunnen manipuleren. We kunnen controle uitoefenen op en over gedrag. Quasi-experimenten zijn gedegenereerde, 'alsof' experimenten waarmee we met enige omzichtigheid ook nog oorzaken op kunnen sporen.
Naast methodologische ernst spreekt uit het (quasi) experimentele ontwerp optimisme om de samenleving en het individu te verbeteren. De eerste publicaties over quasi-experimenten zijn uit de jaren 60. Deze auteurs nemen aan dat het onderzoek bijdraagt aan het verbeteren van het welzijn van mensen en hun omstandigheden. Dat paste in die jaren, maar het is naïef. Er zijn tegenvoorbeelden: Als effect van begeleiding van criminele jongeren in inrichtingen wordt bijvoorbeeld beweerd dat er winst is geboekt omdat bij gelijkblijvende criminaliteit het geweld is afgenomen. Heeft begeleiding tijdens het verblijf ertoe geleid dat de criminelen hetzelfde bereiken maar met minder geweld? Het bureau HALT dat overtredingen met drank bij jongeren aanpakt met een *treatment* behaalde geen succes. Jongeren gingen na de verplichte training meer drinken.

Effect van behandelingen Het is zaak bij behandeling op te letten wat er precies gebeurt en wat er uit komt. Het kan tegenvallen: verwachtingen worden niet ingelost en het effect kan zich tegen je keren: iatrogeen. Niettemin tonen behandelaars een aanstekelijk optimisme. Als gezegd, worden programma's van de overheid om werkgelegenheid en schoolsucces te bevorderen zelden geëvalueerd. Misschien heeft de overheid de neiging met door haar goedgekeurde programma's door roeien en ruiten te gaan: die sterke minister, staatssecretaris! Er zijn echter *iatrogene* effecten van interventies. Recidives liegen er niet om en dat is het geval in bijna alle landen ter wereld van Nederland tot China en India. Had Skinner dan toch gelijk toen hij voor afschaffing van straffen pleitte?
Dit optimisme noem ik naïef. Er zijn goedmoedige en kwaadaardige voorbeelden van experimenteren. Beide gaan uit van maakbaarheid. Pedagogen en filosofen schrijven regels voor die het gedrag van de mens zouden verbeteren. Ze geloven dat het werkt, of ze

beschouwen het als *self-evident*: het is in zichzelf, met zichzelf en door zichzelf evident en goed. Dat is niet wat met *evidence-based* bedoeld wordt want daar is onderzoek voor nodig. Een voorbeeld is het Curriculum Schoolrijpheid van de pedagogen Dumont en Kok (1973). Toen onderzoek verricht werd naar de opeenvolging van de cognitieve stappen en die voor een deel niet empirisch gevonden werd - de stappen in enkele domeinen vormden geen Guttmanschaal - werd gezegd: 'Dat kan niet'. De auteurs waren niet bereid om hun logica te laten sneuvelen voor een empirische toets.

Is behandelen altijd verbeteren? Een *treatment* bedoelt het goede te bewerkstelligen. Dat is boven iedere twijfel verheven. Empirisch onderzoek naar effecten van behandelingen matigt de verwachtingen bijna altijd. Ze zijn feitelijk bescheiden tot gemiddeld. Nu moet je niet op iedere slak zout leggen, maar toch enkele feiten: Men juicht na succesvol experimenteel onderzoek. Het wordt gepubliceerd, er zijn significante effecten, redelijke effectgroottes en het publiek is enthousiast. Bij nader inzien kan de stemming wijzigen: de effectgroottes van om het even welke therapie zijn gering tot gemiddeld (r = gemiddeld .30 en d = gemiddeld < 0,50). Een voorbeeld: In studies van Cuijpers et al. (2010) worden effecten van zeven behandelingen voor depressie vergeleken: cognitieve gedragstherapie, non-directieve ondersteuning, gedragsactivering, psychodynamische behandeling, probleemoplos- en interpersoonlijke therapie en sociale vaardigheidstraining. Ze verzamelden 53 studies over milde tot ernstige depressie bij volwassenen. De studies werden geselecteerd en alleen methodologisch degelijke werden opgenomen. Alle behandelingen hadden een zwak effect. Interpersoonlijke therapie was effectiever (d = 0,20) dan ondersteunende (d = 0,13). Het maakte niet uit of er sprake was van milde of ernstige depressie want er was geen therapie die voor de ernstigere problematiek minder of meer effectief bleek. Dit resultaat is minder optimistisch dan Amerikaanse studies van bijvoorbeeld Weisz et al.(1995). Ze vermelden d-waarden tussen de 0,45 en 0,90. Het verschil met enkele Amerikaanse resultaten schrijven Cuijpers en collega's toe aan de publicatie *bias* van redacteuren van tijdschriften.

Is succes het significant verschil tussen wel of geen behandeling? Er kan bovendien op gewezen worden dat succes gelijkgesteld wordt met de kans op de alfa fout: p < .01, < .05, dus met significantie. Alfa fouten zeggen niet veel. Daar had de ontwerper van de Anova, Fisher (1955) al oog voor. Het is het zinvol bij productcontrole. Je kunt je niet veroorloven dat een auto een kans van 5% heeft op een ongeluk door mechanische gebreken. Er zijn redenen te bedenken voor het verschil in therapiesucces tussen Europa en de VS: therapietrouw, geloof in therapie, optimisme en striktheid van uitvoering. Nog een voorbeeld: De Zwitserse onderzoekers Barth et al. (2014) vergeleken zeven depressietherapieën: interpersoonlijke, cognitieve gedrags-, probleemoplos-, psychodynamische-, sociale vaardigheid therapieën, ondersteunende counseling en gedragsactivering. Ze deden een netwerk meta-analyse op basis van 198 studies bij meer

dan 15.000 cliënten. De basisvergelijking: therapie-geen therapie toonde dat therapie helpt. Het effect was gering/bescheiden tot een enkele keer gemiddeld. De therapieën verschilden onderling niet: ze waren even effectief voor verschillende leeftijdsgroepen en hetzelfde gold voor verschillende uitlokkers van depressie zoals ziekte, geboorte van een kind, verlies van een geliefde. De auteurs wijzen erop dat ze gemiddelden vermelden. De effecten hoeven niet voor iedere cliënt te gelden. Tot hier ging het over therapiesucces. Het succes van onderwijsexperimenten verschilt daar niet van. Verderop vermeld ik Hattie's omvangrijke meta-studie over effectieve ingrediënten voor onderwijsresultaten.

Samenvatting en conclusie

Een *treatment* is een complexe ingreep die je niet met een variabele kunt beschrijven. Toen ze Meehl vroegen waarom hij nooit therapie effect onderzoek gedaan had zei hij dat hij niet zou weten hoe. Nogal complex onderzoek dus. Het experiment is de procedure om te bewijzen dat er een oorzaak (de onafhankelijke variabele) is voor een gedragsverandering (de afhankelijke variabele). Behandelaars en ook methodologen (Shadish et al., 2002) gaan uit van de maakbaarheid van het gedrag. De eersten prijzen dan ook enthousiast hun interventies aan. De overheid getuigt ook van optimisme over haar ingrepen en projecten. Twee verschijnselen temperen dat optimisme. Er zijn iatrogene effecten en het succes is bijna altijd bescheiden voor een steekproef. Nederlands onderzoek laat geringe d-waarden zien. Amerikaans onderzoek geeft een optimistischer beeld. De effectgroottes zijn echter gering en op individueel niveau is er weinig zekerheid of de *treatments* werken.

4. *Evidence-based* behandelingen

De diagnose komt er bij de op-behandeling-gerichten onder de DBC-beoefenaars soms matig af: de meerwaarde kwestie. Dat ervoer Kouwer in de kritiek van de psychiater Kuiper (1965) op zijn *Spel van de Persoonlijkheid*. Het maakt ze eigenlijk niet zoveel uit. Een beperkt aantal verdedigt D, bijvoorbeeld Barendregt (1974). Een ander aantal stelt geen lastige vragen onder hen docenten die D-onderwijs verzorgen en het een geprotocolleerd diagnostisch proces uitvoeren. De behandeling komt er in onderzoek ook niet zonder kleerscheuren af: de effecten zijn bescheiden, worden overschat en behandelaars tonnen soms een niet gerechtvaardigd optimisme. Dat wil niet zeggen dat gebruikers er geen heil in zien. Denk bijvoorbeeld aan door scholen gewaardeerde pestprogramma's die afgewezen worden omdat onderzoek ontbreekt of geen resultaat laat zien. Is dat een placebo effect en zou dat niet bij onderzoek betrokken moeten worden? Hoe vreemd het ook voor ons causaal efficiënt verstand ook mag zijn placebo's blijken soms te werken, ook in de geneeskunde. Sommige huisartsen schrijven ze voor en met *meer* in hun hoofd dan de slogan: 'baat het niet, dan schaadt het niet'. Deze stand van zaken zou voor onderzoekers begiftigd met een Streng Methodologisch Geweten een uitdaging kunnen zijn. Ze zouden kunnen proberen uit te vinden waarom, onder welke condities (covariaten) d- en r- waarden tot stand komen en of de limiet van die waarden bereikt is.

Het kaf en het koren Deze stand van zaken lokt acties uit om bij behandelingen het kaf van het koren te scheiden. Als je nu alleen *Evidence-Based Treatment Programs* (EBTPs) toestaat, zou dan de kou niet uit de lucht zijn? Dit heeft geleid tot een categorisering van therapieën. Ze verschillen met betrekking tot het empirisch bewijs voor hun effectiviteit. Chambles en Ollendick (2001) onderscheiden:

I 'Well-established/ efficacious and specific
II Probably efficacious
III Promising

Niet onderzochte zou IV kunnen zijn en Mislukte V. Placebo's hebben geen plaats in hun ordening. Je zou de controleconditie in experimenteel onderzoek als zodanig kunnen opvatten. Ze verdelen soorten behandelingen over problemen, zoals gezondheid, eten, pijn als gevolg van kanker, hoofdpijn, migraine.
De meeste behandelingen komen in categorie II terecht: mogelijk effectief; verbetering nodig, maar er is perspectief. De kinderombudsman stond waarschijnlijk de categorie *Promising* toe voor programma's ter bestrijding van pesten. Het is mogelijk dat hij alleen programma's uitsloot waar (nog) geen onderzoek naar gedaan is of die vooralsnog *face valid* (mijn IV) zijn. De V categorie zal niet gepubliceerd worden. Dat is jammer, je leert er immers van.
De *evidence-based* eis heeft ertoe geleid dat medicijnen zoals Ritalin (methylfenidaat) en antidepressiva gevestigd zijn in de zin van hun sporen verdiend hebbend. Ze zijn werkzaam inclusief bijwerking en er is onzekerheid over het lange termijn effect. Cognitieve gedragstherapie en multisysteem therapie komen af en toe voor in het rijtje van *evidence-based* producten.
Meta-studies laten zien dat behandelingen geringe tot gemiddelde resultaten opleveren in experimenteel onderzoek onder laboratoriumcondities. Succes is schaarser voor interventies in het veld. Zijn de I en II categorie interventies in het laboratorium onder ideale omstandigheden of in het veld uitgevoerd? Behandelingen moeten in concrete situaties, zoals gezin, schoolklas, de werkvloer geïmplementeerd worden. De effecten moeten generaliseren naar andere vergelijkbare gedragingen en situaties. Als je label I of II bemachtigd hebt wil dat nog niet zeggen dat het overdraagbaar is naar nieuwe steekproeven, situaties en behandelaars. McHugh en Barlow (2010) hebben een overzicht gemaakt van EBPTs en ze beschrijven implementatie en verspreiding in de VS. Ze stuiten op barrières voor de invoering van de EBTPs. Er is daarom training, supervisie en advies aangeboden. Beschikbaarheid alleen is kennelijk niet genoeg. Kazdin (2008) concludeert dat er al een lijst is met goede EBPTs is. Toch krijgen sommige cliënten niet de behandeling die ze nodig hebben omdat therapeuten hun favoriete behandeling blijven geven. Hij beveelt mini-programma's aan voor behandelaars om zo de impact van de EBPTs te vergroten.

Barlow et al. (2013) hebben al een *update* van EBPTs gemaakt. Ze zijn optimistisch over de effectiviteit van behandelingen, vooral farmacologische aangevuld met psychologische. Ze ontlenen hun optimistische kijk voor een deel aan resultaten uit dieronderzoek. Twijfel blijft: was er een belang van de farmaceutische industrie? En, als het succes zo gemakkelijk te behalen is, waarom is het dan niet veel eerder gebeurd? Waarom neemt het aantal behandelingen niet af en blijven een paar goede over? Het tegendeel is het geval, er komen er meer bij.

Werkzame ingrediënten van B Als er zo veel *treatments* zijn en ze alle een beetje effectief (bescheiden, maar steeds significante d-waarden) zijn kunnen we dan niet de werkzame ingrediënten opsporen van deze interventies? Hoe werkt deze voedselmetafoor uit? Cuijpers en collega's suggereren dat succesfactoren aspecifiek zijn: ingrediënten maken alle soorten maaltijden (*treatments*) lekker zoals vroeger bij de afhaalchinees. Ze passen niet precies bij een therapie. Voorbeelden: Shirk en Karver (2003) melden op basis van 23 studies dat bij adolescenten de kwaliteit van de relatie cliënt-therapeut het best de uitkomst voorspelt. Het gaat om een open, luisterende, niet-veroordelende houding en om uitleg geven in de taal van de cliënt. De relatie tussen de ingrediënten en het resultaat werd gemodereerd door inhoudelijke factoren: het probleem van de cliënt en factoren als *timing* en bron (wanneer gaf wie informatie), inschatting van de kwaliteit van de relatie en soort en bron voor het resultaat. Cuijpers et al. en Barth et al. wezen op enkele kenmerken van de interactie zoals de cliënt aanmoedigen om buiten de therapiesessies om iets te ondernemen, bieden van een begrijpelijk kader voor hun depressie, een verklaring geven voor wat de cliënt dwarszit en dat zo vertellen dat de cliënt het *frame* aanvaardt zodat vanuit een gedeeld frame inzicht verschaft wordt en een behandelingsvoorstel gedaan kan worden. Deze auteurs zoeken het verschil niet in de soorten therapie maar in werkzame onderdelen van om het even welke therapie.
Welke de effectieve elementen in *e-Mental Health* zijn moet nog uitgezocht worden. Wellicht gaat het daar evenzeer om algemene principes: bekwaam en accepterend communiceren. Ze lijken op de *unconditional positive regard* van de therapiegoeroe van de jaren 50-70: Carl Rogers. Niets nieuws onder de zon?
De voorbeelden laten zien dat effectieve ingrediënten conceptueel en hermeneutisch gecreëerd worden. Je moet goed kijken wat gemeenschappelijk is aan die interventies en wat hun onderliggende zin en betekenis voor de cliënt is. Er is net zo min als bij de DSM-5 multivariaat onderzoek verricht dat de programma's en hun kenmerken reduceert tot enkele werkzame en onderscheidbare dimensies (de ingrediënten). Dat zou qua onderzoek een *tour de force* zijn en dat is niet te verwachten bij de gewoonte van het presenteren van *least publishable units*.
In theorie en in de ideale wereld zijn er volop EBTPs. Ze geschikt maken is vers twee, ze volgens het boekje uitvoeren is vers drie, de goede uitvoerders vinden is vers vier, de goede klinische groep bedienen is vers vijf en er zijn misschien nog meer verzen. Effectieve

ingrediënten blijken niet specifiek voor een psychotherapie of begeleiding. Het zijn algemene kenmerken die te maken hebben met een open stijl, mensen ertoe brengen iets te ondernemen en bekwaam en adequaat communiceren. Hoewel gemakkelijk gezegd en op het oog eenvoudig toe te passen zijn uitvoering van- en *timing* bij behandelingen niet simpel. Ze vergen ervaring en alertheid van de hulpverlener. Het is vergelijkbaar met de *performance* en *timing* van de cabaretier/*performer.* Hij moet zijn publiek kennen en aanvoelen om het bij de les te houden en te amuseren: behandeling, interventie is kleinkunst. Onze cabaretier heeft heel wat *try outs* nodig voor hij met zijn voorstelling de boer op kan: de behandelaar als *stand up comedian* van de hulpverlening. Dit is niet denigrerend bedoeld. Het drukt waardering voor zijn vakmanschap uit. En, succesvolle hulpverleners op de Amerikaanse TV zijn *performers*.

Onderwijsverbetering Efficiëntie van onderwijsprogramma's en curricula is uiteraard ook nagegaan. De vernieuwingslust van onderwijskundigen is niet te stuiten. Misschien lijden ze onder te snel voortschrijdende inzichten. Geldt misschien ook hier dat de kwaliteit van de leraar, zijn enthousiasme en communicatievaardigheden de belangrijkste ingrediënten zijn? De Australische/Nieuw Zeelandse onderwijsonderzoeker Hattie heeft in zijn van het internet te *downloaden* boek van 2009: '*Visible learning: a synthesis of over 800 meta-analyses relating to achievement'* meer dan 50.000 studies uit 816 meta-studies samengevat. Eén van zijn vele conclusies: klassengrootte is een *minor factor* voor het prestatieniveau. De roep om kleinere klassen is niet te rechtvaardigen vanuit de gemiddelde prestatie. Om andere redenen kan klassenverkleining overigens aangenaam en nuttig zijn. Hattie noemt als belangrijkste variabelen: verwachtingen van leerlingen, afwezigheid van storende leerlingen, stimuleren van gepast gedrag in de klas, kwaliteit en motivatie van de leraar, verbeteren van taalvaardigheid en -begrip, de voorafgaande prestaties van leerlingen, relatie tussen leraar en leerling, feedback aan de leerling over zijn prestaties en gedrag. Een samenvatting van zijn aanbevelingen staat op het internet: *tgls.pdf: Visible learning, Tomorrow's Schools, the Mindsets that make the Difference in Education.*

Behandeling: het glas is halfvol en halfleeg Ondanks de reddingoperaties van interventies door selectie de *evidence-based* behandelingen en de zoektocht naar effectieve ingrediënten kunnen we er niet meer van maken dan dat bij Behandeling het glas zowel halfvol als halfleeg is. De stemming over het eerste lid, de D van de DBC was al gematigd: het kan vriezen en het kan dooien: we stelden vast dat de prestaties op het domein van de D bescheiden zijn. Niettemin blijkt uit vele nieuw ontworpen interventies dat er optimisme heerst en dat is naïef. Zeggen dat het wanhoopspogingen zijn gaat te ver, maar de twee leden D en B spelen vooralsnog niet de sterren van de hemel en hun vertegenwoordigers hebben elkaar weinig eisen te stellen, laat staan verwijten te maken.

Samenvatting en conclusie

Bij behandelaars heerst een gedooghouding ten opzichte van de diagnose. Een enkele keer wil een behandelaar wil uithalen naar de diagnosticus, zoals Kuiper tegen Kouwer. De diagnose is steeds geen gelopen race. De resultaten zijn bescheiden tot gemiddeld in Chens (1988) kwalificaties. De behandeling komt er evenmin zonder kleerscheuren af. Soms werken placebo's wel en interventie niet of nauwelijks. Er is niet uitgezocht hoe dat kan. Er zijn veel *treatments*. Dat is mogelijk omdat er enthousiaste en behulpzame uitvinders zijn en omdat er een markt voor is. Om te kiezen tussen de vele is experimenteel onderzoek nodig dat toont of een interventie *evidence-based* is. Publicerende Amerikaanse klinisch psychologen zijn ervan overtuigd dat er voldoende *evidence-based psychological treatments* zijn. Ze moeten goed geïmplementeerd worden. Het bewijs voor een *treatment* is echter evenals in onderzoek voor de farmaceutische industrie verkregen in een labotratoriumsetting bij een beperkte steekproef door expert onderzoekers. Interventies in het veld heeft men niet zo goed in de hand. Als zoveel *treatments* enig effect hebben kun je gaan zoeken naar effectieve ingrediënten. Dat levert een beeld op dat Rogers al in de jaren 50-60 propageerde: een open luisterende houding, niet veroordelen, opdrachten uit laten voeren en cliënten aanmoedigen zelf iets te ondernemen. Het beeld van effecten *treatments* uit de hulpverlening herhaalt zich bij interventies in het onderwijs. Beide maken duidelijk dat effecten van *treatments* beperkt zijn. Dat ze steeds weer omarmd worden en nieuwe gretig aangegrepen worden gebeurt niet op basis van hun effectiviteit, maar op basis de hoop op verbetering. Hoop doet leven. Daar kan geen sociaalwetenschappelijk onderzoek tegenop.

5. Maatwerk: een DBC voor elke cliënt?

De DBC voor een specifieke cliënt ligt zo voor de hand dat er oude voorbeelden moeten zijn, al is het ambitieus. Een klassieke DBC is de opvatting dat psychoanalyse geschikt zou zijn voor hoogopgeleiden. Gedragsmodificatie zou passen bij laagopgeleiden. Dat is discriminatie en dat voelt niet goed. Er is vast een proefschrift van een bevlogen wetenschapper misschien uit een *well to do family* die deze veronderstelling in twijfel trekt. De cultuurcriticus/ psychiater Daniels (2012 pseudoniem: Theodore Dalrymple) besteedt uitvoerig aandacht aan het feit dat de Engelse onderklasse door de aanpak van praten en helpen - pamperen zegt hij - gestimuleerd wordt om niets te ondernemen. Die aanpak lijkt wel wat op de neoanalytische hulpverlening met dat vleugje rationaliteit: ik kan toch uitleggen, dat ze niet de hele dag voor de TV moeten liggen, werk moeten zoeken, de fles moeten laten staan en vooral een goed naar jezelf moet kijken. De oude opvatting is niet verdwenen.

De hulpverlener loopt voorop bij het maken van BDCs Uit dit voorbeeld blijkt ook dat de hulpverlening *voorop* liep en loopt in combineren van diagnostiek en behandeling. De therapeut, onderwijs- en leerlingbegeleider voegde altijd al zijn behandeling naar

kenmerken van de cliënt/leerling en zijn sociale context. Pedagogen maken een behandeling op maat na diagnose van het kind, zijn ouders en de school. Zij noemen het *Handelingsgerichte Diagnostiek.* Ze werken handelingsgericht. De onderwijspsychologen benutten kennis over het functioneren van kinderen op domeinen als rekenen, taal, aandacht, motivatie, zelfconcept en emotionele en gedragsproblemen aangevuld met informatie over de opvoedings- en onderwijscontext. Zij noemen het *Diagnostiek in de Leerlingbegeleiding.* Naar welke concrete acties dat verwijst is moeilijk precies te omschrijven want in de praktijk van de uitvoering zijn er altijd vrijheidsgraden. De feitelijke praktijk is het resultaat van de combinatie van programma's x uitvoerders x situaties x leerlingen. Bovendien is er weinig onderzoek dat bewijst dat Diagnostiek + Behandeling tot gedragsverbetering leidt. Dat vergt een experimentele en controle groep op het niveau van de steekproef en n = 1 onderzoek bij een cliënt. De methodologie en bewijsvoering verschillen.

Risico van face-valididity bij de DBC Het maken van DBCs lukt wel. Er is creativiteit en fantasie voor nodig. Meestal is een behandeling samengesteld als een stappenplan en bevat verschillende ingrediënten. DBCs lopen het risico *face valid* te zijn. De therapie of training wordt vormgegeven en ingezet naar aanleiding van een waargenomen gelijkenis met, en een vooronderstelde effect van de behandeling op het (on)gewenste gedrag. Dat zijn vooralsnog hypothesen. *Face validity* is als de liefde op het eerste gezicht. Ze duidt op de overtuiging dat deze test/behandeling precies meet/bewerkstelligt wat je nodig hebt om het gedrag op een criterium te voorspellen/een gedrag te veranderen. Ze berust op een waargenomen of hermeneutisch geconstrueerde gelijkenis tussen het werk, de taak/het probleem en de inhoud van de test/behandeling.
Er zijn instrumenten waarbij we aanvoelen dat ze een gedrag voorspellen. In de jaren 50 van de vorige eeuw werden chauffeurs geselecteerd. Goede chauffeurs zijn belangrijk omdat ze duur materiaal vervoeren en op tijd bij klanten moeten afleveren. Er werd een chauffeursstoel geplaatst in een kamer met films die verkeerssituaties lieten zien. Individuele verschillen in gedrag op de stoel bleken niet gecorreleerd met het criterium: levertijd en aantal ongelukken. Het rijden in het verkeer was niet te vergelijken met de situatie in de stoel. Dit voorbeeld komt uit Drenths *Testtheorie* (1967). Vroeger maakten leraren zelf toetsen voor de leerlingen. De Schriftelijk Overhoren cijfers correleerden nauwelijks met tentamencijfers op de hogescholen en universiteiten. We herinneren ons de teleurstelling van Wissler (1901). Als leerling van McKeen Cattell deed hij onderzoek naar de relatie tussen zijn *Mental Tests* en schoolprestaties van leerlingen (n = 90 tot n = 252) van een naburige *high school* en studenten van de universiteit van Columbia. De resultaten waren mager. Schoolprestaties correleerden r = .18 met reeksen getallen onthouden, r = .08 met de sterkte van de handgreep en r = .02 met kleuren benoemen. De items vertoonden weinig samenhang: de taken vormden geen ééndimensioneel en inhoudelijk te interpreteren construct. Hij stapte teleurgesteld over naar culturele antropologie. Nog een

voorbeeld: In de jaren 60 en 70 werd een zogenoemde draadbuigtest gebruikt om het technisch inzicht op eenvoudig niveau te voorspellen. Een voorbeeld op papier moest met een soepele ijzerdraad op schaal nagemaakt worden. Het resultaat voorspelde het uitvoeren van eenvoudig technisch werk nauwelijks.
Dit zijn voorbeelden waarbij de constructeurs van diagnostische middelen, van tests geloven dat ze het bedoelde criterium voorspellen: *quod non*. Binet en ook De Groot (*Vijven en Zessen,* 1967) wezen op de onbetrouwbaarheid van schoolcijfers en oordelen van leraren.
Ze zijn naar verhouding eenvoudig: het gaat om het voorspellen van een criterium. Bij een behandeling gaat het om het voorspellen/bewerkstelligen van complex gedrag na een ingewikkelde interventie. De DBCs zijn op het oog weliswaar aannemelijk maar kunnen - evenals bij tests die een criterium zouden moeten voorspellen - *face valid* zijn. Er is experimenteel onderzoek met controle groepen nodig om resultaten van de complexe B op een aantal afhankelijke variabelen aan te tonen.
Verschueren en Koomen (2007) en Braet en Bögels (2014) voeren de redacties van boeken met zulke DBCs. Het eerste mikt op diagnose in functie van de toewijzing van zorg en op het verbeteren van de afstemming tussen onderwijsomgeving en individuele kenmerken van de leerlingen. Dit is het ATI onderzoek waar Cronbach en Snow (1977) in vastliepen. Het tweede bevat onder meer 42 protocollaire behandelingen voor kinderen en adolescenten met psychische problemen. Voor beide geldt dat bewijs voor effect van de behandelingen geleverd moet worden. Het enthousiasme van deelnemers aan DBC symposia en van uitvoerders van DBCs is niet genoeg.
Face validity kan ook de andere kant op werken. Er zijn tests/behandelingen met significante en voldoende predictieve validiteitscoëfficiënten/effecten. De autoriteiten kunnen deze echter opzij schuiven als niet terzakedoende omdat zij het criteriumgedrag op school of bedrijf niet terugvinden in de tests/behandelingen. Een recent voorbeeld is een studie van Boelema (2014). Zij heeft gedurende enkele jaren het gevolg van wekelijks alcoholgebruik van adolescenten op concentratievermogen, impulscontrole en geheugen onderzocht. Ze vergeleek drinkende jongeren met minder- en niet-drinkende jongeren. Zij vond geen verschillen op de variabelen tussen zes groepen die ze onderscheidde op basis van mate en tijdstippen van alcoholgebruik. Ze heeft haar data drie keer opnieuw geanalyseerd om zeker zijn van de uitslag van geen verschil. Tegen het protest in van onder meer van artsen die het comazuipen op de TV brachten bleef ze haar empirische resultaten trouw. Dat doet een onderzoeker en dat hoort ook zo. In een interview zei ze dat de instrumenten wellicht niet gevoelig genoeg waren maar ze hield zich aan het onderzoeksresultaat. De algemene *face valid* overtuiging is dat alcohol hersencellen beschadigt bij adolescenten, *quod non* in deze studie bij adolescenten.
Als we dit naar een DBC transponeren betekent het dat een behandeling wel het bedoelde effect heeft maar dat op *face valid* gronden niet gepikt wordt. Waarom zou ik iedere dag activiteiten ontplooien zoals een uur wandelen per dag, geen TV kijken na 23.00 uur, sociale

contacten onderhouden, iets nieuws ondernemen, alcohol zeer met mate, enzovoort, als ik niet inzie wat ze te maken hebben mijn gedrag thuis en op het werk.

Het ATI onderzoek van Cronbach en Snow Dit onderzoek is een bekende voorloper van de DBCs. Het gaat over kenmerken van de cliënt x zijn sociale context x *treatment*. Er is theorievorming over individuele verschillen, ontwikkeling en de context. *Treatment* is in dit verband de manipuleerbare context in de vorm van behandelprogramma's. In zijn *Presidential Address* voor de Amerikaanse psychologenvereniging (APA) heeft Cronbach (1957) correlationeel en experimenteel onderzoek omschreven als *The two Disciplines of Scientifc Psychology*. Weer zo'n tweedeling maar ze schiep duidelijkheid. Later heeft hij ze in zijn boek met Snow van 1977 gecombineerd in het ATI systeem. Het idee is dat het effect van een behandeling afhangt van het niveau op een aantal individuele verschillen variabelen van personen en hun contexten. Het effect van een onderwijsprogramma hangt bijvoorbeeld af van de intelligentie, motivatie en creativiteit van de leerling en van de context: ouders, school, buurt, SES. Hattie heeft er intussen veel over bijeengezet.
Het ATI systeem kan ook op behandelingen worden toegepast met als gevolg dat ze aangepast worden aan de individuele cliënt. De verwachting is dat afstemming tot betere resultaten leidt dan wanneer de gemiddeld genomen meest werkzame behandeling wordt ingezet. Cronbach en Snow vonden weinig empirische steun in hun boek met voorbeelden uit onderwijs en therapie. Hun hoofdstuk 12 gaat bijvoorbeeld over *Personality x Treatment interactions.* De behandeling van angst is het thema. In *Handelingsgerichte Diagnostiek* en *Diagnostiek en Leerlingbegeleiding* wordt niet naar dit standaardwerk over ATI en DBC verwezen. Synergie, interactie, afstemming tussen diagnose en behandeling zijn niet in een handomdraai bewerkstelligd. Het vergt veel n = 1 en groepsonderzoek.

Is er na Cronbach en Snow iets veranderd? Lukt na 40 jaar die synergie, afstemming, integratie wat meer? Ik neem als voorbeeld de drie theoretische oriëntaties in de psychologie: individuele verschillen, ontwikkeling en sociale context. Is er inter- en sub-disciplinair onderzoek en is het resultaat synergie? Als we de drie oriëntaties van psychologische theorievorming twee bij twee combineren blijkt dat Individuele Verschillen x Ontwikkeling door dominantie van de eerste tot stabiliteits- en continuïteitsonderzoek heeft geleid. Typerend gedrag en intellectuele prestaties komen steeds als stabiel uit de bus. Stabiel verwijst naar absolute stabiliteit (gelijke gemiddelden over leeftijd), rangorde (de stabiliteitscoëfficiënten van tests en vragenlijsten), structureel (bijvoorbeeld vergelijkbare factorpatronen door de tijd heen), ipsatief (binnen een persoon dezelfde rangorde: weinig onderzocht) en processtabiliteit (overgangskansen van verschillende gedragingen blijven gelijk door de tijd heen: weinig onderzocht). Continuïteit wordt gereserveerd voor stabiliteit van gedragingen die er weliswaar verschillend uitzien maar dezelfde betekenis hebben, denk aan de hechtingcategorie op jeugdige leeftijd en de kwaliteit van persoonlijke relaties in de volwassenheid. We leren keer op keer hoe stabiel persoonlijkheid en cognitie zijn.

Onderzoekers zijn er kennelijk niet op uit ons te leren hoe personen veranderen terwijl veranderingscoaches volop werk hebben en succes claimen. Hoe kan dat?
Hoe gaat het met de combinatie Individuele Verschillen x Sociale Context? Integratie, synergie, afstemming? Het persoon-situatie debat is voorbij. De trektheoretici hielden vast aan trans-situationele trekken en de situationisten probeerden de variantie in criteria te verbinden met variatie in situaties. En, een beetje theorie brengt ze van meet af aan bijeen zoals Eysenck al deed met situatie-specifieke responsen als onderste laag in zijn hiërarchisch persoonsmodel. De combinatie levert niet meer op dan dat een bescheiden hoeveelheid variantie in een criterium door situatie en trek gebonden kan worden.
Ten slotte laat Ontwikkeling x Context dominantie van de te manipuleren context zien. Piaget zag de sociale context voornamelijk als *aliment* voor de cognitieve ontwikkeling. In onderzoek werd geprobeerd die ontwikkeling te versnellen. Een voorbeeld is het bevorderen van de overgang van pre-operationeel naar concreet operationeel door training van deeltaken bij de conservatieproeven. Het gaat daarbij niet om Piagetiaanse structuren, maar om leer- en informatieverwerkingsprocessen. Die kunnen we immers - op papier - analyseren, isoleren, in een stappenplan onderbrengen en dan trainen. Dit onderwerp is nu van de onderzoeksagenda verdwenen. Het succes was beperkt en het onderzoek van korte duur. Loevinger (1997) werd de psychometrie ontrouw en gaf zich aan de persoonsontwikkeling. Ze beweerde dat het effect van een training eerst zichtbaar kan worden als een bepaald ontwikkelingsniveau bereikt is. Een plausibele gedachte maar het is geen populair onderzoeksthema en bijgevolg is er niet veel over bekend.
Er zijn congressen met de belofte goede DBCs te laten zien. U kunt er in de winter van 2015/2016 naar toe in mooie Zwitserse oorden, of U bent er inmiddels al geweest. Een titel als 'MAATWERK, de weg naar *precision therapy* voor uw patiënten' laat zien hoezeer de DBC - in dit geval met de nadruk of de B - in de belangstelling staat. Ja, de ene patiënt is de andere niet. Wanneer psychologen het initiatief nemen tot een dergelijk congres ligt de nadruk op de D zonder de behandeling uit het oog te verliezen.

Wat nu? Je mag en kunt vaststellen dat voor het uitzoeken en onderzoeken van het effect van DBCs nog werk te doen is. De praktijk loopt voorop in de toepassing. Bestuurders zien DBCs wel zitten. Het ligt immers intuïtief zeer voor de hand. Niettemin is maatwerk op basis van ATI onderzoek en integratie van theoretische oriëntaties niet serieus van de grond gekomen. De dominantie van het experiment is gebleven zoals blijkt uit de nadruk op *evidence-based* behandelingen. Antipestprogramma's zijn daar een voorbeeld van. De kinderombudsman wil wildgroei in deze programma's beperken. Er zijn er meer dan 100 in Nederland. Er moeten een paar bewezen effectieve overblijven. Hij zegt daarbij te steunen op wat de wetenschap - zijn adviseurs - hem verteld heeft. Dat is één element van de DBC: de kwaliteit van behandelingen. Om effecten van de Bn te bewijzen is er in de empirisch-analytische traditie maar één weg: het *true experiment*. Fisher heeft dat uitgelegd met veldjes groente. De binnenvariantie moet in de hand gehouden worden en de

tussenvariantie moet op basis van de experimentele variabele het verschil maken. De binnenvariantie wordt groter naarmate de steekproef op allerlei variabelen verschilt. Men kan ze proberen weg te filteren met covariaten maar in de werkelijkheid kun je de groepen niet zo kiezen dat ze covariaatvrij zijn. En, je doel is meestal niet iets te ontwerpen voor een zeer specifieke homogene groep. Je maakt het programma voor een gemêleerde groep. Je sluit geen meisjes of jongens, ouderen of jongeren, lage-hoge SES groepen uit. Dat zijn de bekende covariaten. Het experiment is geschikt om utilitaristisch vast te stellen wat het beste programma is voor een zo groot mogelijke groep. We kennen de d-waarden en dat leidt tot bescheidenheid over de effecten van de Bn.

Maken we het onszelf niet te moeilijk met de DBCs? Misschien worden de DBCs te complex opgezet: te veel covariaten, risico en beschermende factoren en qua duur en omvang te complexe behandelingen. Er kan onderweg zo veel uit de hand lopen dat je door de bomen het bos niet meer ziet. Lukt het met bescheidener opzetten om iets meer over een DBC te weten te komen? Misschien is een variant op adaptief testen een mogelijkheid om zicht te krijgen op DBCs. Het wordt gedaan met 'Rekentuin' en 'Taalzee' van de UVA. Ze zijn te vinden op: *http//:www.oefenweb.nl.* De afhankelijke variabele kan geschaald worden als een individuele verschillen dimensie van een groep of als een ontwikkelingsdimensie van een leerling. De behandeling bestaat uit vele precies gekozen en uitgeteste opdrachten die passen bij de plaats op de latente trek of de ontwikkelingsschaal van de individuele leerling. Je hebt er een grote voorraad aan precies metende taken en opdrachten voor nodig.
Misschien is het mogelijk dat zorgvuldig opgezette en gemonitorde *casestudies* stukje bij beetje zicht geven op DBCs. Deze kunnen gegeneraliseerd worden naar soortgelijke *cases*. Je komt ze in tijdschriften niet tegen. Gigerenzer (2008) noemt dit een verlies. Hij vertelde dat hij enkele artikelen in het *Journal of Experimental Psychology* van de jaren 20 tot 30 van de vorige eeuw had gelezen. Hij merkt op dat er iets verloren is gegaan: verschillende (statistische) methoden, precieze rapportage van individuele gevallen, zorgvuldige selectie van proefpersonen, nagaan of de sekse van de proefpersoon of proefleider van invloed was. Dit zijn precies de covariaten die er bij voorbaat uitgehaald worden om succes van een behandeling te bepalen.
Ondertussen kun je practici, diagnostici en behandelaars aanmoedigen door te gaan met zoeken naar de succescombinatie bij individuele cliënten. Er valt van hen iets te leren als ze vastleggen hoe en waarom ze de DBC in een concreet geval zo vormgegeven hebben en wat het resultaat was op korte en middellange termijn. De werkgever/organisatie zou hen kunnen toestaan om één casus per maand zo volledig mogelijk uit te werken, te bespreken en te rapporteren. Daar horen ook *niet* succesvolle DBCs bij. Misschien is er zelfs een tijdschrift of website te maken dat ze wil publiceren. Je moet wel de moed hebben om je niets aan te trekken van de slagzin: *Scientia non est individuorum*. Aan de hand van DBCs van individuele cliënten kan een kennisbestand opgebouwd worden. Natuurlijk is dit eerder

geprobeerd en het resultaat was beperkt. Toch kun je pleiten om het met de kennis van nu nog eens te doen.
De overtuiging en de intuïtie van bestaan van synergie tussen diagnose en behandeling blijven, ook al heb je niet veel bewijs in handen. Het verdient uitgezocht te worden. Je kunt daarbij buiten het ontwerp van *true experiment* met een experimentele en controle groep gaan en onderzoek van *cases* publiceren. DCBs zijn in de medische wetenschap aanvaard en daar spiegelt een diagnosticus zich aan. DBCs zijn ook aantrekkelijk om de bijna marxistische gedachte dat we wel veel weten maar er weinig mee doen. Is kennis er om de wereld te veranderen, je gedrag te verbeteren? Ja, een beetje en het lukt misschien ook wel een beetje.

Samenvatting en conclusie

DBCs liggen voor de hand. Er zijn voorlopers, bijvoorbeeld de opvatting dat freudiaanse therapie geschikt is voor hoogopgeleiden en gedragsmodificatie voor de gewone man. Wat later is er Cronbach en Snows *Aptitude-Treatment- Interaction* (ATI) model. Hoewel dat de basis is voor de DBCs wordt door hulpverleners niet naar dit model verwezen. Ze hebben dan ook niet kennis genomen van het feit dat er weinig steun was voor het ATI model. De hulpverlening liep en loopt niettemin voorop in het verbinden van kenmerken van de cliënt en de soort behandeling. In Nederland zijn er voorbeelden voor leerlingbegeleiding en voor kinderen met gedragsproblemen. Deze DBCs zijn meestal ontworpen op basis van een waargenomen verband tussen de stoornis en de behandeling. Dit is vooralsnog een hypothese. Je zult met experimenteel bewijs moeten komen. De hypothese vraagt complex experimenteel onderzoek om op steekproefniveau te tonen dat de samengestelde interventie verschillend uitpakt bij cliënten met verschillende cognitieve en persoonskenmerken uit uiteenlopende sociale contexten. Dat bewijs ontbreekt nog voor een groot deel. Het effect van een behandeling wordt als *self-evident* of logisch wordt beschouwd. Omdat de basis voor de DBC een waargenomen overeenkomst is tussen (on)gewenst gedrag en de behandeling is er het risico van *face validity.* Dit verschijnsel kan ook de andere kant opwerken: stel je vindt een DBC maar de autoriteit die de invoering ervan bepaalt wijst de combinatie af omdat hij het verband niet ziet. Er is volop werk om DBCs te ontwikkelen die bewezen effectief zijn. Misschien moet je de weg teruggaan en vanuit zorgvuldige *case studies* DBCs opbouwen. Dit past overigens niet in experimenteel onderzoek dat steekproeven vergelijkt. Misschien moet je dat naast je neerleggen en inductief DBCs voor individuele cliënten opbouwen.

6. Reflectie en evaluatie

De diagnosticus moet zijn plaats bepalen en veroveren in het DBC debat. Dat gaat over de relatie tussen Ds en Bn en die tussen diagnostici en behandelaars. Beide(n) hebben te maken met hun kennisbestanden. Deze dwingen tot bescheidenheid over de resultaten en leiden soms tot twijfel. Bij *daten* moet je echter niet te veel twijfelen om op stoom te

komen. Je moet de ander de ruimte geven. Als de behandelaar geen spoor van twijfel vertoont over zijn behandeling en verwacht dat de diagnosticus voedsel verstrekt om zijn behandeling te verbeteren, hoef je geen synergie te verwachten. Dit was de houding van de psychiater/behandelaar Kuiper ten opzichte van Kouwer, de diagnosticus/twijfelaar naar aanleiding van zijn boek *Het Spel van de Persoonlijkheid*.

De *treatment* verstrekker is hulpverlener. Uit de aard van zijn werk is hij handelingsgericht. In het boek van De Bruyn et al. (2015) is ruimte gereserveerd voor handelingsgericht werken. De behandelaar wil/kan niet toekijken. Hij wil iets doen aan een nood. Dat brengt hem in een speciale positie ten opzichte van de cliënt. Hij geeft, de cliënt ontvangt. Wat hij geeft komt uit zeker een goed hart maar het moet ook tot verbetering leiden. Doorgaans gelooft hij in zijn behandeling (denk aan de psychiater Kuiper, hierboven). Deze is immers logisch en aandacht en zorg zijn niet te ontkennen werkzame ingrediënten om het gedrag van de cliënt te veranderen of draaglijk te maken.

Als de behandelaar geconfronteerd wordt met de wetenschapper die eerst bewijs wil dat zijn program effect sorteert kan hij bogen op zijn ervaring en eerdere successen. Als dat wordt afgedaan als anekdotisch en klinisch kan hij terugvallen op de *evidence based* interventies. Als de wetenschapper zegt dat het effect van interventies meestal bescheiden is (en op het niveau van een steekproef) dan laat hij dat voor wat het is. Er moet immers iets gedaan worden. Moet ik dat probleem dan op zijn beloop laten? Moet ik die vraag dan onbeantwoord laten? Nee toch. Als de behandelaar nauwelijks zou geloven in zijn interventie voert hij die niet of kwansuis uit. Hij heeft nochtans enig - niet al te naïef - geloof nodig om op gang te komen en te blijven. Hij doet er van alles aan: interventies maken en benutten, zoeken naar effectieve ingrediënten, uitkijken voor iatrogene effecten en ontsporing en hij voert *true experiments* en klassieke n = 1 studies uit om meer aan de weet te komen. Hij is op de hoogte van het Mattheus effect dat zegt dat behandelingen verschillend effect hebben binnen subgroepen en wel zo dat de 'rijken rijker en de armen armer worden'. Als de diagnosticus twijfelt over zijn diagnose kan hij de *date* vergeten. Je gaat niet met een potentiële *loser* in zee. Niettemin is er bij de diagnose kans op twijfel. Zijn categoriseringen en voorspellingen van criteriumwaarden bevatten fouten. En, de behandelaar twijfelt aan de meerwaarde van zijn werk.

Om met de metafoor van relatie te besluiten: D en B beoefenaars hebben kans van slagen als elk erkent niet perfect te zijn, elk niet de dominante partij wil zijn op basis van schoonheid en *good breeding* (pikorde in de wetenschap). Elk streeft synergie na zonder volkomen in elkaar op te willen gaan, maar ook niet strikt hun eigen weg willen gaan. Ze kunnen af en toe door één deur gaan.

Hoe beschouwt de behandelaar de DBCs? Hij vat de relatie tussen *treatment* en gedrag impliciet unidirectioneel op. Zijn interventies zijn willekeurig ontworpen contexten (gedragsmodificatie, therapie, training, curricula). Bij vergelijkbare effecten van interventies kan hij effectieve ingrediënten op het spoor komen en benutten. Onder druk van de autoriteiten moet zijn therapie, behandeling, het curriculum een bewezen effect hebben:

Evidence-Based Psychological Treatments: EBPTs. Hij weet dat medische behandelaars vaker het label *effective* op hun conto schrijven. Zijn psychologische hulpverlening belandt meestal in de *promising* of *doubt* categorie. De behandelaar is cliënt- en probleemgericht en kent de beperking van EBPTs. Ze vermelden gemiddeld succes in een steekproef. Er is geen sprake van maatwerk in de zin van het ATI ontwerp en de gemiddelde effecten zijn bescheiden. Dat is niet prettig maar de feiten zijn de feiten en hij moet betere interventies maken.

Met de interactieve diagnose als behandeling heeft hij niet veel op. Het is een *sort of* therapie. De DBC houdt voor hem in dat hij nagaat onder welke voorwaarden een therapie en welke interventie het beste bij deze cliënt en zijn sociale context past. Daarbij betrekt hij de tijd en inspanning voor hem in voor de cliënt: past de behandeling in het dagelijks leven. De bescheiden d-waarden van *treatments* suggereren dat andere factoren dan de die van de *treatment* werkzaam zijn. Hij heeft daar geen weet van en dus geen controle op. Er zijn onverwachte effecten en die berusten op toeval. Ze zijn voor behandelaar en cliënt een verrassing. De behandelaar probeert open en transparant te zijn ten opzichte van collega's en cliënten. Dat is tricky want de verhouding is niet gelijk: er is een hulpbehoevende en een hulpverstrekker. Het is de vraag of de cliënt gehele transparantie verwacht.

Als behandelaar beschikt hij over een arsenaal aan *treatments.* Hij verwacht van de diagnosticus als die zich aanbiedt om bij de keuze van een *treatment* behulpzaam te zijn. Hij gaat er daarbij impliciet vanuit dat er verschillende behandelingen, procedures zijn die bij verschillende problemen passen. Dat opent de weg voor het vinden van BDCs. Hij vertrouwt doorgaans op zijn behandelingen en al doet hij dat niet, hij voelt zich toch geprest iets te doen, ook als er geen bewezen effectieve behandeling (B) en geen aanvaarde Diagnose-Behandel-Combinatie bekend is. Terughoudendheid in het verrichten van een behandeling op grond van het bescheiden effect is doorgaans geen optie. Een probleem kunnen aanzien is wellicht een competentie die evengoed past als kordaat ingrijpen. Ten slotte een quote van Freud (1925) in zijn Ten geleide van *Verwahrloste Jugend* (Aichhorn *Werken 9*: Amsterdam Boom). Hij geeft de positie weer die de behandelaar inneemt of zich toe eigent. Hij wil drie onmogelijke beroepen tegelijk uitoefenen: *Erziehen, Kurieren* en *Regieren*.

Onderwerpen en namen Hoofdstuk I en II

Idiografische en nomothetische benadering
Begrijpen versus verklaren: hermeneutiek versus empirisch-analytisch
Idiografisch en nomothetisch
Biografische en psychografische methode
Actuarische tabellen
Gecontextualiseerde wetten over gedrag
Kritische dialoog versus discussie en competitie
Objectieve tests en vragenlijsten
Normen
Diagnosticus: determinist
Statisticus: probabilist
Verschil klinische predictie en statistische predictie volgens Meehl
Combinatie van diagnostische gegevens met het hoofd en met de formule
Valse positieven en valse negatieven
Lineair model om voorspellers te combineren
Kleinste kwadraten oplossing
Meta-studie
Effectgrootte (d-waarde: Cohen)
Zwakten van informatie verwerven en integreren:
- *overconfidence*
- *heuristics'* en *biases*
- *illusory correlations*
- *confirmation bias*

Meehl vs Brunswik
Kahneman: heuristieken en vertekeningen
- *Availability*
- regressie naar het gemiddelde

Naturalistic fallacy
Criteriumgedrag
Intuïtie
Compensatoire lineaire modellen
Normatieve regels
Base rate: incidentie van een verschijnsel
Intuïtieve statistici
Compensatoire en conjuncte regel van combineren
Sign rule
Wicked environments
Bootstrapping (ook schoenveter-theorie of Münchhausen dilemma)
Ecologische structuur van de omgeving

Diagnose-Behandel-Combinaties (DBCs)
Meerwaarde
Evidence-based treatments
Werkzame ingrediënten van *treatments*
Aptitude-Treatment- Interaction
DBC als maatwerk
Face validity van voorspellende tests
Face validity van *treatments*
Case studies

III Diagnosticeren en betrouwbaarheid en validiteit

Lezers, die deze onderwerpen uitvoerig bestudeerd hebben in een ander kader, bijvoorbeeld het psychometrie onderwijs kunnen dit hoofdstuk overslaan. De aanpak in dit hoofdstuk bevat een reconstructie van betrouwbaarheid en validiteit vanuit het diagnosticeren als georganiseerde activiteit en dat is meer dan testgebruik of cliënten 'testen'.

Betrouwbaarheid en validiteit komen vooral aan de orde in het kader van tests en vragenlijsten. Ze gelden echter voor elke wijze van diagnosticeren en voor elk sociaalwetenschappelijk onderzoek. Ze hebben betrekking op het vaststellen van correlaties, het effect experimentele manipulaties en treatments, afhankelijke variabelen en beslissingen. Ze hebben ook betrekking op de diagnose en het diagnostisch proces. Over de betrouwbaarheid van de diagnose en het proces is mij niet veel empirisch onderzoek bekend. Herhaling kost tijd en geld en dan komt het er niet van. De validiteit van de diagnose lag onder vuur in de statistisch-klinisch controverse. De validiteit van het diagnostisch proces is in procedures als het HTM een conceptuele aangelegenheid dat wil zeggen worden de stappen helder omschrijven en achtereenvolgens uitgevoerd. Er nauwelijks onderzoek dat verschillende procedures vergelijkt. Er is wel nagegaan of de voorschriften van het HTM gevolgd worden.

Betrouwbaarheid gaat in eerste instantie over de herhaalbaarheid van gedrags- en omgevingsmetingen. Er wordt niet meteen gedacht aan de betrouwbaarheid van de diagnose (bijvoorbeeld nadruk op een second opinion zoals bij medici) en van het proces. Validiteit verwijst in de diagnostiek naar geldige, ware uitspraken over het gedrag van de cliënt en zijn context met het oog op een betekenisvolle beschrijving, voorspelling, verklaring/controle en beslissing.

Dit hoofdstuk organiseer ik met behulp van drie bronnen voor theoretiseren: impliciete, expliciete en alternatieve theorieën en drie theoretische oriëntaties: individuele verschillen, ontwikkeling en sociale context. Welke impliciete opvattingen van betrouwbaarheid zijn er?. Hoe is de expliciete betrouwbaarheidstheorie uitgewerkt? Welke statistische theorieën over testscores zijn er? Hoe worden indices gemaakt om betrouwbaarheid te schatten? Welke informatie geven die indices de diagnosticus? Zijn er naast schattingen van betrouwbaarheid van tests schattingen van betrouwbaarheid voor andere diagnostische procedures? Hoe hoog moet een betrouwbaarheidscoëfficiënt zijn?

Waarom is validiteit van belang voor leken en professionals? Welke validiteits- of waarheidscriteria worden door leken gebruikt en impliciet beaamd? Welke expliciete validiteitsconcepten zijn uitgewerkt voor diagnostiek? Welke typen validiteit worden onderscheiden? Wat laat de geschiedenis van de concepten zien? Is er eenheid, een lijn in te herkennen?

Betrouwbaarheids- en validiteitstheorie nemen de diagnostiek op de korrel: is je diagnostisch onderzoek en je diagnose betrouwbaar en valide? Heeft de diagnosticus er iets aan? Naast testvaliditeit wordt gesproken over research validity. Wat betekent dat? Hoe hoog moet een

predictieve validiteitcoëfficiënt zijn? Zijn er alternatieve opvattingen voor test- en research (onderzoeks)validiteit?

1. Impliciete opvattingen over betrouwbaarheid

In alledaagse taal verwijst betrouwbaarheid naar een kenmerk van personen. Ten tweede, worden in het dagelijks leven - in tegenstelling tot in de wetenschap - fouten niet als een opgevat als toeval maar als iets dat te herstellen is. Een persoon wordt betrouwbaar genoemd als hij iemand is op wie ik kan vertrouwen, die eerlijk, consistent en stabiel is. Het begrip wordt ook gebruikt voor voorwerpen en instituties, bijvoorbeeld auto's, politiek, informatie, het weer en voorbehoedmiddelen. Het is een sociaal wenselijk kenmerk van personen. De Franse humanist Montaigne (1533-1592) schreef in zijn kasteeltoren, wachtend op zijn dood, prachtige essays over menselijke eigenaardigheden. Hij noemde *consistentie* de grootste deugd en gaf toe er niet in te slagen consistent te zijn.

Ouders stellen prijs op eerlijkheid, rechtvaardigheid, onafhankelijkheid, openheid en 'respect' van hun kinderen. Dit is sociaal wenselijk en ze weten dat de praktijk anders is. Werkgevers waarderen deze kenmerken bij hun personeel: integriteit, betrouwbaarheid en gewetensvolheid. In een survey onder 3000 werknemers bleek IQ eerst op de vijfde plaats te komen (*Michigan Employability Survey*, 1986). Er is onderzoek naar integriteit van werknemers. Dit kenmerk correleert in een meta-studie van Sackett (1994) met r = .41 (SD = 0,05) met criteriumgedragingen als (on)verantwoordelijk zijn, (contra)productiviteit, disciplinaire problemen, afwezigheid, sabotage en de kantjes eraf lopen (Viswesvaran & Ones, 1999). Schade wordt niet overigens alleen door werknemers veroorzaakt maar ook door CEOs en Raden van Bestuur. Hun integriteit wordt vanzelfsprekend geacht en dus niet getoetst. Na de bankencrisis, de bouwfraude, de Noord Zuid metrolijn, de besturen van woningcorporaties is dat niet meer vanzelfsprekend.

In onderzoek van 1928 werd betrouwbaarheid opgevat als een situatiegebonden individueel verschil. Hartshorne en May deden onderzoek naar die verschillen en concludeerden dat betrouwbaarheid en gewetensvolheid geen stabiele persoonlijkheidstrekken waren, maar veeleer afhingen van de situatie, bijvoorbeeld het verschil tussen thuis, werk of sportclub. De derde factor van de *Big Five* persoonskenmerken Gewetensvolheid wordt evenwel als een stabiele trek opgevat. Termen als methodisch, georganiseerd, efficiënt, verantwoordelijk, betrouwbaar en slordig, onverantwoordelijk, onvoorspelbaar, frivool en vergeetachtig hebben hoge positieve respectievelijk negatieve ladingen op deze factor (Hendriks, 1997). Het is de enige factor van de Vijf die consistent en significant samenhangt met school- en werkprestaties. Bij 974 Utrechtse 1ste jaars psychologiestudenten was bijvoorbeeld de relatie tussen de som van de propedeuse tentamencijfers en de factor Gewetensvolheid r = .19. Deze waarde wordt ook in meta-studies aangetroffen.

In de alledaagse taal wordt betrouwbaarheid nauwelijks onderscheiden van validiteit. De uitdrukking 'de betrouwbaarheid van het ijs in de winter' wijst bijvoorbeeld op validiteit. In de testpsychologie zijn de twee onderscheiden. De eerste gaat over de herhaalbaarheid van het meetresultaat en de fouten die zich voordoen bij de testscores. De tweede verwijst vooral naar wat gemeten wordt en welk gedrag ermee voorspeld kan worden.

Fouten vatten we alledaags op als iets dat voorkomen en hersteld kan worden: je kunt leren van je fouten. In de diagnostiek gaat het om onvermijdelijke fouten door processen waar we geen greep op hebben. Meetfouten zijn willekeurig en daarom niet te sturen en niet te herstellen. Ze zijn niet systematisch zoals bij een verkeerde afstelling (kalibratie) van een thermometer. *At random* of willekeurige fouten doen zich voor als gevolg van niet te controleren omstandigheden, gebeurtenissen en personen. Ze verwijzen naar fluctuaties in aandacht, lichamelijke conditie, efficiëntie, concentratie, impulsiviteit en onzorgvuldigheid. De betrouwbaarheidsbepaling van diagnostische instrumenten laat zien in welke mate we staat kunnen maken op geobserveerde scores. De enige manier om een beeld van de meetfouten te krijgen is deze te schatten door herhaalde metingen.

Samenvatting en conclusie

In het leven van alledag wordt van personen, instellingen en verschijnselen gezegd dat ze in meer of mindere mate betrouwbaar zijn. Het is een sociaal wenselijk kenmerk en wordt opgevat als een individuele verschillen variabele. Eerst werd ze als situatief bepaald beschouwd. (Hartshorne & May). Later als stabiele trek, bijvoorbeeld als factor III van de *Big Five*. In het dagelijks spraakgebruik wordt, in tegenstelling tot in de testpsychologie, nauwelijks onderscheid gemaakt tussen betrouwbaarheid en validiteit. Fouten worden niet als *random* verschijnselen opgevat maar als iets dat voorkomen of hersteld kan worden. Meetfouten zijn onvermijdelijk, want willekeurig. Oorzaken worden gezocht in fluctuaties binnen personen, de context en de onnauwkeurigheid van het meetinstrument. Betrouwbaarheids- en validiteitstheorie bevatten kenmerken waaraan de diagnosticus moet voldoen wil hij een geldige uitspraak over het gedrag van de cliënt doen. Hij kan er niet omheen. Onbetrouwbare en niet valide uitspraken kun je terzijde schuiven als onjuist.

2. Expliciete opvattingen over betrouwbaarheid

In IV wordt het idee van herhaling bij de schatting van betrouwbaarheid van een meting besproken. Herhaling levert geen identieke uitslagen op en de variatie is een index voor betrouwbaarheid van een test. Dit idee is uitgewerkt is voor tests en vragenlijsten en minder voor diagnose en diagnostische procedures. Testtheorie (KTT en IRT) heeft het monopolie op de bepaling van relaties tussen de items van een test. Dit heeft geen theoretische grond want elke multivariate techniek is geschikt om de samenhang tussen items, tests en criteria te bepalen.

Indices Deze worden gemaakt door

(1) herhaalde meting met dezelfde items bij de dezelfde persoon of steekproef

(2) herhaalde meting binnen de test met items die dezelfde trek meten, bijvoorbeeld de even-oneven verdeling van items.

De eerste manier leidt tot test-hertest coëfficiënten. Verdeling in alle mogelijke helften levert interne consistentie schattingen op. De test-hertest procedure vooronderstelt dat de antwoorden stabiel zijn over een zekere periode. Test-hertest coëfficiënten zijn een zinvolle index onder specifieke condities want een lage waarde kan ook het gevolg zijn van verschillen in gedragsverandering bij elementen van een steekproef. In dat geval is de index niet geschikt. Om te bepalen of er differentiële verandering is opgetreden moet men iets weten over verandering en veranderlijkheid van het gedrag van de cliënt. Dat is geen meetprobleem maar een psychologisch vraagstuk. Bij berekening van interne consistentie coëfficiënten gaan we ervan uit dat de items één en dezelfde trek meten. Elk item wordt beschouwd als een (niet) lineaire functie van hetzelfde theoretische attribuut. Als er twee verschillende, maar betrouwbaar gemeten trekken in de items gerepresenteerd zijn die niet perfect gecorreleerd zijn is de test niet onbetrouwbaar. Hij meet verschillende trekken.
Instrumentenmakers moeten het hebben van homogene unidimensionele tests die het optellen van de items toestaan. Betrouwbaarheidstheorie is een groot woord voor enkele assumpties die nodig zijn om meetfouten te schatten. Ze resulteren in indices, zoals test-hertest, paralleltest en interne consistentie coëfficiënten. Sommige auteurs, bijvoorbeeld (Lucke et al., 2005) laten het idee van de unidimensionele onderliggende trek los. Ze pleiten voor heterogene tests omdat *'... the complexity of (prosocial) behavior requires tests that are heterogeneous, measuring more than one attribute'* (p. 65). Deze opvatting houdt rekening met de samengesteldheid van criteriumgedrag. Het voorstel van Lucke et al. leidt tot een samengestelde betrouwbaarheid die bestaat uit de betrouwbaarheden van de verschillende attributen in de test. In de moderne testtheorie wordt de unidimensionaliteit getoetst.
Het is aannemelijk dat zelfs bij het minigedrag van het antwoord op een item niet één maar meer trekken of onderliggende vaardigheden betrokken zijn. We moeten immers bij het beantwoorden van een stelling uit een attitudeschaal de bewering begrijpen, een handeling uitvoeren om het antwoord te geven en in ons geheugen nagaan wat we vinden van de stelling.

Soorten indices De Kuder Richardson (KR) 20 is een maat voor interne consistentie. Het getal 20 verwijst naar het aantal. Er waren er toen al veel. Dat kan omdat er veel manieren zijn om relaties tussen antwoorden op items te bepalen. We beperken ons tot de grondvorm: herhaling. Dat wordt gerealiseerd door een test twee of meer keren af te nemen bij dezelfde persoon of dezelfde steekproef op verschillende tijdstippen: *occasions*. Dit levert een stabiliteitscoëfficiënt op. Men vooronderstelt dat de trek stabiel is door de tijd heen. Als zich verandering voorgedaan heeft leidt het tot een lage stabiliteitscoëfficiënt, maar dat wil niet zeggen dat de test onbetrouwbaar is. De test-hertest procedure past in dat geval niet. Daarvoor moeten we weten of er zich in de steekproef een differentiële verandering heeft voorgedaan. Dat is een verandering die zich bij sommige wel en bij anderen niet voordoet. Herhaling is ook te verwerkelijken door parallelle of alternerende

items: parallel en even-oneven betrouwbaarheid en met behulp van parallelle methoden. De werkwijzen berusten op hetzelfde concept: *tau equivalentie* maar leveren iets verschillende waarden op bij de procedures als de GLB: *Greatest Lower Bound* van Bentler & Woodward (1980), de lambda2 van Guttman (1945), een samengestelde coëfficiënt op basis van een structureel vergelijking model (SEM) en de bekendste: Cronbachs alfa (1951). Een combinatie van de wijzen van herhalen is de basis van de generaliseerbaarheidstheorie van Cronbach et al. (1970). De variantie in testscores is samengesteld uit een ware score en drie bronnen: consistentie, of generaliseerbaarheid over items; stabiliteit, of de generaliseerbaarheid van de scores over *occasions*. Zij voegen generaliseerbaarheid over oordelaars toe. Dit slaat op de overeenstemming tussen beoordelaars bij het toekennen van een reeks gedragingen aan een categorie bijvoorbeeld van de DSM-IV-TR, DSM-5, ontwikkelingsstadia, observatiecategorieën en observatie- en *rating*schalen. Generaliseerbaarheidscoëfficiënten worden zelden in testhandleidingen aangetroffen. Het is veel werk en als testgebruikers en -uitgevers tevreden zijn met de andere indices waarom dan nog die bewerkelijke generaliseerbaarheid bepaald? Men loopt daarbij ook het risico dat er verschillende waarden gevonden worden in vergelijking met de andere methoden. Dat roept meteen vragen op ook al is de hoogte van de coëfficiënt acceptabel. De waarde van de interne consistentie coëfficiënten loopt bovendien niet erg uiteen. Er is dan ook geen reden om de ene boven de andere te kiezen. Test-hertest waarden zijn - zo blijkt uit ervaring - lager: r tussen de .05 en .15.

De Item respons theorie (IRT) biedt een nieuwe vorm van betrouwbaarheidsbepaling en maakt het mogelijk een betrouwbaarheidsinterval voor elk item te bepalen. Een item is gekenmerkt door een itemkarakteristieke curve of responsfunctie (ICC of IRF). Hoe steiler de curve des te beter deze de positie van een persoon op de latente schaal schat. Het item is in die zin betrouwbaar want het geeft de positie op de latente schaal van de persoon aan zonder veel spreiding. Elk item heeft zo'n informatiefunctie die zijn betrouwbaarheid weergeeft. De testinformatiefunctie is de som van de iteminformatiefuncties. Deze vorm van betrouwbaarheid komt voor bij schoolprestatietoetsen en wordt steeds meer bij intelligentie- en persoonlijkheidstests toegepast.

In principe zijn veel multivariate technieken te gebruiken voor betrouwbaarheidsbepaling van items en tests. Cheng et al. (2011) vergeleken indices gebaseerd op factoranalyse en IRT. De IRT betrouwbaarheid voor ja-nee items is iets lager dan de indices van het factor model. IRT maakt een gewogen somscore. De IRT stelt hogere eisen aan items dan de KTT en factor analyse. Voor predictie maakt het weinig uit.

Ik veronderstel dat personen kunnen verschillen in variabiliteit rond hun ware score. De een heeft mogelijk een kleinere betrouwbaarheidsinterval rond die score dan een ander. In IRT termen: de ene persoon heeft steilere IRFs dan een andere. De betrouwbaarheid van de testscore wordt zo gepersonaliseerd. Misschien heeft een opportunist, die telkens 'zijn kansen grijpt als het moment rijp is' een grotere spreiding rond zijn ware score dan een consciëntieus persoon. Wellicht is de een gevoeliger voor de situatie dan de ander. Dit zou

kunnen leiden tot bepalen van gepersonaliseerde en gesitueerde betrouwbaarheidscoëfficiënten. Dat gebeurt vooralsnog niet. We moeten het - om het niet al te ingewikkeld te maken - doen met één coëfficiënt die geldt voor de hele range scores van de steekproef. Wel zo gemakkelijk. Er zijn wel argumenten voor personaliseren gegeven. Bauer (2011) benadrukt dat er individuele verschillen zijn in psychologische processen. Hij pleit evenals Molenaar (2004) voor de idiografische aanpak. Waarom geen idiografische uitwerking voor betrouwbaarheid? Cliënten verschillen in accuraatheid, oprechtheid, interesse en geloofwaardigheid bij het beantwoorden van interviewvragen en het invullen van vragenlijsten en tests. Dat verschil kan ook gelden voor verhalen die ze de diagnosticus vertellen.

De diagnosticus komt meerdere indices tegen: de standaardmeetfout van een persoon (zelden), test-hertest bij steekproeven (regelmatig), paralleltests (vaak), interne consistentie (ook bepaald door middel van multivariate technieken) en stabiliteitscoëfficiënten (beide het meest), variantiecomponenten voor items, over tijdstippen en beoordelaars (zelden: generaliseerbaarheidstheorie) en de test informatiefuncties van de IRT modellen (steeds meer). Dit laat zien, dat

betrouwbaarheid en validiteit niet categorisch onderscheiden zijn. Bij test-hertest moet het gedrag stabiel zijn, bij interne consistentie moet er sprake zijn van één en dezelfde trek evenals bij de IRT.
er veel coëfficiënten zijn. Wittman (1988) noemde het een sport van psychometrici om steeds nieuwe te verzinnen. De bronnen zijn ook gevarieerd: parallelle tests, IRT en statistische technieken, die tot kleine verschillen leiden.
er pragmatische redenen zijn om een beperkt aantal betrouwbaarheden te onderscheiden op basis van parallelle tests. Dat houdt het berekenen in van inter-itemrelaties met statistische procedures en multivariate technieken want items zijn variabelen die niet alleen met behulp van de KTT en IRT kunnen worden geanalyseerd.
overeenstemming tussen beoordelaars de betrouwbaarheid van lineair geordende, getrapte procedures kan worden bepaald. Dit wordt zelden gedaan. Het is mogelijk bij Toulmins stappenplan van correct argumenteren, het HTM, de MAUT en het opstellen van regressievergelijkingen met gewogen predictoren.
betrouwbaarheid van de *diagnose* vrijwel nooit bepaald wordt door bijvoorbeeld herhaalde diagnoses of door bepaling van de interne consistentie. De *second opnion* komt vrijwel nooit voor bij de diagnose van een cliënt. Is dit op basis van het vertrouwen in de diagnose of vanwege de ondergeschoven positie van de psychologische diagnostiek: het maakt toch niet zoveel uit voor de behandeling en het is ook veel te kostbaar.
betrouwbaarheid van het *diagnostisch proces* evenmin bepaald wordt door een herhaling. Er wordt niet nagegaan of verschillende diagnostici op gelijke wijze handelen bij de diagnose van een cliënt.
interne consistentie niet bepaald wordt. Er zijn recent studies verricht waarin uitgezocht wordt of een voorgeschreven procedure gevolgd wordt, bijvoorbeeld het HTM. Dat blijkt beperkt het geval te zijn. Het is niet eenvoudig en bewerkelijk om betrouwbaarheid van complexe procedures en processen te bepalen. Het is echter wel zinvol. Gegeven het vele werk en de vermoedelijke uitslag

dat het toch niet zo repliceerbaar is als je vooronderstelt en zou willen, is dit type onderzoek bij de huidige publicatiedruk onwaarschijnlijk.
test-hertest coëfficiënten in de KTT beperkt zijn tot testscores maar stabiliteit en consistentie kunnen ook voor complexe gedragingen berekend worden.
IRT item- en testinformatiefuncties informatie geven over de betrouwbaarheid van het item en zijn constructvaliditeit. Het item meet immers volgens het model één latente trek.
de diagnosticus zélf moet uitzoeken welke coëfficiënt passend is en of hij hoog genoeg is voor de vraag/probleem van de cliënt. Bij voorspellen zijn test-hertest coëfficiënten geschikt. Wil hij weten of de items en taken verwant of onderscheiden zijn dan is informatie over interne consistentie zinvol, heeft hij een criterium nodig om op een schaal onderscheid te maken, bijvoorbeeld tussen wel of niet toelaten van kandidaten *(cut off* score, afbreek- of stopregel) dan is de testinformatiefunctie geschikt; de betrouwbaarheidsintervallen rond de verwachte kunnen ook behulpzaam zijn, maar de IRT modellen zijn preciezer.

Samenvatting en conclusie

Betrouwbaarheid is gebaseerd op herhaalbaarheid en consistentie van testscores. Een theoretisch idee is de standaardmeetfout bepaald bij een persoon. Theoretisch omdat men niet iemand steeds dezelfde vragen kan voorleggen. Er zijn concrete indices gemaakt gebaseerd op parallelltests, test-hertest en interne consistentiematen. Deze vooronderstellen dat gedrag stabiel is in de steekproef en dat items samenhangen door af te stammen van één unidimensionele trek. Generaliseerbaarheidstheorie vat de klassieke testtheorie in één kader: generaliseren over items, gelegenheden en beoordelaars. De IRT levert item- en testinformatiefuncties. Betrouwbaarheid en constructvaliditeit zijn in de IRT niet onderscheiden. De diagnosticus moet zelf uitzoeken welke coëfficiënt past bij de vraag van zijn cliënt. De indices van betrouwbaarheid stellen in staat om te zeggen binnen welke marge de ware score van een persoon zich bevindt. Over het bepalen van de betrouwbaarheid van andere diagnostische procedures, zoals interviewen, documenten, gedragsneerslagen (verhalen, dagboeken) stadia van ontwikkeling, observatiecategorieën, checklists is (nog) niet veel uitgezocht.
Betrouwbaarheid van de diagnose en het diagnostisch proces zijn nauwelijks gethematiseerd en bijgevolg niet empirisch onderzocht.

3. Diagnose en betrouwbaarheid

Diagnostiek valt niet samen met testen en testtheorie gaat vooral over item- en testscores. Een diagnosticus wil ook weten hoe het met betrouwbaarheid van categoriseren, bepalen van ontwikkelingsstadia, interviewen, *case formulation*, HTM en Checklists en Richtlijnen. Er zijn wel enkele voorbeelden maar het is niet standaard om bij deze procedures betrouwbaarheid te bepalen. Overeenstemming van diagnostici die checklijsten gebruiken, is gemakkelijk na te gaan maar het gebeurt zelden. Beoordelaarsovereenstemming bij categorietoekenning wordt bepaald maar dat geldt niet voor complexe procedures als HTM en *case formulations*.

Overeenstemming tussen beoordelaars is een index voor betrouwbaarheid. Er zijn latente categorische modellen. Deze zijn er vooral voor en door statistici. De diagnosticus moet het doen met tussen-beoordelaars-overeenstemming bij categorietoekenning van DSM- en observatiecategorieën, ontwikkelingsstadia, interviewonderwerpen, verhalen en gedragsneerslagen, zoals werkstukken, *files* van de cliënt of significante anderen zoals dagboeken of rapporten. De eenvoudigste bepaling is: het aantal overeenstemmende oordelen gedeeld door het aantal overeenstemmende vermeerderd met niet-overeenstemmende. Om te corrigeren voor kans is Cohens kappa (1960) geschikt. Stel, twee beoordelaars A en B verdelen honderd cliënten over drie categorieën: 1, 2 en 3. Beoordelaar A plaatst 50% in 1 en beoordelaar B 40% in 1, dan is er .50 × .40 = .20 overeenstemming op basis van kans. Het toepassen van de formule veronderstelt dat de oordelaars gelijkwaardig en competent zijn. Dat wordt zelden nagegaan. Cohens kappa ziet er zo uit:

(k)appa = P (waargenomen) – (verwacht) / 1 – P (verwacht)

Het ligt voor de hand om te controleren voor de *base rate* of incidentie. Als veel cliënten tot één categorie behoren dan is de kans op overeenstemming groot voor die categorie. Het niveau van overeenstemming tussen oordelaars, interviewers, observatoren en uitvoerders van protocollen is van belang om idiosyncratisch scoren en rapporteren te voorkomen. Cohens idee is verder uitgewerkt door Bakeman en Gottman (1997) en Goodwin (2001: zie zijn hoofdstuk 4). Voorbeelden:

(1) Mumma en Smith (2001) wilden de opzet en standaardisatie van *case formulations* onderzoeken en verbeteren. Tweetallen clinici kregen videotapes te zien met half gestructureerde interviews van vier cliënten met stemmings- en/of angststoornissen. De tweetallen formuleerden onafhankelijk twee tot drie *cases*: *Cognitive-Behavioral-Interpersonal Scenarios* (CBIS). Tien experts schaalden elk CBIS op vijftien dimensies met 9-puntsschalen bijvoorbeeld cognitie, affect, verschillende symptomen en interpersoonlijk functioneren. De betrouwbaarheid (r) was voor alle dimensies > 0.83. De CIBS van de clinici bevatten duidelijke informatie om ze betrouwbaar te kunnen schalen op vijftien dimensies. Betrouwbaarheidsbepaling van *case formulations* is tijdrovend maar nuttig. De diagnosticus leert of zijn formulering door andere clinici gedeeld wordt. Er zijn geen directe vergelijkingen tussen *case formulations* gevonden. Het voorbeeld van Mumma en Smith is een uitzondering. Men gaat er mogelijk vanuit dat het protocol uniformiteit garandeert. Of dat het geval is kan alleen blijken uit empirisch onderzoek.
(2) Onderzoek naar tussen-diagnostici-overeenkomst in het *hypothese toetsend model* is zeldzaam. Groenier et al. (2011) deden zo'n een studie. Acteurs speelden de cliënten en een aantal diagnostici interviewde hen. De diagnostici waren getraind in het toepassen van het HTM. Het interview werd op video opgenomen. De diagnostici schreven eerst een rapport en daarna werden ze ondervraagd over hun activiteiten en overwegingen gedurende het onderzoek. Er werd gevraagd naar de klachten van de cliënt, mogelijke verklaringen en behandelingen. De diagnostici waren het vaker eens over categorisering van de klachten dan over verklarende factoren. De studie liet ook zien dat ze

voortdurend door de voorgeschreven geordende stappen heen en weer gingen en niet de opgelegde volgorde aanhielden. Ze switchten van oorzaak naar behandeling en van behandeling naar de klacht en weer naar een hypothese. Ze herzagen hun beslissingen regelmatig en lichtten dat toe vanuit wat er tijdens het onderzoek gebeurde. HTM diagnostiek is in de praktijk een dynamisch proces waarbij het protocol niet strikt gevolgd wordt. Het viel op dat de hypotheseformulering minder centraal stond dan voorgeschreven en verwacht. Bovendien waren er individuele verschillen tussen diagnostici en het probleem van een cliënt leidde niet altijd tot een identieke diagnose.
(3) Checklijsten en richtlijnen bevatten ja-nee vragen. De overeenstemming tussen antwoorden van diagnostici kan bepaald worden. Als het om meer dan een schaal gaat kan de correlatie tussen beoordelaars berekend worden. Dit ligt voor de hand ligt. Er zijn echter nauwelijks betrouwbaarheidsbepalingen van checklijsten en richtlijnen.

Overeenstemming tussen activiteiten van diagnostici en hun beslissingen bij dezelfde cliënten worden kortom zelden empirisch bepaald. Het onderzoek is administratief en tijdrovend, maar het leert over idiosyncrasieën van diagnostici en haalbaarheid van de complexe procedures en protocollen.

Hoe hoog moet een betrouwbaarheidscoëfficiënt zijn Tests zijn producten die verkocht worden. Daarom wordt soms niet vermeld of alle onderdelen van een test voldoende betrouwbaar zijn. De diagnosticus kan een eigen koers varen waarbij de hoogte niet voor elk onderzoek dezelfde hoeft te zijn. Een leraar kan bijvoorbeeld taken gebruiken om het niveau van een leerling te schatten en zijn onderwijs aanpassen. Deze werkwijze hoeft niet perfect betrouwbaar te zijn, alleen al omdat het onderwijs zelf meteen informatie verschaft over het feit of hij er naast zit met zijn toets of schriftelijke overhoring. Hoge betrouwbaarheden zijn vereist bij het toelaten van kandidaten tot opleidingen, bij operators die complex en duur materieel bedienen, bijvoorbeeld piloten, computerdeskundigen en kraandrijvers. Het doel is vaak het beschermen van het materieel. Het is van minder belang of er kandidaten ten onrechte afgewezen worden. Nunnally en Bernstein (1994) formuleerden vuistregels voor de hoogte van betrouwbaarheidscoëfficiënten:

r < .80 is onvoldoende, r tussen .80 en .90 is voldoende, en r > .90 is goed bij tests en procedures om belangrijke beslissingen te ondersteunen op individueel niveau, zoals personeelsselectie, toelaten tot scholen, opnemen of ontslaan uit de kliniek. Belangrijke beslissingen zijn vaak onomkeerbaar en moeten soms worden genomen zonder dat de persoon toestemming kan geven, bijvoorbeeld bij opname in een psychiatrische kliniek.
r < .70 is onvoldoende, r tussen .70 r .80 is voldoende en r > .80 is goed voor minder ingrijpende beslissingen, bijvoorbeeld, voortgang bijhouden, beroepskeuze, therapie-indicatie. Daar is herstel mogelijk.
r < .60 is onvoldoende, r tussen .60 en .70 is voldoende en r > .70 is goed bij wetenschappelijk onderzoek van steekproeven en voor experimenteel gebruik van tests en andere procedures. Dezelfde waarden kunnen gehanteerd worden voor generaliseerbaarheidscoëfficiënten. De verwachting is dat de waarde iets lager is dan bij elke coëfficiënt afzonderlijk.

Samenvatting en conclusie

Indices voor tussenbeoordelaarsovereenstemming worden gebruikt om de betrouwbaarheid van categorietoekenningen te bepalen. Cohens kappa is een formule die de mate van overeenstemming berekent, gecorrigeerd voor overeenstemming op basis van kans. Betrouwbaarheid in de vorm van overeenstemming tussen beoordelaars bij complexe diagnostische procedures (formuleren van casussen, HTM, beslismodellen maken en toetsen, regressieformules opstellen met predictoren) wordt zelden onderzocht. Mogelijk wordt verondersteld dat het protocol overeenstemming garandeert. Empirische gegevens over interdiagnostici-overeenstemming bij gebruik van complexe procedures zijn nuttig voor de diagnosticus. Misschien wordt te gemakkelijk aangenomen dat de voorschriften protocollen evident en logisch zijn. Het is gerechtvaardigd om empirisch onderzoek te vragen dat verheldert hoe betrouwbaar diagnostici de formeel correcte voorgeschreven procedures, protocollen en programma's uitvoeren.

Nunnally en Bernstein hebben vuistregels voor de gewenste hoogte van de coëfficiënt voorgesteld. Voor belangrijke beslissingen met gevolgen voor het individu, zoals selectie en plaatsing is r > .90 goed; voor het bijhouden van gedragingen door de tijd, bijvoorbeeld op school is r > .80 goed, en voor onderzoeksdoeleinden is r > .70 voldoende.

4. Alternatieve concepten

Het concept betrouwbaarheid is gemunt door de klassieke en moderne testtheorie. De coëfficiënten zijn variaties op het herhaalprocédé. Alternatieven zijn een onderwerp voor psychometrici. Zij blijven strikt binnen hun kader en stellen af en toe een nieuwe coëfficiënt voor (Wittmann, 1988) of benadrukken het nut van een oude (lambda2: Sijtsma, 2009). Deze hebben weinig gevolgen voor de praktijk. Het ziet er ook niet naar uit dat de testtheorie een onderdeel van multivariate technieken wordt. Er is een afwijkend, gedateerd voorbeeld gevonden. Het is de opvatting van Lumsden (1978) dat testitems perfect betrouwbaar zijn. Ze zijn wat ze zijn, ze liggen als objecten Sartriaans 'verpletterd op zichzelf'. De enige bron van variatie en onbetrouwbaarheid is de persoon. Elk item staat op een plaats, punt van het attribuutcontinuüm, bijvoorbeeld verbale capaciteit. Elke persoon is gekenmerkt door een verdeling van attribuutlocaties door de moment tot moment fluctuaties. Deze verdelingen zijn normaal en de SD van de verdelingen kan van persoon tot persoon verschillen. Zo kan A een geringe SD hebben en dat duidt erop dat hij dat item op dat punt van het continuüm correct beantwoordt en weinig items die verder weg liggen. B heeft een naar verhouding grote SD. Hij zal meer items in de buurt correct beantwoorden. De Spearman-Brown formule geldt hier niet want als een test verlengd wordt zal de betrouwbaarheid niet toenemen. Is dit alternatief een slecht idee?

Samenvatting en conclusie

Er zijn nauwelijks alternatieven voor de testtheorieën, hoewel de samenhang tussen items en criteria met vele technieken bepaald kan worden. De testtheorie maakt uit hoe betrouwbaarheid gedacht en bepaald wordt. De coëfficiënten zijn variaties op het thema 'herhalen'. Alternatieven zijn een interne psychometrische aangelegenheid. Lumsden formuleerde in jaren 70 een alternatief waarbij hij ervan uitging dat items perfect betrouwbaar zijn. De fluctuatie, onbetrouwbaarheid is aan de persoon te wijten. En, personen verschillen in de SDs rond het item op dat punt in het continuüm dat ze correct beantwoorden en hun niveau op de trek uitdrukt. Het alternatief speelt geen rol in de huidige betrouwbaarsheidstheorie en - bepaling.

5. Impliciete opvattingen over validiteit

Validiteit, waarheid, geldigheid laten leken en professionals niet onverschillig. Ze hebben uitgesproken opvattingen over de geldigheid van beweringen over gebeurtenissen, verschijnselen en gedrag. Het concept wordt gemakkelijk gelijkgeschakeld met objectiviteit. We vatten objectiviteit hier op als recht doen aan het gedrag van de persoon (hoofdstuk 1). Dit mag vaag klinken maar dit is gekozen om open te staan voor betekenissen die gedrag en antwoorden op testvragen kunnen hebben. Objectiviteit betekent meer dan objectief scoren van items en tests.

Box Geen epistemologische stroming heeft de waarheid in pacht

Dat waarheid ons aanbelangt wordt pregnanter gezegd door dichters dan wetenschappers. Emily Dickinson (Fr1263, 1872):

Tell all the truth/ But tell it slant-/ Success in Circuit lies/ Too bright for our infirm delight/ The Truth's superb surprise/ As lightening to the Children eased/ With explanation kind/ The truth must dazzle gradually/ Or every Man be blind. In vertaling: *Zeg de Waarheid zijdelings/ Een omweg voert naar het doel/ Te fel is Waarheids grootste schok/ voor ons krank lustgevoel/Als bliksem rustig uitgelegd/Aan het beangstigd kind/ Moet waarheid lichten gaandeweg/ of anders maakt zij blind.*

Ze zegt dat waarheid moet en niet vanzelf en snel tot stand komt. Er is een omweg nodig, begrip, geduld en uitstel van de oplossing, anders zien we niets. De dichter Nijhoff zegt in 'Awater', waarin de hoofdpersoon op zoek is naar een vriend: *'...en het toeval nam een binnenweg naar het doel'*: het bereiken van een doel, een begeerd resultaat hoeft niet noodzakelijk rationeel, protocol- en procesgestuurd afgedwongen te worden. Dichters hebben het niet voor het zeggen in de diagnostiek; wetenschappers en professionals wel.

Er is natuurlijk meer dan één waarheids- en objectiviteitsconcept in de diagnostiek. Deze stand van zaken lokt naar goede westerse gewoonte dominantie van één of van een hiërarchie uit. Methodologische waarheid en objectiviteit, dat wil zeggen zich ontdoen/aan banden leggen van subjectiviteit - denk aan de klinisch statistische controverse en het HTM - is er slechts een, maar het

telt zwaar in de diagnostiek. De empirisch-analytische stroming domineert en is met kennistheoretische posities van empirisme en rationalisme verbonden in de samenstelling 'logisch positivisme'. Dat is een slim niet naïef empirisme. Deze stroming loopt het risico de onmiskenbare subjectiviteit van de cliënt te reduceren, soms tot in het extreem *zero subjectivity*. Zij maakt de diagnosticus tot een inwisselbare klerk. De cliënt wordt serieus genomen met zijn impliciete opvattingen. Bovendien blijkt de diagnosticus niet aan banden te leggen door protocollen en de twee gaan een verhouding aan en daar gelden de communicatieve *maxims* van Grice van non-redundantie en relevantie. De relatie kan verder gaan (Hermans: de relatie van mede-onderzoeker en helper. Deze stroming heeft ook 'hinder' van de contextgevoeligheid van het gedrag. Met die grilligheid kun je geen regel of wet formuleren want in iedere context komt er iets anders uit. Er is zelfs niet één betrouwbaarheidscoëfficiënt voor een test. Sommigen moeten dan ook niets hebben van zo'n objectiviteitsopvatting. Ze zoeken het bij een andere stroming, bijvoorbeeld de fenomenologie, hermeneutiek of kritische psychologie (zie Smaling, 1987, pp. 166-172). Ze steunen op een houding of bron voor objectiviteit: sceptische, ervaring gebonden, dialogische, hermeneutische objectiviteit en gebruik van eigen waarden en normen om het object van onderzoek recht te doen.

Er is voor de diagnostiek een minder nadrukkelijke stroming met een uitloper naar objectiviteit als het object van studie recht doen. Dat is de interpretatieve waar hermeneutiek, constructionisme en het gebruik van metaforen onder vallen. Diagnostiek en psychologie ontdoen zich niet geheel van hermeneutiek. Hermeneutisch te werk gaan berust op de *bereidheid* - er is geen dwingend argument - eenzelfde betekenis te leggen onder voor de waarneming verschillend gedrag. Een nog minder aanwezige en in de psychologie gewantrouwde stroming is de kritische. Deze is met het historisch en dialectisch materialisme verbonden is. Marx is nooit helemaal van het toneel als het om objectiviteit gaat. De drie stromingen hebben achtereenvolgens technische beheersing (voorspellen en controleren), communicatie (het eens worden over een betekenis, bijvoorbeeld bij het begrip van het ontwikkelingsstadium) en emancipatie (de Marxistische arbeider) als *kennisbelangen*.

Subjectief en objectief is in de diagnostiek een *muddy dichotomy*. Objectief wordt soms opgevat als overeenstemming tussen iedereen of velen, of als de unanieme mening van rationele discussianten: experts. Subjectief wordt beschouwd als het tegenovergestelde. Objectief betekent ook betrouwbaar en valide en subjectief is dan onbetrouwbaar en niet valide (bijvoorbeeld het klinisch oordeel volgens Dawes). Weer een andere betekenis duikt op als men het opvat als een houding, stijl of vaardigheid en dan is het een gewaardeerd kenmerk, een aantal deugden zoals terughoudendheid, openheid, authenticiteit en betrokkenheid.

Vruchtbaar operationaliseren gaat gepaard met betrokkenheid bij het object van studie. Objectiviteit wil ook zeggen het vermijden van een tunnelvisie duiden en het streven op zijn minst meer gezichtspunten betrekken bij het object van studie. We hebben geen *régard survolante* dus ontkomen we niet aan een tunnelvisie. De drie driedelingen van hoofdstuk 1: elementen, bronnen en theoretische oriëntaties zijn bedoeld om vanuit verschillende, niet-hiërarchische gezichtspunten naar de vraag van de cliënt te kijken in een poging een *rich problem field* te maken.

Objectiviteit is op te vatten als balans tussen betrokkenheid en afstand, partijdigheid en onpartijdigheid. Ten slotte kan men objectiviteit tot stand zien komen in een dialoog waarbij de deelnemers gelijkberechtigd zijn en eenieder waarachtig en authentiek is. Kom daar maar eens om op een vergadering of bij het beoordelen artikelen, toekennen van onderzoek subsidies. Bij Hermans' zelfconfrontatie methode (ZKM) heeft diagnostiek dat dialogische karakter. Kouwer heeft

het gesprek centraal gesteld als werkwijze om de cliënt als persoon te beschrijven. Een object praat niet terug, een subject wel.
Objectbetrokkenheid, het object recht doen vraagt dat de diagnosticus open staat voor al die betekenissen. Hij laat zich niet vangen in een van de stromingen. Objectiviteit heeft geen absolute, zekere of onfeilbare basis en kan niet uitsluitend via een weg, bijvoorbeeld de logisch positivistische tot stand gebracht worden. Objectiviteit en waarheid hebben vele vaders, stromingen en draagvlakken nodig.

Begrip van validiteit wordt niet alleen verworven wordt door erover te lezen en tentamens in af te leggen. We leren erover in het alledaagse leven en het is een verworvenheid van de cognitieve en sociale ontwikkeling.

Alledaagse concepten van waarheid Waarheid is verbonden met redelijkheid, zuiverheid, deugdelijkheid, gerechtvaardigdheid en houdbaarheid van beweringen over de fysische en sociale werkelijkheid en over gedrag van personen. Validiteit en waarheid verwijzen naar wat, hoe en wie objecten en personen werkelijk zijn. We verdisconteren ook haar formele karakter, want het gaat ook om correct beschrijven en logisch redeneren en om tonen dat er over een stand van zaken (geen) twijfel is, overeenstemming, harmonie tussen beweringen over objecten en personen en wat en wie deze zijn. Er worden vier basisintuïties vermeld over kennisverwerving: intuïtionisme, empirisme, rationalisme en pragmatisme. In impliciete opvattingen van validiteit zijn daar elementen van terug te vinden.

Validiteit en waarheid laten ons niet onverschillig We streven we er naar objectief te zijn, dat wil zeggen het *object recht* te doen zoals het is, erover te spreken, zoals het/hij in elkaar zit. Het is geen subjectieve en door eigenbelang ingegeven beschrijving van objecten, verschijnselen en gedrag. We noemen uitspraken over personen en verschijnselen bijvoorbeeld ongeldig als die voortkomen uit het op eigen voordeel uit zijn. We spreiden geen eeuwige achterdocht ten toon. We verdragen de gedachte niet dat we voortdurend bedrogen en voor de gek gehouden worden. Het is geen abstract kenprobleem want we ervaren dat intellectuele inspanningen bedreigd worden door eigenbelang, halve waarheden en leugens. De kenhouding dat er geen waarheid bestaat en dat alles subjectief is, is niet vol te houden. We pikken het niet dat machtigen in politiek en wetenschap de waarheid definiëren zoals het hen uitkomt (Williams, 2002). Er zijn maar enkele radicale cynici voor wie geen enkele inhoud telt. Een voorbeeld:

Vasili Grossman - de Russische schrijver en oorlogscorrespondent (WO II) – schreef in februari 1962 een brief aan de Russische partijleider Chroesjtsjov om zijn boek *Leven en Lot* (nieuwe Nederlandse vertaling: 2014, pp. 953 - 957) gepubliceerd te krijgen. Hij beschreef literair en precies wat hij gezien had bij het oprukkende Sovjet leger in WO II tot en met de bevrijding van twee Nazi concentratiekampen. Uit zijn brief: '... *mijn fysieke vrijheid is zinloos als het boek waaraan ik mijn*

leven heb gegeven zich in gevangenschap bevindt. Ik heb het ten slotte geschreven; ik heb er geen afstand van genomen, en dat zal ik ook nooit doen. Het is nu twaalf jaar geleden dat ik aan dit boek begon te werken. Ik geloof nog steeds dat ik de waarheid heb geschreven, uit liefde en medelijden, omdat ik in mensen geloof. Ik verzoek u mijn boek de vrijheid te geven'. De geheime dienst probeerde het manuscript te vernietigen, maar het is uiteindelijk vertaald en in het Westen verspreid. Op zijn brief kreeg hij nooit antwoord.

De cynische houding leidt ertoe dat sociologie van de kennis die zegt dat waarheid eigendom is van de groep met de macht de validiteit van gedrag schept. Zo streven naar macht is volgens Williams (2002):

'... *one of the reasons, why, at the present time, the study of humanities runs a risk of sliding from professional seriousness, through professionalization to a finally disenchanted* (betovering wegnemend, ontluisterend, ontgoochelend) *careerism'.*

Bernard Williams was moraalfilosoof. Zijn laatste boek gaat over waarheid en moraal (2002). Hij is bekend vanwege zijn ironische kritiek op het utilitarisme, dat volgens hem uitgaat van een onpartijdige waarnemer, die alwetend, belangeloos en zonder emoties is, maar verder 'normaal' (1985) is. Mensen kunnen volgens hem niet 'gezichtspuntloos' zijn.

Bellettrie is meer nog dan de wetenschap een bron voor het denken over waarheid. Voorbeelden:

(1) Kingsley Amis van *Lucky Jim* (1953) was een universitair literatuurdocent. Hij beschrijft de slangenkuil die de werkkring van zijn hoofdpersoon Dixon is. Deze haat zijn professor, is zich bewust van de trivialiteit van diens onderzoek, maar hij voegt zich naar de Britse universitaire mores, die ieder laat vechten voor zijn carrière. Hij is bereid daarvoor een liefdeloze verhouding aan te gaan.
(2) De Amerikaanse literatuurprofessor John Williams schreef de roman *Stoner* over het universitaire leven. Volgens een *reviewer* sloot het boek aan bij zijn eigen loopbaan. De roman (1^e druk, 1965) gaat over de niet glanzende loopbaan van een universiteitsmedewerker. Hij verzorgt zijn onderwijs met passie, heeft de obligate affaire met een PhD studente, en wordt er door een medestudent, die het tot decaan schopte, uitgewerkt. Zijn taak wordt opgeheven. John Williams held neemt inhoudelijk werk serieus. Stoner zag de romans en gedichten die hij besprak niet als iets om tentamen over te doen: ze moesten de studenten iets leren en door hen ervaren worden.

(3) De zes boeken van Voskuil over het Meertens instituut zijn een Nederlands voorbeeld van de - in de ogen van zijn hoofdpersoon Maarten Koning - venijnige sociale structuur van een wetenschappelijk instituut en het triviale onderzoek dat daar verricht wordt.

Bernard Williams beweert dat waarheid een *waarde* is. Naast *accuracy*, de objectieve, valide beschrijving, onderscheidt hij *sincerity*: oprechtheid. Dat houdt onder meer in het afwijzen van *free riders*, anderen niet willen bejegenen vanuit eigenbelang en in staat zijn tot geven en ontvangen. Deze begrippen impliceren weerstand tegen *wishful thinking*, zelfbedrog en fantasie. Dit lijkt op Royce's *authoritarianism.* We zijn afhankelijk van anderen om de betekenis van gebeurtenissen, verschijnselen en menselijk gedrag te begrijpen en te verklaren. Die ander heeft iets te bieden als hij *accurate* en *sincere* is en niet alleen reclame maakt voor zijn gedachten en producten.
Ontkenning van de mogelijkheid tot waarheid en pure onoprechtheid doen iemand belanden in een vreemde wereld. Burgers, werknemers, studenten gaan er vanuit dat de overheid, leraren en bestuurders waarachtig zijn en niet liegen en uit eigenbelang handelen. Daar kan men zich in vergissen zoals ondermeer blijkt uit de gaswinning in Groningen. Williams argumenteert dat we de alledaagse waarheid met respect tegemoet moeten treden. We kunnen daar echter niet stoppen. Feiten dienen *detached* gezocht en beschreven te worden. Ze staan bovendien steeds open voor nieuwe interpretaties.
Waarheid, geldigheid, accuraatheid en oprechtheid spelen een rol in het alledaagse leven. Enkelingen houden een totaal scepticisme vol. Anderen zoeken het in een strikte ideologie, die geen herinterpretatie toestaat en geen tegenspraak duldt. Dit duidt erop dat waarheid en geldigheid gefabriceerd kunnen worden en de jeugdige en volwassen intuïtie dat 'iets niet klopt', bedolven kan raken onder een zee van culturele regels en conventies van wetenschappelijke dwingende theorieën, ideologieën, religies en sekten.

Epistemologische ontwikkeling Voor een uitwerking van het begrip validiteit kunnen we te rade gaan bij de manier waarop kinderen, studenten en professionals kennis verkrijgen over de waarheid en geldigheid van beweringen. Epistemologische ontwikkeling wordt door psychologen bestudeerd. Hoe verwerven we kennis over objecten en mensen en hoe weten we of die kennis waar, geldig is? Kinderen moeten geïnformeerd worden door anderen. Ze denken oprecht de waarheid te vertellen over dingen die ze nooit gezien en ervaren hebben. Ze moeten op anderen afgaan maar zijn geen passieve luisteraars die alles slikken. Voorbeelden:

(1) Harris (2007) vond dat vierjarigen bij voorkeur vragen stelden aan een informant die volgens hen accuraat was. Ze gebruikten aanwijzingen om zich ervan te vergewissen of hij betrouwbaar en waarachtig was. Ze schatten zijn geloofwaardigheid in en hielden rekening met wat de informant over zichzelf zei, bijvoorbeeld dat hij bekwaam en eerlijk was. Fernbach et al. (2012) lieten zien dat kinderen vanaf ongeveer vier jaar hun causale beweringen aanpasten aan de kennis die ze hadden over de oorzaak.

(2) Als men kinderen tussen de acht en twaalf jaar ondervraagt over wat liegen en de waarheid spreken inhoudt en hoe ze weten of een bewering over een persoon of gebeurtenis waar is, dan geven ze antwoorden die we in uitgewerkte vorm in onze tekstboeken aantreffen. Het is waar 'omdat het logisch is', omdat 'het echt gebeurd is', 'omdat je het kunt zien'. Dit is de correspondentie tussen een bewering en een stand van zaken in de wereld. Ze zeggen bijvoorbeeld 'omdat je weet dat het waar is'. Ze voerden ook aan: 'iemand heeft me het eerder verteld', en 'het is hetzelfde verhaal'. Dit lijkt op het hanteren van het coherentiecriterium. Ze kunnen ook tot de bewering: 'het is waar' komen, omdat 'een ander het heeft gezien', 'anderen vertelden hetzelfde'. Dit lijkt op empirische steun en interbeoordelaarsovereenstemming. Er waren ook pragmatici: 'het helpt je als je het weet', 'dan weet ik wat ik moet doen'. Er waren zelfs enkele jonge sceptici: ''dat weet je nooit' en intuïtionisten: 'ik weet het gewoon'; 'daarom'. De kinderen gebruikten ook kenmerken van de boodschapper als criterium voor validiteit: 'hij trekt zich nergens iets van aan, hij is zorgeloos, hij vertelt het zonder blozen, hij krijgt een rood gezicht, hij is serieus, hij vertelt het zonder te stotteren'. De laatste antwoorden kwamen meer voor bij de 8- dan bij de 12-jarigen (Van Houdt, 1994).
(3) Tenney et al. (2011) onderzochten aanwijzingen die kinderen en volwassenen gebruiken om een oordeel te geven over de vraag of iemand een betrouwbare bron van informatie is. De aanwijzingen waren accuraatheid, vertrouwen en kalibratie dat wil zeggen hoezeer het vertrouwen van de informant de kans voorspelt dat het waar is wat hij beweert. Het verschil tussen de 5-6 jarigen en volwassenen bestond vooral uit het in rekening brengen van de kalibratie van de informant.
(4) William Perry Jr. (1913-1999) volgde de epistemologische en ethische ontwikkeling van een groep Harvard studenten. Dit inspireerde Kitchener en King (1994, 2002) tot onderzoek naar de ontwikkeling van reflectief denken bij studenten en volwassenen.

Box Perry en de epistemologische ontwikkeling bij zijn studenten

Perry was studentenpsycholoog op Harvard en bestudeerde epistemologische en ethische ontwikkeling. De twee gebieden sluiten aan bij Kant: 'Kritik der reinen Vernunft' en *'Kritik der praktischen Vernunft'*: hoe weet ik dat wat ik weet valide/waar is. Hoe weet ik wat ik moet doen? Perry wilde niet over stadia van ontwikkeling spreken. Dat is nu en was toen niet populair. Hij noemde het *posities.* Het was niet alleen interne groei of ontwikkeling maar ook een antwoord op verwerken van conflictsituaties. Hij volgde 140 studenten gedurende een aantal jaren door middel van interviews. In zijn tijd vlak na WO II zo zegt hij was onderzoek gericht op individuele verschillen in de autoritaire persoonlijkheid. Hij koos er voor om de intellectuele reis *(intellectual journey*) van zijn studenten te observeren. Hij sprak over een *Intellectual Pilgrim's Progress* en gebruikte met enig gemak Bijbelse metaforen. Er is nu kennis van goed en kwaad: de kennis van waarden en de mogelijkheid tot een geldig oordeel in *'...a world devoid of Eden'*. Het zag een ontwikkeling van absolute waarheid naar een houding gekenmerkt door respect voor de context en *commitment* (Moore, 2003). De onderzoeker blijft niet buiten schot. Hij zet door zijn onderzoek zijn reputatie, kunde en waardigheid op het spel. In de materialen van de studenten onderscheidde hij negen posities, gecomprimeerd tot vier categorieën. De eerste werd vooral gevonden bij de 1e jaars- en de laatste bij PhD-studenten: Dualisme (posities 1-2): De eerste positie kwam zelden bij studenten voor. Het is de tuin van Eden (het Paradijs): een ongereflecteerd zicht op Absolute Waarheid. Ouders zijn de autoriteit; een ander gezichtspunt wordt niet getolereerd. In de tweede positie is er sprake van

verschillende perspectieven maar het blijft eenvoudige tweedeling: het is óf waar óf vals.
Meervoudigheid *(multiplist*: posities 3-4): Het bestaan van onzekerheid wordt erkend en de posities zijn: waar-vals-nog niet bekend. Het is het afscheid van de absolute positie. In positie 4 geldt: *anything goes*: niets kan worden beslist: 'je ziet maar wat je doet'.
Contextueel relativisme (positie 5): Verschijnselen en gedrag zijn afhankelijk van de context. Het is geen pseudo-relativisme, want de student is een actieve relativist. Bij alles wat er gezegd wordt, zegt hij: 'dat vind jij dus, ik hoef dat niet te vinden'.
Commitment (betrokkenheid) met relativisme (posities 6-9): Er wordt gekozen uit aannemelijke alternatieven. Er is echte twijfel maar ook een actieve en verantwoorde keuze. De student verovert een plaats in een wereld waarin meer perspectieven mogelijk zijn en gebruikt voor zijn standpunt logische en methodologische regels die aanvaard zijn in de wetenschappelijke gemeenschap.

Perry's werk leidde tot de stadia van King en Kitchener, tot instrumenten om de stadia te meten en tot voorstellen om intellectuele groei te bevorderen (West, 2004; Dawson, 2004; Marra & Palmer, 2004). King en Kitchener ondervroegen 15 - 46 jarigen om uit te leggen hoe ze wisten of beweringen waar of vals waren: *'Can you ever know for sure that your position on this issue is correct?'*. Ze legden de deelnemers twee perspectieven vroegen hen om de gezichtspunten te integreren. Dit materiaal werden gereconstrueerd als zeven stadia in reflectief denken.

Box Stadia van reflectief denken volgens King & Kitchener

Pre-reflectief denken: Stadium 1: Kennis is absoluut en ligt daar objectief, buiten mij, voor het oprapen. Kennis wordt niet begrepen als abstractie. Ze kan dan ook verworven worden door te observeren. Opvattingen behoeven geen rechtvaardiging omdat er een correspondentie is tussen opvatting en werkelijkheid: 'Ik weet het; ik heb het gezien'. Stadium 2: Kennis is weliswaar absoluut zeker maar niet onmiddellijk zichtbaar. Je moet er naar zoeken of luisteren naar iemand die het weet: de autoriteit. Er is een antwoord, geen conflict of debat: 'Het is waar, het was op TV, op het journaal'. Stadium 3: Kennis is nog absoluut en alleen voor een tijdje onzeker. Er is tijdelijk een opinie die verdwijnt als we meer weten en dan we het weer zeker. Opvattingen worden ontleend aan autoriteiten en we kunnen er zelf opinies op na houden. Er is geen directe verbinding tussen evidentie en opvattingen: 'Evidentie leidt tot zekere kennis, zo niet, dan is het een gok'.
Quasi-reflectief denken: Stadium 4: Kennis is onzeker vanwege individueel verschillende meningen en het verschil tussen situaties (incorrect rapporteren, gegevens raken zoek met verloop van tijd, ongelijke toegang tot informatie); kennis kan altijd betwijfeld worden en is ambigue. Opvattingen berusten op redeneren en evidentie, maar argumenten zijn idiosyncratisch en passen bij een vooringenomen standpunt: 'Ik wil wel in de evolutie geloven, als ik een onomstotelijk bewijs krijg, maar dat lukt niet; ik denk dat we het nooit zullen weten'. Stadium 5: Kennis hangt van de context af en is subjectief. Ze hangt af van bril die de persoon op heeft en van zijn eigen criteria. We kennen alleen interpretaties van de feiten, gebeurtenissen en gedrag. Opvattingen zijn altijd context specifiek of een compromis tussen interpretaties. We moeten ons oordeel uitstellen: 'Mensen denken verschillend en pakken problemen verschillend aan; alle theorieën zijn waar, elk heeft zijn eigen evidentie'.

Reflectief en methodologisch denken*:* Stadium 6: Kennis bestaat uit individueel verschillende conclusies want problemen zijn slecht gestructureerd en steunen op verschillende bronnen van informatie. Interpretaties worden geaccepteerd als er evidentie geleverd wordt. Criteria worden gebruikt waarbij niet elke even veel bijdraagt. Het nut is belangrijk. Opvattingen blijven overeind als er een vergelijk tussen opinie en evidentie plaatsvindt: 'Het is moeilijk zekerheid te krijgen, er zijn gradaties en je kunt op een punt komen dat de evidentie voldoende is'. Stadium 7: Kennis is de uitkomst van een proces van *reasonable inquiry* waarbij oplossingen voor slecht gestructureerde problemen worden gezocht. De kwaliteit van de oplossing wordt geëvalueerd in het licht van bestaande evidentie. Opvattingen zijn probabilistisch en berusten op verschillende overwegingen: het geschatte gewicht van de evidentie, verklarende kracht, kans op foute conclusies, alternatieve interpretaties. Conclusies worden niet opgevoerd als definitief en onomstotelijk maar als de meest waarschijnlijke. Elk argument wordt van alle kanten bekeken: manier van redeneren, welke empirische steun en consistentie.

Als *scientist-practitioner* geeft de diagnosticus de voorkeur aan positie 7. Maar lukt dat altijd? Als hij zeker is op basis van intuïtie of ervaring zit hij in de buurt van een stadium 3 argument. Als hij op zijn creativiteit en fantasie steunt benut hij een stadium 4 argument. Het 5de tot 7de stadium wordt onderwezen in het methodologie onderwijs.

(5) Hofer et al. (2002) bestudeerden de epistemologische opvattingen van voornamelijk vrouwelijke studenten in het hoger onderwijs. Ze stelden vragen als: *Imagine two people dispute about the interpretation of a poem'. How can you decide who is right*? '*How can you be sure that your view is right?'* De auteurs reconstrueerden de lijn in het materiaal met behulp van drie stadia:

I Realistische dualistische positie: er is één juist antwoord op elke vraag (naïef realisme).
II Subjectivistische relativistische positie ook *multiplist* positie genoemd: er zijn veel gelijkelijk valide antwoorden op elke vraag; wat waar is voor mij hoeft dat niet voor jou te zijn, maar we hebben beiden gelijk.
III Coördinatie van subjectieve theorieën met de werkelijkheid: ideeën worden met objectieve feiten verbonden.

We hechten belang aan de waarheid, al bestaat dé waarheid niet. We willen niet bedrogen worden. Er zijn impliciete ideeën zijn over validiteit en waarheid. Boven is de vergelijkbaarheid van leken en professionals naar voren gebracht. Merleau-Ponty zou daar mee instemmen want elk gesofisticeerd idee over validiteit heeft zijn basis in de geleefde ervaring van ieder van ons. Dit staat op gespannen voet met het werk van Kahneman en Dawes. Zij benadrukken vertekeningen en het irrationele in oordelen en diagnoses. Merleau-Ponty stelde voor om zulke nog niet bewezen opvattingen op te nemen in een verbrede rede. Hij wil de (vage) freudiaanse concepten over het irrationele, onbewuste opnemen in die verbrede rede.

Er is gelijkenis maar ook verschil. Filosofen zijn getraind in een formele discipline en leken niet. Starmans en Friedman (2012) wijzen erop dat leken beamen dat er ware kennis is,

terwijl filosofen erop wijzen dat er steeds tegenvoorbeelden gegeven kunnen worden. Leken onderscheiden 'toevallig' minder van 'gedetermineerd' dan vakfilosofen.
De diagnosticus is gebaat bij kennis over wat de cliënt waar, geldig of pragmatisch behulpzaam vindt. De cliënt verwacht van hem dat hij accuraat en oprecht is. Er zijn ontwikkelings- en individuele verschillen in wat cliënten geldig en waar vinden. Daar kan hij in de mondelinge en schriftelijke rapportage rekening mee houden. Dit betekent niet dat de diagnosticus de cliënt onderwijs in epistemologie moet geven maar ook weer niet dat hij zijn antwoorden en oplossingen moet aanpassen aan wat een cliënt wil horen. Het is niet eenvoudig zo'n balans te vinden.

Samenvatting en conclusie

In alledaagse taal verwijst validiteit naar redelijkheid, robuustheid en houdbaarheid van beweringen over de fysische en sociale werkelijkheid. Het zijn niet alleen abstracte formele, logisch criteria. Ze weerspiegelen ook waarden. We veronderstellen dat we beweringen doen die waar zijn en dat de boodschappers die ons informeren - wetenschappers, politici, programmamakers en leraren oprecht zijn. We gaan er niet van uit dat elke bewering over objecten en personen subjectief is en het belang van een persoon of groep dient.
Impliciete opvattingen over waarheid/geldigheid zijn door ontwikkelingspsychologen onderzocht. Kinderen vanaf vier jaar proberen al een profiel te maken van informanten, schatten geloofwaardigheid in en houden rekening met hun accuraatheid, onwetendheid en onzekerheid. Kinderen tussen acht en twaalf jaar geven op beperkte wijze blijk van een idee over coherentie, overeenstemming en pragmatiek. Ze gebruiken kenmerken van de boodschapper om te bepalen of ze de waarheid spreken. De stadia in reflectief denken bij adolescenten en volwassenen laten een ontwikkeling van naïef realisme naar waarheid als een uitkomst van theoretiseren en toetsen zien. Zekerheid wordt ingeruild voor waarschijnlijkheid. Dit is niet zozeer een filosofisch gefundeerde houding als wel een gevolg van ervaring en onderwijs: *wiser and sadder*.

6. Expliciete opvattingen over validiteit

Validiteit kan toegepast worden op theorie, methodologie en instrumentatie. Validiteitstheorie verwijst vooral naar validiteit van tests. Ze is methodologisch van aard en in de empirisch-analytische en logisch positivistische stroming uitgewerkt. Toch onttrekt ook validiteit zich niet geheel aan de hermeneutiek. Statistische technieken zijn verbonden met soorten testvaliditeit: lineaire regressie voor predictor-criterium relaties en factoranalyse voor constructvaliditeit. Validiteit van diagnostische procedures is nauwelijks onderzocht. Een nieuwe procedure is evenals bij therapieën snel gelanceerd, maar vergelijkend onderzoek om uit te maken of ze iets toevoegen ontbreekt vaak. Quasi experimentele ontwerpen zijn een thema geworden omdat willekeurige toewijzing aan condities niet altijd mogelijk is. Omdat de validiteit van deze ontwerpen niet zo duidelijk is als bij ware

experimenten zijn er vier validiteitssoorten onderscheiden en bij elk zijn bedreigingen van de soorten uitgewerkt.
Eerst wordt de geschiedenis van het validiteitsconcept beschreven. Ten tweede komt het overheersende doel van voorspellen (criteriumvaliditeit) aan de orde. Ten derde wordt ingegaan op het doel van controle. Dat is verbonden met validiteit in (quasi) experimenteel onderzoek. Ten vierde kun je je afvragen of validiteitscoëficiënten gegeneraliseerd kunnen over ongeveer gelijke predictoren en criteria (*validity generalization*) en of het toevoegen van predictoren en covariaten validiteit verhoogt (*incremental validity*). Ten vijfde wordt Cohens vuistregel over de vereiste hoogte van predictieve validiteitscoëfficiënten gebruikt om te zien hoe deze uitpakt in onderzoek. Ik stel me de vraag gesteld of er zo'n regel is voor constructvaliditeit.

Geschiedenis: van predictieve, inhouds- naar constructvaliditeit De belangrijkste kwestie van de diagnostiek is validiteit (Lissitz, 2009). In 1921 rapporteerde Courtis de resultaten van een overleg van een standaardisatiecommissie. De boodschap was dat een test valide is als die meet wat die verondersteld wordt te meten. Een test verwijst dus naar iets dat bestaat en dat kan gemeten worden. Als men niet zeker is over het attribuut, de inhoud, de dimensie, de latente trek en ook niet over het meten daarvan omdat een specifieke structuur ontbreekt dan is er een probleem. In de loop van de tijd is de nadruk verschoven. In het begin ging het om basisprocessen (geheugen, reactietijd) om werknemers te selecteren bijvoorbeeld buschauffeurs en telefoonoperators. Thorndike (1918) noemde een test valide voor elk criterium waarvoor het goede schattingen levert. Zo is begonnen bij het pragmatisch doel van het voorspellen van prestaties. In het toepassen gaat het niet om de geldigheid van een model maar om de geschiktheid voor een bepaald doel (Sireci, 2009).
Predictieve of criteriumvaliditeit is historisch de eerste, of de tweede, als met het prille theoretische begin erbij neemt. Als het criterium tegelijk met de voorspellers gemeten wordt heet het *concurrent validity*. In zijn boek definieerde Gulliksen (1950) het als de kwaliteit van een test om toekomstig gedrag efficiënt te voorspellen. Testonderzoek voor het leger betrof vooral selectie en daarvoor zijn predictief valide instrumenten nodig. Al gauw volgt het testen voor plaatsing in het onderwijs en om te bepalen wat en hoe goed iets geleerd is. De laatste zijn verbonden met de inhoud van curricula voor lezen, rekenen, wereldoriëntatie, enzovoort. Dit is *content validity* en verwijst naar de gelijke kans van elk item uit een omschreven populatie items om in de test opgenomen te worden. Dit is goed te doen voor helder te omschrijven inhouden, bijvoorbeeld alle optelsommen met getallen < 20. Het gaat nog bij prestatietoetsen maar voor onderwijsdoelstellingen is dat al lastiger. De hulp van experts wordt ingeroepen om een domein te omschrijven en items te maken. Dit gebeurt bij de Cito-toetsen.

Construct validity een wolf in schaapskleren? Jonson en Plake (1998) maken aannemelijk dat de invoering van *construct validity* een ingrijpende verandering is geweest. Het begon

met een complexe theoretische bijdrage van MacCorqoudale en Meehl (1948). Zij bespraken hypothetische constructen. Dat zijn hypothesen over het bestaan van eenheden, processen en mechanismen die niet direct waar te nemen zijn. Constructen verwijzen naar onderliggende eenheden, enzovoort. Dit is in 1955 uitgewerkt door Cronbach en Meehl. Ze hadden het concept nodig om recht te doen aan onderzoek dat weliswaar inzicht gaf in testgedrag maar dat niet gedekt werd door predictieve en inhoudsvaliditeit. Ze beschrijven methoden en inferentieregels waarmee men evidentie kan verkrijgen over de hypothetische constructen. Ze leggen de nadruk op inductieve én deductieve processen. Correlationele en experimentele onderzoeksopzetten zijn bruikbaar. Het onderdeel Begripsvaliditeit van het Cotan beoordelingssysteem van tests volgt dit essay uit 1955 nog steeds. Cronbach is niet expliciet over de ontologische status van de begrippen. Hij vat ze soms op als werkelijke eenheden, processen en *states* maar ook als *inductive summaries*. Dit opende de deur voor interpretaties van hypothetische constructen. Dat is blijmoedig gebeurd bij Messicks uitwerking van validiteit. Het houdt onder meer in dat een concept niet volledig gedefinieerd is door een specifieke operationalisatie of empirische specificatie. Cronbach en Meehl aanvaarden dat er een verschil is tussen theoretische en observationele taal (Slaney, 2012). Psychologische begrippen zijn *funktionierende Begriffe* die we werkendeweg invullen.

Box Psychologisch constructen: vaag maar het kan bijna of eigenlijk niet anders

Logische analyse en empirisch onderzoek geven een construct al doende, werkendeweg vorm en inhoud. Dat onderzoek kan semantisch theoriegeleid zijn, psychometrisch door gebruik te maken van statistische modellen en exploratief. Een gemeenschappelijk kenmerk is dat constructen het gesprek over interessante, belangrijke en terzakedoende psychologische verschijnselen op gang houden. Dat zijn in de praktijk overwegend gedragingen die succes opleveren, waar we bang voor zijn of die ons medelijden opwekken. Wetenschappelijk is er een voorkeur voor een omschrijving met een materieel en/of formeel object. De materiële objecten van de sub-disciplines - waar de diagnostiek op terugvalt - bieden een *mèr a boire* aan gedragconstructen. Over welke dat zijn, raken we niet uitgepraat. In een speciaal nummer van het tijdschrift *New Ideas in Psychology, 31,* (2013) komen twee betekenissen aan de orde (a) het kan gelezen worden als een proces, een mechanisme en verwijzen naar een hypothetische entiteit die individuele verschillen, ontwikkeling en adaptatie aan de omgeving verklaart en (b) als een heuristische samenvatting van een aantal operaties. Dit lijkt op het onderscheid dat Cronbach en Meehl 60 jaar geleden maakten. Het verschil is niettemin duidelijk: de heuristische versie veronderstelt dat een construct een nuttige samenvatting van *observables* is terwijl het hypothetische entiteiten kamp een onderliggende variabele of genererend mechanisme veronderstelt (Lovasz & Slaney, 2013). Sommigen willen van het hypothetische af. Maraun en Gabriel (2013) stellen dat men genoeg heeft aan de concepten en hun empirische referenties. Dat mag uiteraard maar hier ligt een oude omstreden slagzin op de loer: intelligentie is wat intelligentietests meten. Er is een verzoeningspoging. Markus en Borsboom (2013) willen de inhoud vastleggen in een *behavioral domain theory* en een *causal structure* ontwerpen die de basis is voor het domein waarin individuele verschillen op dimensies beschreven worden. De domeinscore

moet verbonden worden met een bestaand attribuut dat met een meetmodel (Rasch) afgebeeld kan worden. Een dergelijke combinatie doet denken aan Guttmans facet theorie maar de auteurs vermelden deze niet. Dit brengt ons terug bij het debat over de relatie theorie opgevat als een reeks semantische constructen, een *behavioral domain theory* en meten vooral met IRT. Sijtsma (2012) twijfelt of ze goed te verbinden zijn want de constructen (het domein) zijn te vaag om de IRT en andere modellen toe te passen. Trendler (2013) zei dat ze zo wel eeuwig kunnen doorgaan, als ze niet bereid zijn om de pretentie van het vinden van Newtoniaans opgevatte universele gedragswetten op te geven. Dit is een controverse die in allerlei jasjes en op verschillende tijdstippen terugkeert. Misschien zijn ze op te vatten als twee toegangen, perspectieven die elk tot interessante kennis over gedrag kunnen leiden.

Loevinger (1957) beschouwde constructvalidering als een proces van volledige theorietoetsing. Campbell en Fiske (1959) voegen de *multi-trait multi-method* werkwijze aan constructvalidering toe. Deze maakt het mogelijk karaktertrekken te onderscheiden en het aandeel van het gebruik van dezelfde methode (observatie, zelfrapportage in vragenlijsten) vast te stellen. Trekken 'mogen' immers niet correleren vanwege de gebruikte methode. Magnusson (1966) sprak over en *complete validiteitstudie* als we na predictief en inhoudsonderzoek de aanbevelingen van Cronbach en Meehl (1955) en Campbell en Fiske (1959) volgen. *Face validity* ruimde het veld omdat het te veel eer gaf aan wat mensen *denken* dat valide is. Guion (1980) noemt criterium-, inhouds- en constructvaliditeit *something of a holy Trinity* en beschouwt het als een afgerond thema.
Invoering van constructvaliditeit is de grootste verandering geweest voor de *validity theory*. Deze soort staat zo veel interpretaties toe dat het eind zoek lijkt. In 1985 doet zich volgens Jonson en Plake een dramatische verandering voor naar een *unitary* of *unifying concept.* Validiteit beschouwd als een *strategie om testscores te interpreteren*.

Messicks validiteitsconcept: Het *modern unified validity concept* van Messick (1931-1998) had en heeft invloed tot in de voorlopige *Standards for Educational and Psychological Tests* van 2014. Het plaatst interpretatie van testscores in het middelpunt en verwijst naar geschiktheid, nut en zinvolheid van gevolgtrekkingen uit testscores. Hij was verbonden aan de *Educational Testing Service* in Princeton (VS). Het Nederlandse Cito is er naar gemodelleerd. In 1989 schreef hij in een belangrijk handboek over testen een lang hoofdstuk over validiteit (Linn et al., 1989). Hij zei (persoonlijke communicatie, Groningen, september, 1993) dat dit het moeilijkste was dat hij ooit geschreven had. Dit concept is terechtgekomen in de *Standards* van 1999. Ik ging er min of meer vanuit dat het in de voorlopige uitgave van 2014 zou verdwijnen maar dat is niet het geval.
Validiteit is geen kenmerk van een test maar gaat over interpretatie van testscores (p. 13):

Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment'.

Samuel Messick

Hij was een Educational Testing Service (ETS) medewerker die een hoofdstuk over validiteit in een handboek schreef dat het uitgangspunt werd voor validiteit in de *APA Standards* van 1999.

Constructvaliditeit verbindt alle voorgaande soorten en brengt ze tot eenheid. Dat berust op een hermeneutische actie. Het is niet het resultaat van formele analyse of empirisch onderzoek. Er moet een basisidee gevormd worden dat de eenheid tot stand brengt. Het is een open concept want er is geen grens aan interpretaties. Messick voegt in de geest van die tijd *consequentiality* toe. Dat houdt in dat een test die groepen discrimineert niet valide is. Discrimineren wil bijvoorbeeld zeggen dat kandidaten met een gelijkwaardige criteriumprestatie maar met lagere testscores afgewezen worden.

Tabel 1 Messicks validiteitconcept: de integratieve kracht van constructvaliditeit, dat wil zeggen bewijs voor interpretaties en de gevolgen met betrekking tot maatschappelijke waarden.

	Testinterpretatie	*Testgebruik*
empirisch bewijs	Constructvaliditeit	constructrelevantie, nuttigheid, enzovoort
maatschappelijke gevolgen	erbij betrokken waarden	sociale gevolgen

Het voorstel opent de deur voor veel interpretaties en methoden om die interpretaties te ondersteunen.

Tabel 2 Messicks vragen bij het trekken van conclusies uit testscores en *other modes of assessment*. In de rechterkolom staan de klassieke validiteitsoorten: predictief, inhoud en constructvaliditeit.

Vragen	*Aansluiting bij validiteitssoorten*
Is de inhoud evenwichtig vertegenwoordigd in de test, en is het de bedoelde inhoud?	Inhoud
Is niets belangrijks over het hoofd gezien?	Inhoud

Vragen	*Aansluiting bij validiteitssoorten*
Introduceren de manier van testen en de specifieke methode geen irrelevante variantie, waardoor de testscores worden vertekend?	Construct: Campbell en Fiskes 'multi-trait multi-method matrix'
Correspondeert de manier van scoren van de test met de manier waarop de processen in dat domein werken; stemt de structuur van de scores overeen met de structuur van het domein waarover conclusies worden getrokken en voorspellingen worden gedaan?	Construct en Predictief
Welk bewijs is er dat de scores betekenen wat ze pretenderen te betekenen; is bijvoorbeeld de rapportage van persoonskenmerken zinvol bij vragen omtrent personeelsselectie, onderwijsdoelen, therapie?	Construct, Predictief
Zijn er andere, concurrerende interpretaties van de testscores of zijn andere adviezen voor behandeling mogelijk?	Construct, het gaat in tegen de tendens om één-op-één relaties te definiëren en één oorzaak voor een verschijnsel te zoeken: Katzko (2002) noemt deze onjuiste gewoonte: *the uniqueness assumption*
Zijn de scores betrouwbaar en kunnen ze gegeneraliseerd worden over inhouden (domeinen), contexten en groepen?	Generaliseerbaarheidstheorie van Cronbach et al., een samenvatting van de Klassieke Testtheorie (KTT) maar ook inhoud- en constructvaliditeit: betrouwbaarheid en validiteit liggen hierbij in elkaars verlengde
Wordt rekenschap gegeven van de waarde-implicaties van de scores; is er informatie bekend of normen en waarden gerespecteerd worden?	*Consequential*
Is de interpretatie van de testscores zinvol voor school- en beroepskeuze, selectie, behandeling?	Predictief, *consequential* en wat eerder als doel van diagnostiek is toegevoegd: Beslissen
Zijn de scores eerlijk voor verschillende groepen, mannen en vrouwen, subgroepen in de samenleving en verschillen in sociaaleconomische status (SES)?	*Consequential*: discriminatie
Zijn de gevolgen van de resultaten op de	Predictief en een verwijzing naar het

Vragen	*Aansluiting bij validiteitssoorten*
tests in overeenstemming met doelen op korte en lange termijn; zijn er geen nadelige bijeffecten?	criteriumprobleem

Validiteit is vooral testvaliditeit en diagnostiek spiegelt zich daaraan. Het gaat - als ze zich zo en daartoe bepaalt - om voorspellen van een criterium, adequaat representeren van een gedragsdomein en benutten van onderzoek dat niet meteen over criteria of inhoud maar over het construct zelf. Validiteit is gecombineerd en geïntegreerd in een verbreed idee. Dit verloop correspondeert met verandering in theoretische oriëntatie van pragmatisch-empirisch via technisch-psychometrisch naar psychologisch-theoretisch. Het idee dat een kritisch experiment een theorie kan falsifiëren (*justificationism*: Lakatos, 1968) is afwezig. Een theorie wordt niet volledig getoetst door empirische feiten. Messick heeft een *non-justifactionist view* op validiteit. Het onderzoek is nooit af en draagt bij aan een doorgaand proces van theorieontwikkeling en -evaluatie (Strauss & Smith, 2009).

Reacties op Messicks opvatting Integreren is kennelijk een behoefte. We houden van eenheid en eenvoud maar validiteit is een containerbegrip (Slaney & Racine, 2011, p. 8) en die zijn gevuld met verschillende materialen. Het concept kan als een samenvatting van observaties of als een latente variabele opgevat worden. Het geïntegreerde concept drukt een streven naar *'mono-validity'* uit, zoals ook er monotheorisme is. Messicks concept vereist het beantwoorden van veel vragen tegelijkertijd. Dit herinnert aan de richtlijnen bij het diagnostisch proces. Je kunt je afvragen of een testconstructeur zich aan al de aanbevelingen kan houden. Er zijn kritische reacties, voorbeelden:

(1) Sheppard (1993) stelt vast dat de diagnosticus zelf moet uitzoeken welke informatie hij verzamelt gegeven de vraag van zijn cliënt. Zij gaat ervan uit dat diagnostici zelden verder gaan dan direct beschikbare informatie en beperkt rekening houden met de diagnostische waarde van die informatie. Ze zegt dat het toetsen van hypothesen te ver gaat en stelt voor een rangorde te maken van de vereisten en aanbevelingen. Zij neemt aan dat oude concepten blijven en is niet ingenomen met het idee van een nooit eindigend proces: zoveel hypothesen, situaties, groepen, enzovoort. Het moderne geïntegreerde concept van validiteit wil recht doen aan de complexe psychologische realiteit maar het is zaak om maat te houden. Elk construct bevat immers een reductie van de werkelijkheid. De procedures zijn in de diagnostische praktijk niet uit te voeren. Het lot van het *unified concept* lijkt op dat van de generaliseerbaarheidstheorie. Het genuanceerd, elegant en slim bedacht maar moeilijk toe te passen.
In de *Standards* van 1999 is Messicks concept niettemin uitgangspunt voor de omschrijving van validiteit. De nieuwe voorlopige *Standards* (6de editie, 2014) handhaven het concept. De vernieuwing bestaat uit meer aandacht voor de IRT dan in vorige uitgaven. Er is geen paal en perk gesteld aan het brede *unifying* validiteitsconcept. Voor professionals lijkt predictieve validiteit in letterlijke zin de eerste de beste. Het gaat vaak om voorspelling van gedrag van een cliënt. Van de andere zegt Sireci

(2007) *'...that it is extremely difficult to explain to lay audiences'* (p. 478). Men kan men diagnostici en practici toevoegen.
(2) Borsboom et al. (2004) zijn het op theoretische gronden met Messick oneens. Ze houden het bij de opvatting van Courtis dat een test valide is als het een attribuut meet. Maar wat is dat attribuut? Men kan denken aan een factor of een Rasch schaal. Borsboom et al. willen af van het eindeloos interpreteren van de inhoud. Het attribuut bestaat, veroorzaakt individuele verschillen en kan gemeten worden, i.e. is te beschrijven in een model. Als het geen individuele verschillen veroorzaakt dan heeft het geen zin. Er zijn niet zo zeer veel interpretaties als wel meer valideringsprocedures. Deze hoeven geen eenheid te vormen en dat doen ze ook niet.
(3) Een verzoening is na a en b te verwachten. Hoods (2009) beweert dat Messick en Borsboom et al. een essentialistische epistemologie volgen. Messicks bijdrage wordt gereconstrueerd als een methodologische. Hij toont hoe er steun gevonden kan worden voor interpretaties van testscores in een steekproef. Borsboom et al.'s bijdrage gaat over de structuur van latente trekken of theoretische attributen gebonden aan een materieel substraat. Messick is geen constructivist want de interpretaties zijn gebonden aan de empirie. Borsboom et al. zijn realisten want de latente trek bestaat en veroorzaakt individuele verschillen. Hood verbindt de twee. Tests zijn primair instrumenten die berusten op een formeel meetmodel voor attributen (Borsboom et al.) en deze leveren categorieën en schalen die inferenties toestaan (Messick).
Wat heeft de diagnosticus hieraan? Heeft het debat alleen zin op conceptueel-theoretisch en meet-model niveau? Ik vermoed dat ze geen invloed hebben op zijn testgebruik en -interpretatie. De discussie over inferentie versus attribuut raakt de diagnosticus nauwelijks. Ook de *APA Standards* voor testgebruik die Messicks concept uitdragen hebben geringe invloed op diagnostische activiteiten (Jonson & Plake, 1988). Het doel van voorspellen en de behoefte aan inhoudsvalide tests in het onderwijs domineren. Het validiteitsconcept van Borsboom et al. is voor testconstructeurs van belang als het leidt tot betere tests. Feitelijk gebeurt dat ook voor een deel: Cito toetsen bevatten alleen testitems die aan een model voldoen.

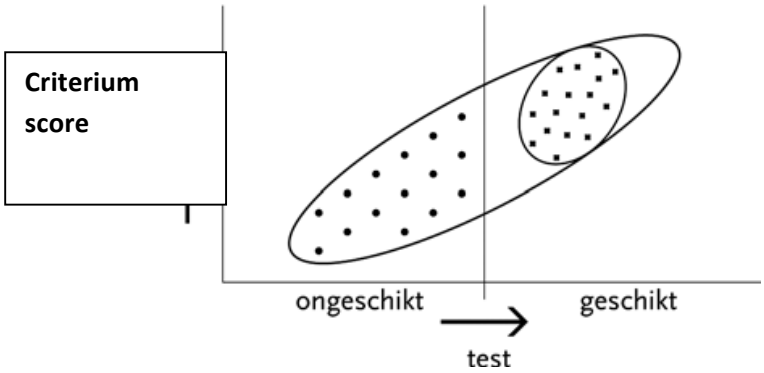
Soorten criteria en restriction of range Drie vragen zijn verbonden met het voorspellen van criteria. Welke soorten zijn er, hoe betrouwbaar zijn criteriummetingen en hoe gevoelig zijn predictor-criterium (p-c) relaties voor de samenstelling van de steekproef?

Soorten criteria: Thorndike onderscheidde in 1949 *specifieke* en *globale criteria*. Een specifieke opdracht die dicht bij de functie en taken ligt is gemakkelijker te scoren dan een globale, bijvoorbeeld voldoen aan criteria als 'een evenwichtig persoon met gevoel voor ambtelijke verhoudingen' of 'kunnen werken in een complexe organisatie'. Een specifiek criterium maakt een grotere kans te correleren met een test dan een globaal. Een *onmiddellijk (immediate) criterium* kan redelijk goed voorspeld worden, bijvoorbeeld welk resultaat je op het eerstvolgend tentamen haalt, gegeven kennis van voorafgaande resultaten. Als er men (weer eens) zou willen selecteren in het hoger onderwijs zou men aankomende studenten kunnen vragen om een tentamen te doen over een representatief eerstejaars boek van de studie van hun keuze. Dit gebeurt bijvoorbeeld bij psychologie in Groningen en de Universiteit van Amsterdam. Een *intermediate criterium* is voorspellen dat 'Daan zijn Masters binnen vijf jaar behaalt'. Dat is al moeilijker te voorspellen. Het

ultimate criterium is relevant maar het moeilijkst te voorspellen; bijvoorbeeld worden Sofie, Marijke, enzovoort goede dokters, ingenieurs, loodgieters, managers, politici?
Criteria en afhankelijke variabelen zijn complexe gedragingen en ze worden niet altijd onderworpen aan psychometrische analyses. Criteriummetingen bestaan ondermeer uit beoordelingen door experts, personeelsfunctionarissen, managers en leken. De betrouwbaarheid van deze metingen is soms niet bekend en regelmatig laag. Het is bekend dat beoordelingen van onderzoeksaanvragen op een aantal criteria tussen de r = .15 en r = .35 correleren. Dat is een matige correlatie volgens Cohen. Dit geldt voor de meeste beoordelingen. De dichteres Ida Gerhardt ergerde zich eraan*: '...Laat mij wél zeggen dat ik om de 'litteraire kritiek' in deze lage landen uiterst weinig geef. Een beurs, waar men handel drijft en wissels vervalst...'* (uit: *Courage!* Brieven bezorgd door B. Hosman en M. Koenen, 2005, p. 442 [598]. Amsterdam: Athenaeum-Polak & van Gennep).

Het is niet eenvoudig betrouwbare en valide criteriummaten te maken. Soms past men de correctie voor attenuatie toe om de p-c relaties te verhogen. Dit betekent dat de p-c correlatie berekend wordt alsof de predictortests perfect betrouwbaar zijn. Neem X als de predictor voor criterium Z en Xrr' en Zrr' als de betrouwbaarheden, dan is de maximale correlatie tussen X en Z: rxz = < √ (vierkantswortel van) rxx' . rzz'. Bijvoorbeeld de maximale correlatie tussen X en Z (rxz) (gegeven rxx' = .81 en rzz' = .16 = .√ 81 x .16 = .90 x .40 = .36. Men mag zich afvragen of zo'n correctie zinvol is. Het is het maximaal haalbare terwijl het gaat om wat je werkelijk voorspelt. *Face validity* wordt hieraan toegevoegd. We nemen aan dat sommige gedragingen goede voorspellers zijn maar er is altijd onderzoek nodig om dat te bevestigen. We kunnen ernaast zitten (zie de DBCs in II).

p-c correlaties zijn gevoelig voor de variantie in de steekproef. In een homogene groep is de p-c-correlatie lager dan in een heterogene groep. Dit verschijnsel heet *restriction of range*.



Figuur 1: *Restriction of range*: de scores van de gehele groep liggen dicht bij de regressielijn (kleinste kwadratensom voor afwijkingen van de regressielijn). De puntenwolk van de scores van de subgroep ligt bijna in een cirkel. Dat levert bij een eventuele regressielijn een grote kwadratensom op.

Restriction of range kan zich voordoen bij strenge selectie. Na training worden kandidaten geselecteerd die de hoogste scores hebben behaald. Hun scores worden gecorreleerd met een criterium bijvoorbeeld het aantal fouten dat ze als straaljagerpiloot, kraanmachinist of

computeroperator maken. De correlatie is laag door de geringe variantie in de predictoren. Als er geen variantie is dan is er ook geen covariantie (correlatie).

Validiteit p<.05, p< .01 en effectgrootte Beschrijving, voorspelling, controle en beslissing. Zijn doelen van diagnostiek. Valide *beschrijvingen* doen - fenomenologisch opgevat - gedragingen recht in hun uniciteit. Beschrijving kan ook verwijzen naar een formeel model dat gepast wordt op gedrag, bijvoorbeeld een Rasch schaal voor een item. De validiteit van *voorspellingen* is uitgewerkt voor rangorde en interval gescoorde tests die criteriumgedrag prediceren. Testvaliditeit wordt aangetroffen als *entry* in tekstboeken voor onderwijs, diagnostiek en psychometrie. *Controle* is het gebied van het experiment. Variatie in de onafhankelijke variabele veroorzaakt de variatie in de afhankelijke. De eerste verklaart doorgaans een bescheiden deel van de variantie in de afhankelijke variabele. We zijn in empirisch onderzoek tevreden met een significant hoofd- en interactie-effect na toetsing van de nulhypothese dat de onafhankelijke variabele niets uithaalt. Formeel gezegd: de gemiddelden van experimentele en controle condities komen uit dezelfde populatie.
Validiteit in 'true' experimenten is geregeld door binnenvariantie ten opzichte van tussenvariantie te toetsen. Validiteit van experimenten wordt geborgd geacht door de bepaling van de fout van de eerste soort, de alfa: $p < 0.5$ of $< .01$. Deze waarden worden vermeld ook al zijn er andere mogelijkheden (Simmons et al., 2012, zie hoofdstuk 1). De alfa schat de kans dat de nulhypothese ten onrechte verworpen wordt, of het vinden van een significant verschil dat er niet is. Een significant verschil tussen de experimentele en controle groep wordt gelezen als: de manipulatie van de onafhankelijke is de oorzaak van verandering in de afhankelijke variabele. Er is nadruk op de fout alfa. Waarom de verwaarlozing van de fout van de tweede soort: bèta: het ten onrechte concluderen dat de onafhankelijke variabele de verandering niet veroorzaakt of het missen van een verschil dat wel bestaat?
De reden is dat de alfa fout past bij kwaliteitscontrole van producten. Om risico's te vermijden wordt de kans op een fout van de eerste soort laag gezet (Fisher, 1955). Daar kun je je iets bij voorstellen. Een auto moet het 'altijd' goed doen, bijvoorbeeld 5 jaar garantie. Misschien is dit voor psychologisch onderzoek anders.
De kans op een bèta fout is niet onafhankelijk van de alfa: de kans op een fout van de tweede soort is groter als de steekproef klein is, de alfa laag gezet is en er een eenzijdige toets wordt gebruikt.

Kritiek op NHST is oud Er is al gedurende 90 jaar kritiek op de nulhypothese en significantietoetsing. Lambdin (2012) vat de historie van twijfel en kritiek samen aan de hand van 29 eerste auteurs van Boring (1919) tot Lindsay (2008). Vanaf 1930 zeiden methodologen dat significantietoetsing niet zo belangrijk is in onderzoek: p-waarden zijn een *dirty little secret* van de psychologie (p. 67). Ze waarschuwen tegen de toetsing, omdat het je wegtrekt van de gegevens zelf. Het sust in slaap in plaats van dat het helpt. Het kan

tot absurde conclusies leiden: r = .03, maar significant door het omvangrijke sample. Zo kun je significante verschillen in IQ tussen mannen en vrouwen aantreffen bij een zeer gering verschil in aantal punten. Als de p-waarde significant is en de betrouwbaarheidsinterval groot dan kan de p-waarde misleidend zijn. Daar komt bij dat in experimenteel onderzoek een variabele geïsoleerd wordt. In de werkelijkheid is komt een variabele niet geïsoleerd voor. Als het aantal meeteenheden groot is dan kunnen er verwaarloosbare (*negligible:* Cohen) correlaties significant zijn. Cohen (1992, p. 84) zegt dat *eyeballing the data* zinvol is. Dit gaat in tegen wat geleerd is: vertrouw je eigen ogen en oordeel *niet,* baseer je op statistische toetsen. Toch hebben we ogen én toetsen nodig. Het durven beslissen wanneer precies kenmerkt de ervaren diagnosticus en onderzoeker.
Schmidt (1996) probeerde de gevolgen van het inzicht in de beperkingen van nulhypothesetoetsing voor het *onderwijs* uit te leggen. Hij daagde docenten uit om hun cursussen te veranderen zodat studenten begrijpen dat de nadruk op de H_o toetsing de opbouw van cumulatieve kennis vertraagt. Hij wil studenten laten inzien dat de voordelen van de toetsing beperkt zijn, dat significantietoetsing vervangen moet worden door puntschattingen en betrouwbaarheidsintervallen en meta-studies zinvol zijn.
Nulhypothese toetsing en p-waarden kunnen daarnaast verkeerd geïnterpreteerd worden. Lambdin (2012, p. 73) vermeldt twaalf van zulke fouten, bijvoorbeeld, de p waarde zegt in welke mate de gegevens op basis van kans tot stand gekomen zijn en de nulhypothese is
'...*the probability of obtaining the results in hand, assuming that the statistical null hypothesis is true in the population*' (p. 74).

Inferenties moeten niet automatisch plaatsvinden door kijken naar de p waarde. Deze argumenten mogen verstandig klinken, maar p-waarden worden geëist door tijdschriftredacties. Dit wordt duidelijk in een studie van Fritz et al. (2012) waarin 30 samenvattingen met meer dan 6000 artikelen zijn geanalyseerd. Steeds worden p waarden vermeld, in 38% van de gevallen aangevuld met effectgroottes. In één op de tien artikelen wordt een betrouwbaarheidsinterval vermeld en in 3% wordt de statistische *power* vermeld. In onderzoeksartikelen blijven p waarden, de fout alfa en significantietoetsing domineren.

Reminder over de nulhypothese: De Groot (1961) noemde de toetsing een *'ignorant thing'*. Cohen (1994) sprak over een ritueel: het vertelt ons niet wat we willen weten, want wat we willen weten is:

'...*given these data, what is the probability that the hypothesis is true'.* Het toetsen zegt: '...*given the null hypothesis is true, what is the probability of finding these or even more extreme data?*'

De p waarde beantwoordt de vraag hoe groot de kans is om met dit onderzoek de geobserveerde gegevens of een extremer resultaat te vinden. De p waarden zeggen ook niet

dat een replicatiestudie ook weer met 95% zekerheid een significant resultaten oplevert. Miller en Schwartz (2011, p. 359):

'...the initial result actually says hardly anything about what percentage of replication attempts should be successful'.

Dezelfde resultaten vinden is onwaarschijnlijk met grillige subjecten en nooit helemaal te controleren omstandigheden.

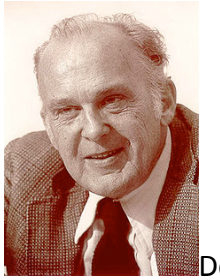
Niettemin moet een diagnosticus het hebben van robuuste vindingen. Ioannidis (2005) wijst er in zijn artikel met de provocerende titel: *Why most published research findings are false* op dat kleinschalig onderzoek met zwakke maar opwindende resultaten kan berusten op kanseffecten. Een onderzoeker mag niet over een nacht ijs gaan: replicaties zijn nuttig, want effecten en samenhangen moeten herhaald gevonden worden (Cohen, 1994; Schmidt, 2009). Er wordt gesproken over *slow science* (Alleva, 2006), waarbij de tijd genomen wordt om te kijken of resultaten overeind blijven over tijd en onder andere condities. *Slow science* waardeert replicatie- en meta-studies en toepassing: dat is de *proof of the pudding*. Mummendey (2012) bepleit deze houding en analyseert de Zeitgeist die vraagt om onverwachte nieuwe vindingen en veel publicaties. Het ligt voor de hand om zich te verbinden aan de uitkomst van eigen onderzoek. Verder is het vermelden van ESs naast of in plaats van significantieniveaus nuttig en van betrouwbaarheidsintervallen rond de correlaties en effecten. Bijvoorbeeld het verschil tussen de experimentele en controle groep is vier eenheden, de populatie correlatie is .30 met de confidentie-interval of het percentage mannen bij IT bedrijven is 75%, ± 10%.

Replicatiestudies Een themanummer van *Social Psychology* van april 2014 (redactie Lakens & Nosek) bevat replicaties van klassiek onderzoek. Daarbij wordt niet alleen het finale artikel met resultaten beoordeeld maar ook de onderzoeksopzet die vooraf ingediend moet worden. De oorspronkelijke resultaten werden niet steeds gerepliceerd. Het liet ook zien dat klassieke artikelen die vaak in handboeken vermeld worden niet altijd precies gelezen en geïnterpreteerd zijn. *Slow science* gaat niet over de onderzoeksopzetten zelf - bijvoorbeeld hun complexiteit en bijbehorende gesofisticeerde data analyse technieken. Ze zou kunnen bevorderen dat onderzoeksopzetten eenvoudiger worden, ecologisch valide en data analyses beter te doorzien zijn. Het laatste is niet meer het geval bij grote hoeveelheden geneste variabelen zowel aan de kant van de voorspellers als de criteria.
Er zijn al jaren bezwaren tegen nulhypothese-significantietoetsing en het vertelt ons iets dat we niet willen weten. Toch blijven redacties ze eisen. Als we naar resultaten van nulhypothesetoetsing kijken dan blijkt bijna iedere studie in de psychologie te slagen: 85% van de hypothesen is 'juist'. In de psychiatrie is het percentage zelfs 92% (Fanelli, april, 2010 in het vrij toegankelijke tijdschrift *PloS* (Public Library of Science) *One* en in

Sociometrics van januari, 2012). Ze stelt vast dat tussen 1990 en 2007 het percentage studies met positieve resultaten met 22% is toegenomen. Is dit inflatie van onderzoeksresultaten? Moeten we eruit afleiden dat psychologen veilige hypothesen formuleren en doorgaan met onderzoek dat goed te publiceren resultaten oplevert? Worden ze beter in het stellen van vragen? Publiceren ze de studie niet als effect ontbreekt? Is het terecht dat ze van negatieve resultaten niets kunnen leren?
De praktijk van rapportage wijzigt wel iets. Vooral gemakkelijk toe te voegen statistieken (ze rollen uit SPSS) worden vermeld. Effectgrootte voor verschillen (d-waarden) en voor correlaties (r-waarden) zie je vaker dan voorheen. De formules staan op al het internet: *d (effect size difference) = $M_1 - M_2$ / gepoolde varianties.* M_1 en M_2 zijn de gemiddelden van de experimentele en controlegroep. Na lang aandringen gaf Cohen de volgende vuistregel: d > 0,20 is een kleine, d > 0,50 een gemiddelde en d > 0,80 een hoge ES. In deze tekst zijn waarden tussen .20 en .50 bescheiden genoemd. Soms worden lagere waarden genoemd: tussen .20 en .37 gemiddeld en > .50 groot. Het is mogelijk om de d-waarde om te rekenen naar het feitelijke verschil. Bijvoorbeeld een halve standaarddeviatie winst op een IQ test met een gemiddelde van 100 en een SD van 15 is 7,5 IQ punten.

Terughoudendheid geboden Je mag concluderen dat validiteit van experimenten niet sluitend geregeld is door toetsing van nulhypothesen. Er zijn erg gauw significante resultaten en er wordt ten onrechte van uitgegaan dat een negatief resultaat je niets leert. Dit maakt misschien de weg vrij voor het (hermeneutisch) causaal modelleren in een *case study*, de Bayesiaanse werkwijze om uit te maken of een cliënt wel of niet in een categorie past en voor beschrijven en van een model voor het gedrag van een steekproef.
Klassieke experimenten in de sociale psychologie zijn herhaald om te zien of resultaten robuust zijn. De ESs bieden de diagnosticus een realistische schatting van de effecten van onafhankelijke variabelen of van behandelingen en van relaties tussen predictoren en criteria op basis van meta-studies. Als daaruit blijkt dat de variantie van de effectgroottes groot is, en dat is vaak het geval is die schatting overigens moeilijk.

Validiteit van quasi experimenten Campbell en Stanley schreven in 1963 een klassiek geworden tekst over *Experimental en Quasi-Expermental Designs.*


Donald W. Campbell was een praktisch methodoloog. Hij maakte samen met Stanley en Cook plaats voor proefopzetten die bedoeld zijn om causale relaties te bepalen

zonder dat er sprake is van een *true experiment*, dat wil zeggen geen willekeurige toewijzing van subjecten/onderzoekseenheden aan experimentele en controle conditie.

De uitwerking van validiteit bij quasi experimenten (QE) heeft een verruiming van het denken over validiteit opgeleverd (Shadish et al., 2002). Quasi experimenteren komt vooral voor bij veldonderzoek. Bovendien kan een *true* experiment een quasi experiment worden, als het niet lukt te randomiseren, dat wil zeggen de experimentele eenheden (proefpersonen) willekeurig verdelen over de condities, of de randomisatie gedurende het onderzoek in stand te houden. Dat gebeurt bijvoorbeeld door selectie, deelnemers hebben zichzelf opgegeven of ingeschreven of er is een deel dat halverwege niet meer aan het programma meedoet (*attrition*). Bij experimenten gaat het erom te achterhalen of een ingreep effect heeft en of de ingreep de oorzaak van het effect is. Effecten van *social experimentation* zijn complex en niet precies te bepalen. Hoe kun je uitmaken of het sluiten van pubs in het VK tussen de middag effect heeft op het aantal verkeersongelukken? Of hoe stellen we vast dat het studiehuis een betere aansluiting op het hoger onderwijs heeft bewerkstelligd dan het oude systeem of de BAMA structuur de studieduur bekort en betere studenten aflevert?
Cook en Campbell (1976, 1981) hebben quasi experimentele onderzoeksopzetten beschreven. Bij elk van die opzetten vragen ze zich af welke de bedreigingen zijn voor *vier* soorten validiteit. Van elke opzet wordt nagegaan wat de zwakke en sterke punten zijn met het oog op de geldigheid van de uitspraak of een ingreep effect heeft en wat daarvoor de oorzaak is. De soorten overlappen voor een deel met testvaliditeit en vullen ze aan.

Interne validiteit verwijst naar de vraag of de gedragsverandering niet aan een andere oorzaak toegeschreven kan worden als aan de bedoelde. Neem een interventieprogramma voor schoolkinderen: we denken dat vooruitgang in schoolprestaties aan het programma kan worden toegeschreven terwijl er op televisie een educatief programma was te zien waar bijna alle kinderen naar gekeken hebben. Bronnen van niet bedoelde, maar verklarende oorzaken zijn: kenmerken van de niet willekeurig aan de experimentele condities toegewezen proefpersonen, ongelijkheid door selectie, rijping, selectieve uitval van proefpersonen en interacties daartussen. Instrumenten en situaties kunnen ook bronnen van niet-bedoelde verandering zijn. Bijvoorbeeld statistische regressie naar het gemiddelde bij een extreme groep, herhaald testen, instrumenten die bij herhaald gebruik iets anders gaan meten en ten slotte gebeurtenissen met invloed op één van de twee groepen die niet met het interventieprogramma te maken hebben. De bedreigingen voor de interne validiteit zijn vooral het gevolg van de opzet, het *design*. Het gaat om de verklaring van een effect. Campbell (1986) wees erop dat de ingrepen molair zijn. Hij heeft het niet over laboratoriumonderzoek met moleculaire, precies te omschrijven en te manipuleren onafhankelijke variabelen. Bovendien wordt de uitslag in een specifieke context gevonden. Het resultaat is lokaal. Hij gaf *internal validity* een nieuw label: *local molar causal validity*. Bij deze beschrijving van validiteit zien we de tendens terug om vooral geen fout van de eerste soort te maken, dus beweren dat X de oorzaak is terwijl die het niet is.

Statistische-conclusie validiteit: In een experiment gaat het er in variantie analytische termen om zo te handelen dat de binnencelvariantie (variatie tussen proefpersonen, de experimentele eenheden) beperkt blijft en de tussencelvariantie (effect van de onafhankelijke variabele) aanzienlijk is. Als de F-waarde - de verhouding tussen tussen- en binnencelvariantie - flink groter dan 1 is(de verwachte waarde) dan kom je tot een effect. Deze validiteitsoort is ontleend aan Fisher (1918, 1e druk; 1970, 16e druk) die de variantieanalyse bedacht heeft om effecten van manipulaties te kunnen bepalen, bijvoorbeeld de invloed van verschillende hoeveelheden licht en water op de opbrengst van groenteveldjes. Deze validiteitsoort wordt bedreigd door gebeurtenissen waardoor de binnencelvariantie niet goed kan worden geschat, zoals te weinig of heterogene proefpersonen, slordige implementatie van onafhankelijke variabelen waardoor sommige proefpersonen meer van een experimentele training profiteren dan andere en slordig monitoren van de omgevingen van de controle- en experimentele groep. Het gaat erom verantwoorde schattingen te maken van de F-waarde door zicht te houden op allerlei bronnen die de binnencelvariantie onbedoeld en ongewenst kunnen verhogen. Deze soort is niet direct terug te vinden bij testvaliditeit.
Externe validiteit: Het doel van onderzoek is generalisatie van resultaten naar andere groepen personen en andere situaties. We doen geen onderzoek om hier en nu iets te ontdekken dat verder nergens werkt. Bedreigingen zijn onder meer dat het resultaat specifiek is doordat bijvoorbeeld twee behandelingen A en B interacteren en we niet weten wat A en B zelf waard zijn. Dit kan zich voordoen als gevolg van interacties van de behandeling (implementatie van de onafhankelijke variabele) met de specifieke samenstelling van de steekproef, met een specifieke situatie en tijdstip. Er wordt bijvoorbeeld een lesprogramma gegeven om de verschillen tussen kinderen kleiner te maken vóór ze naar de basisschool gaan. Het programma blijkt vooral te werken voor de groep die het 't minst nodig heeft. Het gevolg is dat de verschillen nog groter worden dan ze al waren. Later gaf Campbell dit type validiteit de naam van *proximal similarity* om te brede claims wat betreft generaliseren in te dammen. Op grond van wat werd geobserveerd maakte hij inferenties over behandelingen, ingrepen en omstandigheden die dicht bij het uitgevoerde onderzoek lagen.
Constructvaliditeit: Dit slaat bij quasi experimenten zowel op de onafhankelijke als op de afhankelijke variabele. De auteurs wijzen op bedreiging van deze validiteit door zwakke operationalisaties en matige empirische specificaties van constructen. Het gedragsbegrip moet niet over-, maar ook niet onder gerepresenteerd zijn en convergeren met gelijksoortige constructen en zich onderscheiden van andersoortige constructen. Het moet niet gevoelig zijn voor verschillen in methoden van gegevensverzameling. Dit is bekend van de *multitrait-multimethod-strategy* van constructvalidering (Campbell & Fiske, 1959). Constructvaliditeit verwijst ook naar gedragingen van de proefleider en proefpersonen. De laatsten kunnen immers een vermoeden hebben van de hypothese en zich daarnaar gedragen. De proefleider kan door zijn verwachtingen gedragingen van proefpersonen beïnvloeden. Een proefleider is geen *mechanical device*.

Discussie over quasi experimenteren Cook en Campbell combineren concepten van testvaliditeit met die van experimentele en research validiteit. De relatie tussen de vier soorten kan in praktijkonderzoek op gespannen voet staan. Hoge interne validiteit vereist een perfecte organisatie van veldonderzoek. De interventie moet gestandaardiseerd zijn voor statistische conclusie validiteit. Practici en proefleiders zijn echter geen machines of computers. Het doel van het onderzoek bepaalt de rangorde van de vier soorten. De

discussie over validiteit bij quasi experimenten kan uitgebreid worden naar ware experimenten want die kunnen degenereren tot quasi experimenten.
Het publiceren over quasi experimenten begon in de jaren 60 en liep door tot de jaren 90. De samenvatting is van Shadish et al. (2002). Deze is niet alleen methodologisch. Hij ademt de geest om de samenleving te verbeteren. Er wordt uitgegaan van maakbaarheid (Cook & Shadish, 1994) en dit is gekritiseerd. De auteurs '*…focus on real-life problems from a deep, but common sense perspective that led to guidance for causal inference*' (Rubin, 2010). Dit zou ertoe leiden dat ze op causale analyse in gerandomiseerde laboratorium experimenten niet op waarde schatten. Er is geen uitleg over correlationele ontwerpen, bijvoorbeeld regressie analyse, pad analyse en dergelijke omdat er geen onderscheid gemaakt wordt tussen het doel van inferentie (wetenschappelijke uitspraken) en wat onderzoekers leren van wetenschap. De discussie lijkt op wat Trendler (hoofdstuk 1) zei over de onmogelijkheid van meten. Als er geen Cartesiaans klaar en onderscheiden idee is van wat men meet (bijvoorbeeld intelligentie; effect van behandeling) dan is het zinloos om over statistische methoden te spreken. Formeel is dit juist maar als men er zich strikt aan zou houden, zou kennis over effecten van behandelingen in levensechte situaties en over samenhangen tussen voorspellers en criteria ontbreken.
Behalve dit technische commentaar wordt gewezen op niet correct uitvoeren en interpreteren van quasi experimenten. Aussems et al. (2011) verrichtten een inhoudsanalyse op quasi experimenten in 18 tijdschriften. Ze stellen vast dat een aantal niet goed opgezet en geanalyseerd is. Ze noemen vooral de *selection bias.* Het is lastig om dit te voorkomen want het gaat erom de proefpersonen te verleiden om mee te (blijven) doen. Vrijwilligers zijn geen aselecte steekproef. Ze vonden geen verschillen in kwaliteit tussen tijdschriften op basis van hun *impact rate*.

Empirische vergelijkingen Resultaten van QE ontwerpen kunnen empirisch vergeleken worden met experimentele en observationele studies. Ik heb geen vergelijking tussen *true* en QE gevonden, wel tussen QE en observationeel en *true* en observationeel. Voorbeelden:

(1) Jaffee et al. (2012) vergeleken QE en observationele studies bij antisociaal gedrag. De meeste studies over antisociaal gedrag van jongeren zijn correlationeel. Men zoekt naar voorspellers van bijvoorbeeld agressie, i.c. risicofactoren in gedragskenmerken van de jeugdige en zijn context. De voorspellers zijn onder meer harde, dwingende discipline, mishandeling, roken gedurende de zwangerschap, echtscheiding, pathologie bij de ouders en afwijken van leeftijdgenoten. De QE studies toonden kleinere effectgroottes dan de correlationele. De auteurs denken dat de hogere waarde gevolg is van de selecte steekproeven. Sommige risicofactoren zoals roken gedurende de zwangerschap en alcoholgebruik door ouders hebben mogelijk geen door de omgeving gemedieerde effecten. De studie suggereert dat de uitslagen verschillend kunnen zijn tussen de twee soorten onderzoeksopzetten. Het gaat er natuurlijk om of de verschillen zich binnen de betrouwbaarheidsintervallen rond effecten en correlaties bevinden. Een schatting mijnerzijds is dat ze klein zijn.

(2) Cook et al. (2008) identificeerden twaalf *within-study* vergelijkingen: *true* experimenteel versus correlationeel met dezelfde groepen. Drie daarvan stonden een vergelijking tussen een waar experiment en een regressiediscontinuïteit QE ontwerp toe. Het laatste verwijst naar het verschijnsel dat in het geval van een effect van de ingreep de intercept van de regressielijn verandert na de implementatie van de interventie. Shadish et al. vonden vergelijkbare resultaten. Een volgende vergelijking tussen *true* en QE (met *matching* van groepen op de pretest) leverde vergelijkbare resultaten op, vooral als matching lokaal is, dat wil zeggen als de groepen in elkaars buurt leven. De verschillen zijn klein. De gemiddelde waarden komen dichterbij elkaar te liggen als de omstandigheden (bijvoorbeeld door middel van correctie door middel van covariaten) erbij betrokken worden. Mijn conclusie is dat verschillen op basis van de opzetten klein zijn. Dit verleent de QE opzet een plaats in (sociaal) experimenteren. De impliciete overtuiging is dat men de uitslagen van ware experimenten het meest kan vertrouwen, daarna de QE opzetten en ten slotte de correlationele. Cook et al. vonden dat ze vergelijkbaar zijn als men de condities en de omgeving verdisconteert bijvoorbeeld door middel van covariaten.

Box Het verbeteren van gedrag en context door sociaal experimenteren

Cook en Campbell richten zich op de validiteit van sociaal experimenteren en nemen aan dat het onderzoek bijdraagt aan het verbeteren van het welzijn van mensen en hun omstandigheden. Sommigen vinden dat naïef. Er zijn goed- en kwaadwillende voorbeelden te vinden. Sint Augustinus, de bisschop van Carthago (Noord-Afrika) schreef in zijn 'Bekentenissen' in de 4de eeuw dat mensen betrokken zijn bij experimenten: '...*nos autem in experimentis volvimur*': wijzelf zijn verwikkeld in experimenten. Hij had daarbij niet controle en vooruitgang op het oog die uit sociaal experimenteren spreekt. Zijn boodschap is het ware geloof te behouden en afstand te nemen van de heidense Romeinse cultuur. De pedagoog Comenius zei in 1639 de beroemde woorden: *Tempus est* te vertalen met 'Nu moet het gebeuren' of 'Nu!'. Een populair magazine van de jaren 80 en 90 heette bijvoorbeeld 'Ouders van Nu'. Zijn werk ademt het idee dat mensen volmaakt gemaakt kunnen worden. Ze moeten er drie 'boeken' voor lezen: het dikste boek van God: de schepping, dat is zichtbare wereld, het boek dat de mens is want die is naar Gods beeld geschapen en het boek dat God de mens gegeven heeft: De Heilige Schrift (Voigt, 1997). Moderne pedagogen en filosofen schrijven regels voor die ons bestaan en gedrag verbeteren. Deze schrijvers overschatten effecten van regels en interventies. Ze geloven dat en zijn er niet op uit om het empirisch te verifiëren. Ze tonen een aanstekelijk optimisme, zoals ook Theodore Roosevelt deed in 1897 op het tijdstip dat de VS een wereldmacht werd. Hij zette zich in voor de *'great uplifting of mankind'*. Economen en onderwijskundigen wijzen er op dat overheidsprogramma's om werkgelegenheid en schoolsucces te bevorderen zelden geëvalueerd worden. Er zijn voorbeelden van sociaal experimenteren die na een veelbelovend begin *ontaarden.* De communistische leider voor Lenin, Trotski werd vermoord in opdracht van Lenin. Hij hield in 1932 een toespraak voor Deense studenten waarin hij mensen lichamelijk en psychologisch incompleet *(half made, half developed*) noemde. Training zou deze deplorabele toestand moeten veranderen. Ideologie en religie hebben veel gezichten. De Taliban wil het leven veranderen en dat was ook het geval bij de culturele revolutie in China, het regime van Pol Pot en de gedwongen heropvoeding in totalitaire staten.

Validiteitsgeneralisatie Van Berkel (1984) onderscheidde zes soorten: criterium, inhoud, latente trek, (quasi) experimentele, *face* validiteit en een rest categorie. Hij kwam tot 77 validiteit labels. We beperken ons tot de drie en voegen twee toe die af en toe in de literatuur opduiken. Validiteitsgeneralisatie is uitgewerkt door en voor organisatiepsychologen. Schmidt en Hunter (1993) vroegen zich af wat de *true* correlatie tussen intelligentie en persoonlijkheidskenmerken zoals integriteit, eerlijkheid, onkreukbaarheid en werkprestaties was. Intussen was er veel onderzoek over de relatie. Ze vatten de p-c correlaties op als eenheden van observatie. Validiteitscoëfficiënten kunnen gegeneraliseerd worden over verwante, gelijke criteria. Ze verwachtten dat de waarden dichtbij elkaar zouden liggen maar er was grote variatie: r-waarden tussen de r = .10 en r = .65. Ze weten deze variatie aan administratieve fouten vooral onzorgvuldigheid. Ze probeerden de variantie in p-c correlaties te binden aan verschillen in situaties, bijvoorbeeld soorten bedrijven en aan meetfouten. Na correctie voor de administratieve fouten schatten ze de *ware* correlatie. Verschillen tussen bedrijven en instellingen deden er nauwelijks toe. Er bleven substantiële waarden over na correctie voor onvolledige gegevens, scoringsfouten en statistische artefacten. De laatste betroffen te kleine en homogene steekproeven, gebruik van onjuiste correlaties (non-Pearson correlaties, bijvoorbeeld bij discontinue variabelen), ontbreken van correctie voor *restriction of range* en niet compenseren voor onbetrouwbaarheid van criteriummaten.
In hun meta-studie die 80 jaar onderzoek naar voorspellen van werkprestaties samenvatte (Schmidt & Hunter, 1998), bepaalden ze de ware correlatie tussen *general mental abilities* en criteria zoals werk samples, uitslagen van gestructureerde interviews, *peer ratings* en werkprestaties. De geschatte waarde was r = .50. De meta-studie laat zien hoe hoog de correlaties *kunnen* zijn als het onderzoek gecorrigeerd wordt voor administratieve en meetfouten. Dat is ideaal geen werkelijkheid. Validiteitsgeneralisatie kan nu als een onderdeel van meta-studies beschouwd worden die de robuustheid van samenhangen tussen voorspellers en criteria en verschillen tussen experimentele en controle groepen schatten.

Incremental validity verwijst naar de omstandigheid dat bepaalde tests of *other modes of assessment* iets toevoegen aan wat so wie so al bekend is en de p-c correlatie kan verhogen. Dit concept was een tijdlang niet prominent maar het is nieuw leven ingeblazen in een speciaal nummer van het tijdschrift *Assessment* (Hunsley, 2003). Garb (2003) vergelijkt instrumenten om psychopathologie vast te stellen en beweert dat gestructureerde interviews, MMPI, persoonlijkheidsvragenlijsten en *self-report* angstschalen *incremental validity* - toevoegende waarde - hebben ten opzichte van projectieve technieken vooral van de Rorschach. Toevoegende validiteit is gedefinieerd als de mate waarin een meting een verschijnsel ‘verklaart’ in verhouding tot andere metingen. Garb (1998) heeft een omvangrijke monografie geschreven over klinische versus statistische predictie en concludeert dat de statistische methode met objectieve tests het meest opleverde.

Johnston en Murray (2003) zijn op zoek naar instrumenten en procedures die het voorspellen van prestaties en stoornissen bij kinderen verbeterden. Ze komen niet verder dan multiple metingen aan te bevelen. Is dit het idee van hoe meer en hoe gevarieerder, hoe beter? Hunsley en Meyer (2003) wijzen op factoren die tot *incremental validity* kunnen leiden: verhogen van de betrouwbaarheid van de metingen, het aggregeren van data over meetmomenten, situaties, methoden en bronnen van informatie. Dit komt neer op het verhogen van *betrouwbaarheid* van de metingen. Toevoegende validiteit is een zinvolle procedure als een nieuw instrument de arena van het voorspellen van een criterium betreedt. Als er winst is weet je meteen wat je er aan hebt. Feitelijk gebeurt dit weinig, terwijl het aantal tests toeneemt. De constructeurs zullen het te veel werk vinden om hun nieuwe instrument te vergelijken met voorhanden voorspellers. Men zou bij het ontwikkelen van tests of voorspellers rekening kunnen houden met de eis iets toe te moeten voegen met het oog op een specifiek criterium, bijvoorbeeld recidive (Haynes & Lench, 2003). Hoewel *incremental validity* een zinvol concept lijkt, is het gedekt door drie al bestaande praktijken:

(1) Diagnostici maken gebruik van uiteenlopende gegevens: demografische, historische, *self-reports*, observatie, symptomen, interview, test- en vragenlijstgegevens. Ze gebruiken de resultaten van de lineaire technieken die de gegevens verbinden om een criterium te voorspellen in een steekproef en passen dat toe op de cliënt. Ervaring heeft geleerd dat het aantal voorspellers dat bijdraagt beperkt is. Tussen de drie en de vijf is meestal genoeg.
(2) Om validiteit te verhogen is het vanzelfsprekend dat betrouwbaarder instrumenten meer bijdragen. Immers, onbetrouwbare variantie kan niet aan criteriumvariantie gebonden worden. Dat is een tautologie.
(3) Verhogen van de validiteit komt voort uit het willen vergroten van de klinische betekenis van instrumenten. Dat is meer dan statistische significantie. Het was al aan de orde bij het criteriumprobleem: een onmiddellijk criterium is gemakkelijker te voorspellen dan een *ultimate* criterium.

Beslissen tussen opties De validiteit van beslissingen is gewaarborgd door een procedure, bijvoorbeeld verantwoord redeneren (Toulmin), MAUT, de regel van Bayes of het HTM. De gevoeligheidsanalyse van MAUT komt tegemoet aan de vraag of het resultaat invariant is onder andere gewichten van de attributen. Er worden gewichten van *dezelfde* attributen gewijzigd. De procedure is een analogon van kruisvalidering waarbij nagegaan wordt of een bevinding in een nieuwe, gelijkwaardige steekproef overeind blijft. Dit geldt ook voor bepaling van de kansen in de Bayesiaanse procedure. Deze kansen zouden evenals in de gevoeligheidsanalyse op goede gronden gewijzigd kunnen worden en men kan bepalen in hoeverre dat gevolgen heeft voor de uitkomsten. De validiteit van het HTM wordt geborgd beschouwd door het volgen van de stappen, van het protocol. Het is gerechtvaardigd om te vragen naar vergelijkende validiteitstudies tussen het HTM en spontane professionele procedures. Deze studies zijn zeldzaam. Ze vormen geen uitdaging voor onderzoekers,

omdat een echte tegenstelling en strijdende partijen ontbreken zoals het geval is bij klinische tegenover statistische predictie. Er is wellicht ook een winnaar te verwachten zoals bij de klinisch-statistisch controverse. Die kan breed uitgemeten worden, maar valt in de praktijk mee. Ten slotte is het tijdrovend want complexe procedures moeten vergelijkbaar gemaakt worden. Dus komt het er niet van.
De diagnosticus is gebaat met informatie over de differentiële validiteit van procedures. Dat helpt kiezen voor de beste procedure. Daarbij spelen behalve goede resultaten - weinig valse positieven en negatieven - uitvoerbaarheid en het aansluiten van de protocollen bij het spontane proces van diagnosticeren een rol.

Samenvatting en conclusie

Historisch gezien was het meten van geheugen en snelheid van denken het eerst. Daarna ging het om toepasbare kennis: selectie van rekruten en leerlingen. Dit leidde tot de volgorde: 1. predictieve validiteit 2. inhoudsvaliditeit. Er werd daarnaast onderzoek verricht dat niet meteen over deze twee ging, maar wel inzicht gaf in de constructen zelf en bijdroeg aan theorievorming: 3. constructvaliditeit. *Face* validiteit werd buitengesloten. Cronbach en Meehl (1955) zetten constructvaliditeit op de kaart en dat heeft gevolgen tot heden. Loevinger (1957) en Magnusson (1966) vonden de drie voldoende om te spreken over volledige validering van een test. Guion (1980) spreekt over de drie als *a holy Trinity* van de testvalidering.
Messick (1989, 1994) heeft het begrip *constructvaliditeit* uitgewerkt. Zijn opvatting is terechtgekomen in de *Standards* van 1999 en ondanks kritiek niet verdwenen uit de voorlopige van 2014. Het valideringsproces wordt omschreven als strategieën om testscores te interpreteren. Dat zijn er veel en het heeft als doel bestaande concepten te integreren. Het proces wordt zo complex dat het voor de professional aan bruikbaarheid inboet. Sommigen stellen dan ook voor om een rangorde aan te brengen. Daarbij staat predictieve validiteit bovenaan. Messicks omschrijving *all kinds of inferences and meanings of tests* rakelde de discussie over de status van constructen weer op. Is het een *summary index* of ligt er iets aan ten grondslag: een latente trek, een formeel af te beelden structuur, een genererend mechanisme, een biologisch substraat? Volgens Trendler is dit een gebed zonder einde. Borsboom zou daar mee debet aan zijn met zijn definitie: *a test refers to an existing, i.e., ontological anchored latent trait*. Een oordeel wordt in deze tekst uitgesteld omdat betwijfeld wordt of het zoeken naar een materiële oorzaak: biologisch, neurologisch, genen, hersenen, DNA, enzovoort een vruchtbare weg is.
Het criteriumprobleem heeft betrekking op typen criteria. Een criterium kan specifiek of globaal zijn en het eerste is beter te voorspellen. Verder worden onderscheiden onmiddellijke: vorig studieresultaat voorspelt volgende, *intermediate*: de relatie tussen eindexamencijfers en het behalen van het BA diploma en *ultimate*: X wordt een goede dokter gegeven zijn BA-resultaat. Ze voorspellen in de gegeven volgorde het criterium minder goed.

Verder zijn predictor-criterium correlaties gevoelig voor de variantie in de steekproef: *restriction of range.* De betrouwbaarheid van criteriummetingen is soms onbekend of *taken for granted,* niet voor discussie vatbaar, zoals bij beoordelingen door superieuren en managers of door beoordelaars die Holllands topmodel kiezen.
De correctie voor attenuatie is interessant in onderzoek om te zien hoe hoog een samenhang *kan* zijn. Voor de praktijk is het een te mooi om waar te zijn voorstelling van zaken. Schmidt en Hunter (1996) stellen 26 realistische onderzoeksscenario's voor om het veronachtzamen van meetfouten te voorkomen. Ze gaan er daarbij vanuit dat er voor psychometrische theorie geen wiskundige - die spreekt immers vanzelf - en inhoudelijke weerlegging: *substantive rebuttal* (p. 199) bestaat. Daarmee gaan ze voorbij aan de ongemakkelijke verhouding semantische theorievorming en meten.
Controle is het domein van het experiment en zijn validiteit wordt niet volledig gedekt door de alfa: $p < .05$ en $< .01$. Een onderzoeker kan door een hoge alfa de kans mislopen om een feitelijk bestaand verschil te vinden. De nulhypothese wordt gemakkelijk verkeerd gelezen en is niet realistisch. Het is de onwaarschijnlijke vergelijking met een nul casus. Ket is een *ignorant thing*. We doen alsof er niets aan de hand is. In hoofdstuk 1 zijn andere opties voor toetsing van effecten en samenhangen vermeld. Hier is gewezen op effectgroottes van verschillen tussen gemiddelden en robuustheid van samenhangen. Cohen was met de ESs een wegbereider van de hernieuwde aandacht voor robuuste vindingen en voor wat nu *slow science* genoemd wordt. Met enige aarzeling was hij bereid vuistregels te geven voor wat zinvolle ESs en rs waren en er zijn strenge en gematigde versies in omloop. De d-waarden: effectgrootte bij vergelijking van gemiddelden van de experimentele en de controlegroep en de rs: geschatte correlaties over p-c studies zijn meestal bescheiden tot gemiddeld volgens Cohens kwalificatie.
Voor het quasi experiment hebben Cook en Campbell vier typen validiteit beschreven die een combinatie zijn van test- en research validiteit. Deze ontwerpen gelden bij *social experimentation*. Analyse van interne validiteit geeft antwoord op de vraag of de relatie tussen ingreep en gedragingen causaal verklaard kan worden door een efficiënte oorzaak. Statistische conclusievaliditeit verstrekt informatie over het feit of ingreep en de gevolgen statistisch significant samenhangen of een verschil maakt op de afhankelijke variabele. Constructvaliditeit gaat over deugdelijke operationalisatie van interventie- en effectconstructen. Externe validiteit gaat over het verantwoorden van de generalisatie van de causale relatie tussen interventie en gedragseffect naar andere groepen, tijdstippen en plaatsen. Generalisatie heeft te maken met de vorm van steekproeftrekking. Is het populatie-gebaseerd *(probability sampling*), een gelegenheidssteekproef (*convenience sampling*), quota of homogene steekproeftrekking (Bornstein et al., 2013)? Van de trekking hangt af of en hoe er gegeneraliseerd kan worden.
Bij elk van de soorten beschrijven zij bedreigingen. Deze laten zien hoe subtiele invloeden het resultaat van een interventie, of de samenhang tussen voorspellers en criterium kunnen verstoren. Dat wordt weer gekritiseerd omdat het einde zo zoek is (Rubin) en een formeel

model geformuleerd moet worden voor de inferentie over effecten. Deze auteur valt terug op het ware experiment. De uitkomsten van experimenteel, quasi experimenteel en correlationeel onderzoek bij dezelfde onderwerpen worden zelden vergeleken. Er zijn kleine verschillen: de ESs van QE studies zijn lager dan van correlationele. Sociale experimenten berusten op het idee van maakbaarheid van de menselijke conditie. Er wordt soms over het hoofd gezien dat verbeteringen iatrogeen zijn en zelfs in terreur kunnen ontaarden.

De organisatiepsychologen Schmidt en Hunter stelden vast dat het aantal studies over vergelijkbare voorspellende studies toenam. Dit bood de mogelijkheid om de ware validiteitcoëfficiënt (ware is als ware score in de KTT) te schatten van predictoren voor werkprestaties: validiteitsgeneralisatie. Tachtig jaar onderzoek naar p-c correlaties: predictie van werkprestatie leverde waarden tussen de .10 en .65. Dit is een behoorlijke *range*. Ze probeerden die te verklaren door administratieve fouten bij het meten van predictoren en criteria. Ze wijzen er bovendien op dat studies met niet significante p-c correlaties waarschijnlijk niet gepubliceerd worden en tekenen bezwaar tegen deze *confirmation bias*.

Het streven naar hoge p-c correlaties leidde tot het toevoegen predictoren om p-c correlaties te verhogen. bekende factoren, zoals demografische condities, SES en sekse: *incremental validity*. Hoewel het idee sympathiek is, is het al bekend en zijn maatregelen eerder genoemd: verhogen van de validiteit van predictoren en criteria. De laatste 10 jaar is er nauwelijks aandacht voor geweest.

In tegenstelling tot de vergelijking van klinisch tegenover statistisch is er nauwelijks onderzoek naar verschillen in validiteit binnen en tussen de disciplineringsprocedures zoals correct argumenteren, HTM, Bayes-regel en MAUT. Er is geen onderzoek of ze verschillen tonen in valse positieven en negatieven. Er ontbreekt kennelijk één en een duidelijke tegenstelling tussen de procedures. Wellicht wordt vooronderstelld dat de procedures op, door en in zichzelf valide zijn, want het zijn logische werkwijzen. En tegen logica is niets in te brengen. Bij onderzoek is er geen duidelijke winnaar te verwachten; wel inzicht in nuances. Het is veel werk en daarom komt het er niet van. De diagnosticus is gebaat bij informatie over de differentiële validiteit binnen en tussen de procedures. In dat geval kan hij op basis van gegevens kiezen. De kans dat hij daar iets over vindt is niet groot.

In deze sectie is het validiteitsbegrip noodzakelijkerwijs *verruimd* tot diagnostische validiteit, dat wil zeggen validiteit van de diagnose en het diagnostisch proces. Beperking tot testvaliditeit houdt in dat validiteit voor de kleinste eenheden (items, tests) wordt onderzocht. Als dat goed uitpakt kan gegeneraliseerd worden naar complexe eenheden, bijvoorbeeld diagnostische procedures, zoals interviews, observaties, het lensmodel, regels voor correct argumenteren, HTM, MAUT en de Bayes-regel. De atomistische aanpak is niet algemeen aanvaard in de diagnostiek. Enerzijds wordt gepleit voor ‘over de grenzen van de sub-disciplines heen kijken’, interdisciplinair te werk gaan en integratie van theorieën na te streven. Een voorbeeld is Messicks *unifying concept*. Anderzijds gaat het om de kleinste eenheden: items en latente unidimensionele trekken waar de psychometricus goed mee uit de voeten kan.

7. Alternatieve concepten

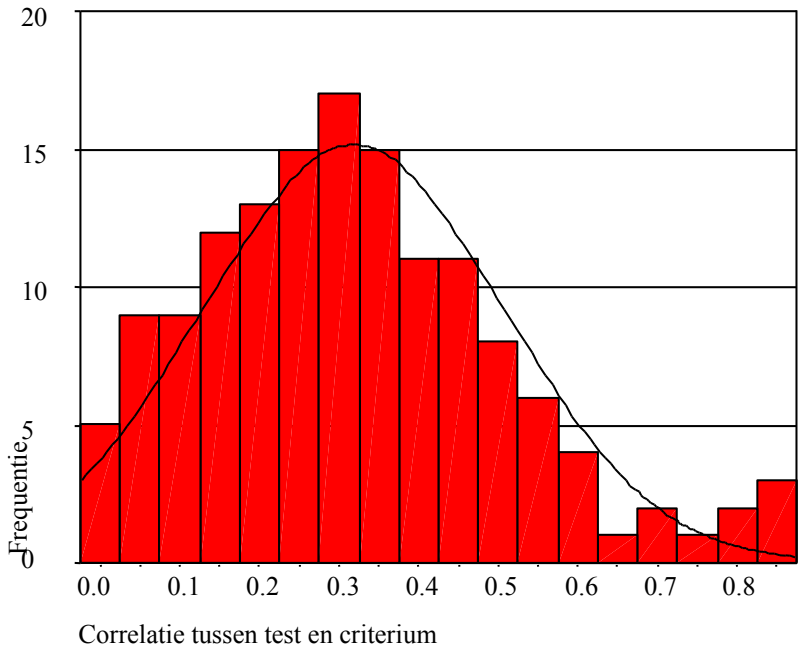
Er zijn wel veel validiteitsconcepten maar nauwelijks alternatieve. Een voorbeeld is het werk van Wheelright (1968) die quasi validiteitscriteria noemt zoals poëtische waarheid, metaforische spanning en waarheid van expressie. Dit is een romantische houding die sommigen niet on-, maar zelfs anti-wetenschappelijk noemen. Een voorbeeld: Mary Shelley (1818-1896) beweerde dat poëzie alle wetenschap omvat en naar poëzie uiteindelijk alle wetenschap verwijst (In: *Frankenstein or Schultze).* Dit wordt hier vermeld omdat men deze attitude een enkele keer in de praktijk tegenkomt. Funder (2001) verwijt de klassieke psychoanalyse soms zulke criteria te hanteren. Hij verbaast hem niet dat Freud meer aandacht krijgt in de faculteit van *Liberal Arts* dan in die van de *Social Sciences*. Deze bewering berust op het westerse idee dat waarheid/geldigheid en onwaarheid/ongeldigheid binair zijn: 'Waar is zeggen wat is en niet waar is zeggen wat niet is'. Dit stamt van Aristoteles en is overgenomen door Thomistische filosofen (Peters, 1957). Andere culturen staan doorlaatbare grenzen tussen waar en niet waar toe. Wellicht past openheid. Waar en niet waar kennen verschuivende grenzen zoals Merleau-Ponty (1945) beweerde: het nog-niet-redelijke opnemen in een verbrede rede.

Samenvatting en conclusie

Er zijn nauwelijks alternatieven voor psychometrische en experimentele/research validiteitscriteria. Afwijkende voorstellen worden opzij gezet als anti- of onwetenschappelijk. Dit weerspiegelt het succesvolle westers rationalisme en het logisch positivisme als basis voor valide uitspraken over gedrag van personen. Mogelijk is dat een weliswaar doorzichtige en controleerbare wijze van werken maar ook een beperkte manier om naar het probleem van de cliënt te kijken.

8. Een aanvaardbaar niveau van predictieve validiteit

Predictieve validiteit heerst in het test- en diagnostiekbedrijf. Daar is een reden voor: een diagnosticus kan iets over het gedrag van een groep en een individu voorspellen. Hoe goed kan hij dat? Niet de 100% waar Allport het over had want er zijn geen perfecte samenhangen. Het is niet realistisch dat te eisen. Voorspellers en criteria zijn niet perfect betrouwbaar te meten. Mensen gedragen zich grillig, zelfs in stabiele contexten. Er is altijd variatie en adaptatie bij een levend organisme. Een diplomatiek antwoord op de vraag is: 'het hangt ervan af': is het voor schoolkeuze, beroepskeuze? Cohen komt ons te hulp met zijn (gematigde) vuistregel: r = < .10 is laag, verwaarloosbaar; r = > .30 is gemiddeld; en r = > .50 is goed/hoog. Dit verschilt *negligibly* van de waarden die corresponderen met de rs, afgeleid van de d-waarden. Naast deze *educated guess* van Cohen zijn er meta-meta-studies over predictieve validiteit van tests voor psychologische en biologische criteria(Figuur 2).



Figuur 2: Meyer et al.'s (2001) predictieve validiteitscoëfficiënten uit 125 meta-studies in een grafiek gezet.

Cohens (gematigde) waarden zijn realistisch als we naar resultaten van bekende meta-studies kijken:

(1) Meyer et al. (2001) combineerden de predictieve validiteitcoëfficiënten van 125 meta-studies. De verdeling van de correlaties is normaal (Kolmogorov-Smirnov-toets: Z = 0,71). Een dergelijke lijst geeft interessante informatie, bijvoorbeeld de correlatie tussen ascal (aspirine) en het voorkomen van een hartaanval is. 02 (een andere studie: r = .034). Dat is laag maar velen krijgen het voorgeschreven. Agressie op televisie en agressief gedrag correleren .14. Waarschijnlijk geldt hetzelfde voor de inmiddels 100 verschillende pestprogramma's. Dit is en zou wel eens laag kunnen zijn maar zolang subsidie blijft binnenkomen voor studie naar gedrag waar we bang voor zijn, gaat dit type onderzoek door. Therapie en zich beter voelen is .32 gecorreleerd, een gemiddeld (medium) resultaat volgens Cohens vuistregel maar deze vorm van hulpverlening staat steeds onder druk. Het gebruik van Viagra en seksueel functioneren hangt .25 samen. Dat is ondergemiddeld en op individueel niveau heeft het gebruik een onzekere afloop. De productie en gebruik gaan onverminderd door. Verder blijkt er geen verschil tussen de voorspelling van psychologische, medische en biologische variabelen. Dit is meteen aangevochten door medici en biologen, maar Meyer et al. (2002) kunnen hun conclusie redelijk overeind houden.

(2) Schmidt en Hunter (1998) voorspelden *overall job performance* op basis van intelligentietests en persoonlijkheidskenmerken. Ze corrigeerden voor administratieve fouten en statistische artefacten. Dit verhoogde de gerapporteerde waarden uiteraard. Na correctie kwamen ze tot een gewogen gemiddelde correlatie van .50. Dit is ongeveer één SD (.32 + .19) boven het gemiddelde van Meyer et al. (2001). Een ander voorbeeld: De karaktertrek Gewetensvolheid (Autonomie) is de beste voorspeller van werkprestaties. Er wordt een gemiddelde r van .31 vermeld in Amerikaanse studies. Mogelijk is deze geflatteerd want in 794 psychologie Utrechtse eerstejaars studenten werd een

waarde van r = .19 gevonden tussen Gewetensvolheid en propedeuse resultaten: de som van de tentamencijfers. Dezelfde gemiddelde waarde vermeldt Poropat (2009) in een omvangrijke meta-studie.

Samenvatting en conclusie

Predictieve validiteit wordt het meest gewaardeerd. Differentiële validiteit tussen de disciplineringsprocedures van het diagnostisch proces is nauwelijks bestudeerd. Cohen formuleerde een vuistregel om de correlaties tussen test en criteria te interpreteren: < .10 is verwaarloosbaar, tussen de .10 en .30, indien significant, matig, > .30 is gemiddeld en > .50 is goed. Deze regel krijgt steun in twee grote meta-studies over de predictieve waarde van tests voor psychologische, medische en biologische criteria.

9. Een aanvaardbaar niveau van constructvaliditeit

Voor constructvaliditeit worden geen getallen genoemd zoals bij predictieve. Hoe hoog moeten items of tests correleren om naar eenzelfde construct te verwijzen? Constructvaliditeit is ondergedetermineerd en open voor interpretaties. De vraag wat een aanvaardbaar niveau is, kan men niet beantwoorden. Hoe hoog moet de lading van een item op een factor zijn? Er zijn misschien ongeschreven regels en afspraken. Strikt genomen zouden de items als parallel opgevat moeten worden. Dit geldt ook voor tests die hetzelfde construct meten. Dit is een te strenge eis en zou een warboel van unieke constructen opleveren. Daarom zijn er regels, bijvoorbeeld deze: een correlatie van ≥ .70 betekent verwijzen naar eenzelfde construct, onder bepaalde condities is een lading ≥ .30 op een factor voldoende om te verwijzen naar die factor, onder de conditie dat er geen hogere lading op een andere factor is. Bij IRT items kan naar de itemreponsfunctie gekeken worden. Er zijn geen expliciete normen voor de aanvaardbaarheid van waarden van correlaties, hoogte van ladingen en vorm van de IRFs maar wel conventies en stilzwijgend aanvaarde regels.

10. Reflectie en evaluatie

Procedures om betrouwbaarheid en validiteit van instrumenten en apparaten te bepalen zijn vooral ontwikkeld voor de industrie. Het proces van meten bijvoorbeeld met een vulmachine moet een stabiel resultaat opleveren. Er mag niet verspild worden en er mag niet te weinig in de flessen, pakjes, doosjes zitten: betrouwbaarheid. Ook moet het product doen waarvoor het gemaakt is: een auto moet vijf jaar rijden zonder mechanisch mankement: validiteit. De nulhypothese past hier met de nadruk op het type I fout. Er wordt geen nieuw product in de markt gezet als 5% niet zou doen waarvoor het gemaakt is. De alfa moet zelfs kleiner zijn. De concepten hebben ook invloed gehad op de productontwikkeling in de psychologie van tests, vragenlijsten en apparaten om reactietijd te meten. Ze hebben nauwelijks onderzoek naar betrouwbaarheid van complexe diagnostische procedures bevorderd.

Betrouwbaarheid wordt uitgewerkt onder de entry *reliability theory.* Dat is het zelfverklaarde monopolie van psychometrici. Ze wijzen diagnostici op fouten in metingen en op veronachtzamen van kenmerken van theoretische attributen. Bij de KTT en IRT horen traditioneel bepaalde analyses waardoor multivariate technieken buiten beschouwing blijven. Daar is geen reden voor. Toetsende factoranalyse *(Confirmatory Factor Analysis*: CFA), variantieanalyse (Anova: Cronbach et al., 1970), structurele vergelijkingstechnieken, en attitude schaling zijn geschikt om kenmerken van items en tests te bepalen.

De betrouwbaarheid van procedures - interne consistentie, stabiliteit en interdiagnostici overeenkomst - krijgt weinig aandacht. Het concept vooral uitgewerkt voor rangorde en interval gescoorde tests en vragenlijsten. Zulk onderzoek kan de bruikbaarheid van de procedures vergroten en aanleiding zijn om ze te veranderen. Validiteit en betrouwbaarheid zijn niet strikt te onderscheiden. Zodra er sprake is van meten en interpreteren van een theoretisch attribuut of latente trek speelt constructvaliditeit een rol.

Betrouwbaarheidstheorie is een statistische exercitie. Objectiviteit is gedefinieerd als uitsluiten van subjectiviteit. Objectiviteit is ook recht doen aan het object van studie. Deze opdracht is moeizaam te verwezenlijken met de empirisch-analytische methodologie en het logisch positivistisch epistemologisch kader. Er is daar weinig ruimte het epistemisch aspect van een gedragstheorie. Dat gaat over relatie tussen de bewering over gedrag en het gedrag zelf. Leken hebben daar geen moeite mee.

Waarheid, geldigheid, validiteit en objectiviteit zijn begrippen die veranderen gedurende de cognitieve ontwikkeling. Andere opvattingen over objectiviteit, zoals de hermeneutisch/ interpretatieve en kritische spelen nauwelijks een rol. Messick ontkomt (ongewild?) met zijn interpretatie van testscores niet aan hermeneutiek. Zijn idee van *consequentiality* zou men kunnen lezen als kritisch. De twee laatste stromingen zijn overigens nauwelijks zichtbaar bij test- en researchvaliditeitsonderzoek.

In het diagnosticeren speelt de persoon van de diagnosticus een rol ook al doet hij zijn best het protocol strikt te volgen en inwisselbaar te zijn. Het getuigt van realiteitsbesef de ‘afwijkingen’ te onderzoeken in plaats van ze als verwijderd te beschouwen. De p-c correlaties worden zonder onderscheid op elke cliënt toegepast. Af en toe wordt de werking van een moderator variabele, bijvoorbeeld sekse, SES of niveau van opleiding in rekening gebracht, omdat ze de p-c correlaties drukken. Dit wordt waarschijnlijk ten onrechte niet uitgewerkt op individueel niveau. Bornstein (2011) vraagt aandacht voor het feit dat cliënten individueel verschillen in de processen die hen tot de antwoorden op items en taken brengen. Hij pleit voor een *Process Focused* (PF) model voor valideren. Het gaat erom vast te stellen hoe en of een cliënt een aantal bekende en voorspelbare psychologische processen laat zien gedurende het diagnostisch onderzoek. Deze worden uitgelokt door de taak en de interactie met de diagnosticus. Dit klinkt aannemelijk maar is moeilijk uitvoerbaar. Bornsteins PF benadering wordt meer gepreekt dan uitgevoerd.

Betrouwbaarheidscoëfficiënten zoals test-hertest en interne consistentie zijn niet geschikt om ontwikkeling te bepalen. Daarnaast tekenen clinici bezwaar aan tegen psychometrische

schalen omdat ze onvoldoende discrimineren tussen patiënten met dezelfde stoornis (Fava et al., 2004). Feinstein (1987) stelde een andere aanpak voor onder de naam *Clinimetrics* om klinische verschijnselen een plaats te geven zoals soorten stoornissen binnen een bepaalde klasse, ernst, ordening, snelheid van ontwikkeling van een stoornis, ernst van erbij betrokken andere symptomen (co-morbiditeit), problemen met lichamelijk functioneren en verloop van dagelijkse activiteiten. Dit klinkt aantrekkelijk maar vraagt (te) veel van het oordeelsvermogen en observatievaardigheid van de clinicus.

Samengevat: betrouwbaarheidscoëfficiënten geven de diagnosticus aanwijzingen over herhaalbaarheid en stabiliteit van categorieën en testscores. Hij kan op basis van de problematiek van de cliënt kiezen welke passend is. Er is weinig bekend over de betrouwbaarheid van complexe diagnostische procedures.

Het alledaagse begrip van validiteit bevat zowel het idee van geldige, ware, objectieve uitspraken als van waarachtigheid. Het 'hoe iets is' en 'hoe iets behoort' liggen dicht bij elkaar. Dit is een categoriefout die leek en professional gemakkelijk maken maar het schrikt filosofen af. In de alledaagse opvatting gaat het zowel om accuraatheid bij het beschrijven van de sociale werkelijkheid als om oprechtheid. We gaan er stilzwijgend van uit dat de boodschappers: expert, diagnosticus en wetenschapper vertellen hoe personen en objecten in elkaar steken omdat ze meer weten en ons niet bewust voorliegen uit eigenbelang. Harris (2007) laat zien dat kinderen vanaf 4 jaar de accuraatheid van informatie en de oprechtheid van boodschappers proberen in te schatten. Messick (1989) voegde *consequential validity* toe aan de klassieke drie. Daarbij leiden onrechtvaardige gevolgen van testen er toe ze *niet valide* te noemen.

De onderwijspsycholoog Perry (1970) onderzocht bij studenten de epistemologische en ethische ontwikkeling. In het gedrag van aankomende wetenschappers en professionals liggen opvattingen van *accuracy* en *sincerity* dicht bij elkaar. Het is goed het onderscheid te kennen maar je hoeft er niet van op te kijken dat ze elkaar bij cliënt, diagnosticus en wetenschapper verbonden zijn.

Er zijn veel validiteitconcepten. Van Berkel (1984) kwam op zeventig uit. Hij telde ook statistische procedures mee. De grote drie: *holy Trinity* (Guion, 1980) dekken redelijk de activiteiten die in valideringsstudies verricht worden. Het heeft voordelen om het aantal beperkt te houden. De Oudtestamentische Mozes was geen geniaal wetgever geworden als hij in plaats van de tien geboden er tweehonderd had gemaakt.

De uitwerking van constructvaliditeit als *unifying force* (Messick) lijkt op Cronbachs poging eenheid te brengen in de testtheorie. Beide zijn elegante, erudiete en hermeneutische exercities maar voor professionals nauwelijks bruikbaar, zoals Sheppard al opmerkte. Dit maakt het volgen van de *Standards* van 1999 en waarschijnlijk ook de voorlopige van 2014 moeizaam voor de praktijk. Ze zijn complex, groot in aantal, betreffen uiteenlopende doelen en zijn onbegrensd. Waar houdt valideren op? Constructvaliditeit kreeg aandacht en vorm in 1955 door Cronbach en Meehls klassieker en bereikte zijn top in 1989 bij Messick. Psychometrici willen terug naar de eenvoud van de latente trek en het item met zijn IRF.

Daar hoort een domeintheorie bij die afgebeeld moet kunnen worden met psychometrische modellen, uitmondend in het schalen van een item en het attribuut. Gigerenzer beweert dat er een dichotomie geforceerd wordt. Trendler beweerde dat het theorie-meten debat en in het kielzog daarvan de construct model-interpretatie van scores eeuwig door kunnen gaan als er geen orde op zaken wordt gesteld. Dat betekent bij hem dat psychologen moeten ophouden te denken dat ze kunnen meten als natuurkundigen (hoofdstuk 1).
Een construct wordt naast *summary index* gezien als veroorzaker van individuele verschillen. Hoe moet die oorzaak geduid worden? Biologisch evolutionair, neurologisch (snelheid, reactietijd) of psychologisch, dat wil zeggen als intenties, doelen en redenen van een individu? Hoe bijvoorbeeld een Piagetiaans stadium te interpreteren? Men kan moeilijk beweren dat het concreet operationeel stadium de efficiënte oorzaak is van het oplossen van conservatietaken (Tacq, 2011).
De Noorse psycholoog Jan Smetslund (1999, 2004, 2009) wijst op een probleem bij p-c studies. Als het criterium qua betekenis onderdeel is van de predictor dan is er een semantische overeenkomst en geen empirische relatie. Een voorbeeld de studie van *happiness*: het prediceren van dat criterium bevat items die welzijn en geluk insluiten zoals positief temperament, kijken naar de zonnige kant van het leven, niet piekeren, welvaart ervaren en vrienden hebben en geld genoeg hebben om een interessante hobby uit te oefenen.
Predictieve validiteitstudies hebben een oceaan aan p-c correlaties opgeleverd. Meta-studies maken af en toe de balans op. Twee omvangrijke laten een gemiddelde van r = .31 (SD .19) en r = .50 (range .12 - .65) zien. De laatste waarde wordt eerst bereikt na correctie voor administratieve fouten en statistische artefacten die kennelijk één SD aan het gemiddelde toevoegen. Zijn dit waarden die men uiteindelijk kan behalen? Waarschijnlijk wel. *Try harder* zal niet helpen. Het is eerder zaak te achterhalen *waarom* dit de grens is. Als voor de validiteit van (quasi) experimenten hetzelfde geldt, komen we in de psychologie ongeveer halverwege in de pogingen tot predictie, controle en beslissen.
Validiteit is uitgewerkt voor tests en vragenlijsten en minder voor andere diagnostische procedures. Vergelijkende studies over de validiteit van de procedures zijn vruchtbaar voor de praktijk. Ze helpen te kiezen tussen procedures als deze gebruiksvriendelijker gemaakt zijn. Daarna kan ook bepaald worden of ze verschillen in validiteit en het aantal juiste diagnosen kunnen verhogen.
Over validiteit bij experimenten is opnieuw nagedacht toen resultaten van gemankeerde, quasi experimenten (QE) geïnterpreteerd moesten worden. Van *true* experimenten wordt aangenomen dat de validiteit door het ontwerp geborgd is. De onderscheidingen van Cook en Campbell: interne, externe, statistische conclusie- en constructvaliditeit zijn een aanvulling. Ze zijn ook nuttig voor het interpreteren van resultaten van ware experimenten. Vergelijkingen tussen de drie typen onderzoeksopzetten, quasi en true experimenten en correlationele om oorzaken te vinden *verschillen minder* dan de *true* experimentalisten lief is (Cook et al., 2008). Als men nauwkeurig kijkt naar de storende condities dat wil zeggen

validiteitsbedreigingen in QE experimentele en correlationele studies blijken de resultaten nauwelijks uiteen te lopen.
In ware experimenten wordt volstaan met een significantietoets met nadruk op de fout van de eerste soort. Dat is past bij industriële producten en bij het vinden van oorzaken voor gedrag en het daardoor kunnen uitoefenen van controle. De fout van de tweede soort is onderbelicht. Je kunt daardoor een zinvolle bevinding over het gedrag mislopen want er wordt gezegd dat een oorzaak die wel invloed zou kunnen hebben er niet toe doet. De fout van de tweede soort is bovendien groter als die van de eerste soort zeer klein wordt gekozen. Vanuit het oogpunt van controle zijn d-waarden informatief. Ze zijn gemiddeld genomen bescheiden tot 'medium'. Kortom, de significante F-waarden en de lage kans op een fout van de eerste soort zijn slechts een bescheiden begin bij het uitoefenen van controle over (on)gewenst gedrag van een cliënt.
Ten slotte zijn de betrouwbaarheids- en validiteitconcepten ontwikkeld voor individuele verschillen op latente trekken en attributen. Ze zijn niet uitgewerkt voor ontwikkeling van gedragingen en voor de invloed van sociale contexten. Theorie en betrouwbaarheids- en validiteitsstudies laten het een en ander liggen dat de diagnosticus zelf moet uitzoeken. Ze bieden echter ook middelen om de kwaliteiten van tests en experimenten te waarderen met oog op het probleem of vraag van de cliënt.
Het valt op dat betrouwbaarheids- en validiteitstheorie in het empirisch-analytisch en logisch positivistisch kader methodologische vereisten biedt. Die zijn niet inhoudelijk en niet op de cliënt gericht. Diagnostiek is n = 1 onderzoek. Er nooit een gepersonaliseerde betrouwbaarsheids- en validiteitscoëfficiënt bedacht. *Scientia non est individuorum*. De tweede wordt ingekaderd in lineaire voorspellingmodellen en (quasi) experimentele ontwerpen om de *causae efficientes* op te sporen in de populatie met behulp van representatieve steekproeven. *Scientia est non individuorum.* Ze wrijven je beide de complexiteit van het doen van een geldige uitspraak over samenhangen en oorzaken van (on) gewenst gedrag in. Dat mag en moet. Wetenschap is immers uit haar aard kritisch. Deze methodologische nadruk legt psychologische diagnostiek onder vuur. Twee relativerende opmerkingen hierbij. Ze schiet haar doel voor een deel voorbij en gaat niet in op betrouwbaarheid en validiteit van het diagnostisch onderzoek van de individuele cliënt. De resultaten van dat betrouwbare en valide onderzoek matigen de zekerheid waarmee (on)gewenst gedrag van een steekproef - laat staan een individu - voorspeld en verklaard wordt. Het vuur op basis van de betrouwbaarheids- en validiteitstheorie wordt gegeven haar prestaties te hoog opgestookt tegen de diagnose en het diagnostisch proces. Hun onderzoekers laten de diagnose van de cliënt en het diagnostisch proces bijna links liggen.

Onderwerpen en namen Hoofdstuk III

Standaardmeetfout

Soorten betrouwbaarheidscoëfficiënten:

- Parallel (theoretisch)
- interne consistentie
- stabiliteit
- overeenstemming tussen oordelaars
- generaliseerbaarheid
- Cohens kappa

Conditionele meetprecisie bij IRT

Vuistregels voor de hoogte van betrouwbaarheidscoëfficiënten

Impliciete opvattingen over validiteit

Impliciete eisen van *accuracy* en *sincerity*

Inhouds-, predictieve en constructvaliditeit

Validiteit: waarheid en objectiviteit in de

- empirisch-analytische,
- interpretatieve,
- kritische epistemologische betekenis

Consequentiality

Unified validity concept

Het criteriumprobleem:

- globale vs specifieke criteria
- onmiddellijke
- middellijke
- uiteindelijke criteria

Social experimentation

Quasi experimentele validiteit: bedreigingen van

- interne
- externe
- statistische conclusie
- constructvaliditeit

Resultaten van 125 meta-studies over p-c relaties

Validiteitgeneralisatie

Restriction of range en validiteit

Effectgrootten voor

- gemiddelden (d-waarden)
- correlaties (r-waarden)

Moderator variabele

IV Testtheorie en Diagnosticeren

De diagnosticus kan niet om de testtheorie heen. In de Nederlandse praktijk moet hij tests gebruiken die een keurmerk van de Commissie Testaangelegenheden van het Nederlands Instituut van Psychologen (Cotan) hebben. Het Centraal Insituut voor Toetsontwikkeling te Arnhem (Cito) ontwerpt test en toetsen voor tentamens, toetsen en examens in het onderwijs van het rijvaardigheidsexamen tot de toets voor groep 8 en in het bedrijfsleven. De diagnosticus draagt zelf ook bij aan de sterke positie van de testtheorie. Geconfronteerd met een probleem van een cliënt vraagt een diagnosticus zich af of 'er een test voor is'. Aan tests en vragenlijsten wordt de voorkeur gegeven boven procedures, zoals observatie, interview en anamnese omdat ze objectief en efficiënt zijn.

Testtheorie gaat over het scoren van items. Zij bevat modellen en functies om constructen af te beelden en de resultaten van meetoperaties te vergelijken met model vereisten. Er zijn twee testtheorieën: klassieke testtheorie (KTT) en moderne of item respons theorie (IRT). En, als er ergens twee van zijn wil het Westers verstand meteen de hiërarchie bepalen. De eerste gaat over de kwaliteit van de somscores van items. Deze hangt af van de meetfout en speelt de hoofdrol bij het schatten van de betrouwbaarheid. De IRT is gericht op de kwaliteit van de items en gaat na of een item past in een IRT model of Item Respons Functie (IRF) en of het onderscheiden is van andere items. Er wordt verder bepaald of ze een unidimensionele schaal vormen. De persoon heeft een positie op een latente trek of attribuut en het item heeft - gegeven die positie - een bepaalde kans om correct beantwoord of beaamd te worden. De latente trek, de functie wordt geïnterpreteerd als een individueel verschillen gedragskenmerk, bijvoorbeeld intelligentie of extraversie.

De IRT kan ook gebruikt worden om te toetsen of er sprake is van een ontwikkelingsschaal en om het effect van een onafhankelijke variabele op de afhankelijke te bepalen. De twee theorieën worden inleidend besproken, vergeleken en geëvalueerd met betrekking tot het beantwoorden van een vraag/probleem van een cliënt. Voor inzicht en uitwerking zijn er psychometriehandboeken en psychometrici.

De diagnosticus kan er niet om heen, zei ik boven, maar moet hij zich veel aantrekken van testtheorie? De klinisch-statistisch controverse, de eis van het maken van effectieve diagnose-behandel-combinaties en de betrouwbaarheids- en validiteitstheorie stellen weliswaar eisen, maar ze hebben hun wetenschappelijke pretenties half waargemaakt. De diagnosticus is zo een gewaarschuwd man (m/v). Hij krijgt geen protocol dat tot een onomstreden diagnose en diagnostisch proces leidt.

Wat levert de KTT op? Zijn meetfouten niet te voorkomen door zorgvuldig te werken? Wat levert de IRT op? Is de latente trek/ theoretisch attribuut een nuttig begrip voor diagnostiek? Waarom zijn er twee testtheorieën en hoe verhouden ze zich? Wat is het verband tussen testtheorie en theorie over individuele verschillen, ontwikkeling en sociale context? Draagt testtheorie bij aan kwaliteit van procedures, zoals interview, observatie, hypothesen formuleren en het diagnostisch proces?

1. Moet een diagnosticus zich iets van testtheorie aantrekken?

Een diagnosticus gebruikt tests en vragenlijsten omdat ze objectief en efficiënt zijn. Dit leidt ertoe dat procedures als interviews, observatie, raadpleging van *significant others,* documenten en files van een cliënt minder aandacht krijgen. Tests en vragenlijsten bestaan uit taken, opdrachten, vragen, items die item- en somscores opleveren. Deze leiden tot waarden op een theoretisch attribuut en voorspellen criteriumgedrag. Is de somscore een zuivere weergave van een psychologische trek, een gedragskenmerk? Als dat zo is kunnen we volstaan met die somscore. Dat is niet het geval want iedere meting bevat fouten. Dat is intuïtief te begrijpen want we weten dat fysieke kenmerken zoals gewicht, gezichtsscherpte en psychologische kenmerken, bijvoorbeeld angst en geluksgevoel niet goed worden weergegeven met een puntschatting die ook nog eens stabiel is door de tijd heen. Omdat we dat weten houden we er in het dagelijks leven ook rekening mee. Iemand is niet altijd en overal angstig, slim, gewetensvol.

Volgens KTT en IRT zijn fluctuaties inherent aan gedrag en meetfouten onvermijdelijk. Je moet een methode ontwerpen om de omvang te schatten. We weten niet of de score verwijst naar het kenmerk, construct of attribuut dat we willen meten. Daar moeten ook procedures voor verzonnen worden. Meetfouten komen uit oncontroleerbare bronnen. Ze kunnen niet vermeden of geneutraliseerd worden. Ze hebben te maken met de items zelf, grilligheid van getesten en veranderlijke contexten waarin ze zich bevinden. De klassieke testtheorie levert een werkwijze om meetfouten te schatten. Het idee is dat een bereik (*range*) kan worden bepaald op basis van geobserveerde scores. Daarbinnen liggen met een zekere waarschijnlijkheid de *ware of verwachte scores* van de onderzochte. Of het bedoelde construct er mee gedekt wordt, is een aparte vraag. Betrouwbaarheidsinterval is de benaming voor dat bereik.

De IRT bepaalt de kwaliteit van de items afzonderlijk door middel van de iteminformatie functie die naar moeilijkheidsgraad en discriminatiewaarde gespreid liggen op een latente trek of theoretisch attribuut, bijvoorbeeld numerieke of verbale vaardigheid. De betrouwbaarheidsschatting van een item en van een test met behulp van iteminformatiefuncties zijn onderdeel van IRT modellen. Deze tonen de positie van een persoon op de latente trek met behulp van een (niet) lineaire functie. De positie drukt de kans op het goede antwoord op een item uit. Stel er zijn personen met verschillende posities op de latente trek bijvoorbeeld grote IQ verschillen maar een item wordt door allen goed beantwoord dan is er iets mis met dat item. Ze meten mogelijk iets anders dan de latente trek waarop de getesten geacht worden te verschillen. Of stel dat de items theoretisch zo gemaakt zijn dat ze een vergelijkbare moeilijkheidsgraad moeten hebben maar toch beantwoorden personen met ongeveer gelijke posities op de latente trek ze verschillend. Ook dan is er iets mis. Er is misschien is er een andere latente trek bij betrokken waarop de proefpersonen wél verschillen. In beide gevallen worden deze items verwijderd. Ze voldoen niet. Ze vallen buiten een vereisten opgelegd door een item-respons-functie (IRF) gebaseerd op een IRT model.

De diagnosticus heeft de KTT en IRT nodig omdat items, tests, vragenlijsten en andere diagnostische procedures meetfouten bevatten. Een schatting helpt hem te bepalen hoe veel staat hij kan maken op een geobserveerde score van een persoon. De psychometricus hoeft er niet vanuit te gaan dat de diagnosticus aanneemt dat zijn vaststelling foutloos is. Hij heeft zijn marges, al zijn dat geen gekwantificeerde bijvoorbeeld in de vorm van een betrouwbaarheidsinterval.

Samenvatting en conclusie

De diagnosticus weet dat zijn vaststellingen niet perfect betrouwbaar zijn. Hij heeft ook zijn marges. Verder gebruikt hij tests en vragenlijsten. Testscores zijn niet zonder meetfouten. KTT bevat manieren om die fout te schatten en de betrouwbaarheid van de test te bepalen. IRT bevat functies (IRFs) om de kans te schatten dat een persoon met een vaste positie op de latente trek een correct antwoord geeft. Testtheorie verschaft de diagnosticus informatie over de psychometrische kwaliteit van test en vragenlijsten. Er is minder kennis over andere diagnostische procedures en het proces. De psychometricus overdrijft als hij aanneemt dat de diagnosticus spontaan uit zou gaan van de perfecte betrouwbaarheid van zijn diagnose en proces.

2. Klassieke testtheorie

Een meting is nooit precies. Als we steeds hetzelfde vragen of telkens eenzelfde opdracht geven is het antwoord van de cliënt niet identiek maar ook niet volstrekt anders.

Schatting van de meetfout In de KTT bestaat het antwoord of de score per definitie - zoals in 2013 nog eens uitgelegd door Trafimow - uit een stabiel *'true* = T) deel en een veranderlijk, flexibel, toevallig (*error* = E) deel

(1) Xj = Tj + Ej

De waargenomen score (X) op de test van Jan (j) is de som van een waar deel (T) van Jan en een veranderlijk, toevallig deel: de meetfout (E: de error van Jan).

Hoe van Tj een voorstelling maken? Het is een stabiele, niet waargenomen waarde van een persoon j (Jan). We kunnen ons voorstellen dat we Jan oneindig vaak dezelfde test voorleggen en er vanuit gaan dat vorige afnames geen spoor nalaten. De verwachte of gemiddelde waarde van al die afnames is de ware score van Jan. Afzonderlijke metingen leveren geen identieke resultaten op. Er is fluctuatie want Jan let niet altijd op, wisselt van stemming en hij leest de items niet nauwkeurig. Het item wordt ook door zijn vorm en bewoording niet altijd precies hetzelfde begrepen. Dat vormt de Ej van Jan. De som van de afwijkingen van de gemiddelde score is 0:

(2) Eexp Ej = 0

De verwachte waarde van de som van de afwijkingsscores is 0; het zijn immers afwijkingen van het gemiddelde.

De standaardmeetfout (Se) is een index voor hoe goed de ware score wordt benaderd door de herhaalde metingen of waarnemingen. De Se^2 is de variantie van de afwijkingen van het gemiddelde. Als die klein is komen de afzonderlijk geobserveerde metingen in de buurt van de ware/gemiddelde score. Een volgende veronderstelling is dat een fout op de ene gelegenheid niet die op de andere gelegenheden beïnvloedt. De correlatie tussen de meetfouten is 0.

(3) r EE = 0

Meetfouten zijn willekeurig dus de eerste heeft geen gevolgen voor de tweede, derde, de tweede niet voor de derde, enzovoort. Errorscores niet gecorreleerd zijn.

Ook nemen we men aan dat de variatie in meetfouten niet samenhangt met het niveau van de ware score. Als Jan bijvoorbeeld een hoge score behaalt op een test dan zegt dat niets over de omvang van de meetfout. Die is niet hoger, lager of gelijk, als hij een gemiddelde of lage score zou behalen. Dit kan betwist worden. Je kunt je afvragen of de meetfout even groot is op alle punten van de schaal. De veronderstelling in KTT is niettemin

(4) r TE = 0

De correlatie (r) tussen de niveaus van de ware score (T) en de grootte van de meetfouten (E) is nul (0).

Feitelijk kunnen we Jan niet steeds dezelfde items voorleggen. De truc van de KTT is om de Se^2 te bepalen in een populatie door een steekproef te trekken. Deze werkwijze leidt er toe dat iedere geteste dezelfde *error variantie* krijgt, afgezien van het niveau van zijn ware score. We kunnen ons afvragen of personen niet verschillen in hun *error score*. De één fluctueert mogelijk meer dan de ander. In de KTT er echter is één betrouwbaarheidscoëfficiënt voor een test. Het is niet ondenkbaar dat de *error range* verschilt tussen personen. In een experiment mag de experimentele groep wel een andere variantie hebben dan de controle groep. Daar wordt voor gecorrigeerd door het geometrisch gemiddelde te nemen. Fluctuatie geldt ook voor fysische kenmerken bijvoorbeeld lichaamsgewicht, bloeddruk, polsslag en huidgeleiding. Dit kan betekenen dat medicijnen geen uniforme uit- en bijwerkingen hebben. Ondertussen gaan farmaceutisch onderzoekers daarvan uit. Ze rekenen immers uit of een medicijn over individuen (een steekproef patiënten) een effect heeft en niet of het effect voor de afzonderlijke patiënten gelijk is. Af en toe zie je een uitzondering, een voorbeeld:

Wang et al. (2012) nemen de verschillende fluctuaties binnen personen serieus als voorspellers van gezondheid. Iemand met een variabele stemming of bloeddruk door de tijd heen is mogelijk wat (on)gezonder dan iemand met een stabiele stemming en bloeddruk. Herhaald meten van dezelfde persoon is praktisch niet uitvoerbaar en men heeft - misschien terecht - nooit gedacht aan identieke tweelingen en in de toekomst aan klonen voor de betrouwbaarheidsbepaling.

De KTT is bedacht door psychometrici. Zij gaan uit van oneindige populaties. Die bestaan niet want het totaal aantal personen dat op aarde leeft en zal leven is eindig. De zon verbrandt ons na ongeveer 10^{14} jaar. Ware scores moeten met behulp van een steekproef geschat worden. Je moet de bezwaren voor lief nemen dat de ware scores werkelijk variëren *tussen* personen en geobserveerde en error scores *tussen én binnen* personen. Het schatten van de meetfout door een steekproef te testen is een oplossing omdat de drie veronderstellingen in de formules 2, 3 en 4 aanvaard worden.
Er kan nog een relatie afgeleid worden. Deze zegt dat geobserveerde variatie van scores in steekproef (sX^2) of populatie (σX^2) gelijk is aan de ware score variantie ($\sigma^2 T$) plus de foutenscore variantie ($\sigma^2 E$).

$$(5)\ \sigma^2 = \sigma^2 T + \sigma^2 E$$

De geobserveerde variantie in een test is de som van de ware variantie en de error variantie, want de derde term van het merkwaardig product (rTE) is immers 0.

Betrouwbaarheid De aannames leiden tot de definitie van betrouwbaarheid. In het vorig hoofdstuk zijn de coëfficiënten vermeld. Dat zijn maten voor de accuraatheid waarmee we de verschillen in ware scores van subjecten kunnen schatten uit geobserveerde scores. We observeren Xj van een persoon/groep maar zijn geïnteresseerd in de Tj van de persoon/groep. Als we aannemen dat T en X lineair gerelateerd zijn dan is de betrouwbaarheidscoëffiënt de correlatie tussen T en X. Die correlatie is de verhouding tussen ware en geobserveerde variantie.

$$(6)\ \rho^2 XT = \sigma^2 T / \sigma^2 X$$

De correlatie van de ware scores in een steekproef/populatie en de waargenomen scores in een steekproef/populatie (ρXT) kan worden weergegeven als de variantie van de ware scores ($\sigma^2 T$) gedeeld door de variantie van de waargenomen scores ($\sigma^2 X$),

Betrouwbaarheid is per definitie de variantie in ware scores gedeeld door de variantie in geobserveerde scores. Als we een waarde van .85 verkrijgen betekent het dat 85% van de variantie 'verklaard, gebonden, gedekt, gevangen' is door de variantie in de ware scores. Als je een index wil maken kun je het idee van een parallel test gebruiken. Een parallel test heeft theoretisch, op papier, gelijke ware scores, varianties en relaties met andere tests en met criteria.

$$(7)\ Tj = T'j$$

De ware scores op een test (Tj) en een parallel test (T'j) zijn identiek - evenals hun varianties en relaties met andere tests en criteria - in dezelfde groep of in twee willekeurige steekproeven uit dezelfde populatie.

De correlatie tussen twee of meer paralleltests drukt de betrouwbaarheid van de test uit

$$(8)\ \rho^2XX' = \rho^2XT = \sigma^2T/\sigma^2X$$

De correlaties tussen een test en een parallel test is de betrouwbaarheid van de test; de verhouding tussen ware en totale variantie.

Als we een index willen hebben voor de betrouwbaarheid van een test moeten we een paralleltest maken. Dat is een vorm van herhaling. Bruno Latour (1977, Nederlandse vertaling, 1994) heeft er een aardige uitdrukking voor: *Repetition is a machine to produce differences with identity*. Hij is een controversiële schrijver met een constructivistische inslag dat wil zeggen met de overtuiging dat wetenschap en wetenschappelijke prestaties gefabriceerd worden. Zijn omschrijving van herhaling past wel bij waaraan bij betrouwbaarheidsbepaling gedacht wordt: door herhaling een score met *identity* produceren. In 1975 schreef de Nieuw-Zeelandse psycholoog en psychometricus Gregson een daarna vergeten boek over de psychometrie van gelijkheid (*Psychometrics of Similarity*). Het ging hem er om psychologische constructen in formele modellen te vangen. Psychologen dienen volgens hem geformaliseerde en kwantitatieve beschrijvingen en predicties te produceren. Normatieve modellen zijn daarbij een eerste stap. Hij heeft niet de illusie dat gedrag logisch, rationeel en consistent is. Formele modellen dienen te helpen psychologische constructen te verhelderen. Bij het vaststellen van gelijkenissen (al een kenmerk van intelligentie bij Aristoteles/Thomas van Aquino) betrekt hij gelijkheid van oordelen (*interjudge agreement*), kleinst waarneembare verschillen, waarnemingen (Gestaltpsychologie), passen in een categorie, van afbeeldingen in een vlak of in meer dimensies, enzovoort. De gelijkheid van scores (betrouwbaarheid) is zo een onderdeel van het generieke probleem van gelijkenis van gedragingen.
Kunnen we in de werkelijkheid herhalen en een getal uitrekenen dat het resultaat van de herhaling weergeeft? Volgens Heraclitus de voor-Socratische filosoof kan dat helemaal niet. Alles verandert en niets blijft hetzelfde maar voor psychometrici is weinig onmogelijk. Herhaling wordt empirisch gerealiseerd (a) door een steekproef subjecten twee paralleltests voor te leggen, gevormd uit twee helften (de even en oneven items) en de correlatie te berekenen. Dit is een index over hoe precies de rangorde van de subjecten overeind blijft in de twee parallelle tests. Dit is de equivalentie coëfficiënt.

$$(9)\ \rho(\text{equivalentie}) = r\ (\text{Testhelft 1, Testhelft 2})$$

We schatten de betrouwbaarheid van een test door de test in twee gelijke helften te verdelen (oneven-even items). We nemen aan dat de twee helften parallel zijn.

Een tweede manier om identiteit empirisch te realiseren is een steekproef subjecten twee keer dezelfde test te laten maken. Er is een redelijk tijdsinterval nodig zodat het spoor van de eerste afname verdwenen is. We rekenen de correlatie uit tussen de scores van de subjecten op gelegenheid 1 en gelegenheid 2.

(10) ρXX (stabiliteit) = r Testafname 1, Testafname 2)

We herhalen dat de ware variantie stabiel is: de T-scores in de populatie zijn stabiel. De *error* variantie fluctueert wel en dus de geobserveerde (totale) variantie. De correlaties tussen de Es zijn 0 en die tussen de Ts en de Es zijn ook 0. Dat betekent dat er geen correlatie tussen T en E hoeft te worden toegevoegd in de formule: het merkwaardig product is immers 0. Deze veronderstelling leidt er toe dat bij het toevoegen van items met dezelfde eigenschappen, *parallelle items*, de betrouwbaarheid hoger en bij weghalen van gelijkwaardige items lager wordt. De ware variantie blijft immers gelijk. Dit heeft geleid tot de *Spearman-Brown formule* voor testverlenging of -verkorting. Omdat deze het eerst voor de even-oneven nummers van de items van een test gebruikt is, wordt de formule meestal uitgeschreven met verdubbeling van het aantal items, of als correctie voor het halveren van het aantal items. Dit houdt in dat de betrouwbaarheid van de gehele test die verdeeld is in twee equivalente helften (twee paralleltests) Y en Y' gelijk is aan twee keer de correlatie tussen de testhelften gedeeld door 1 + de correlatie tussen de twee helften.

(11) rXX' = 2rYY' / 1 + rYY'

De betrouwbaarheid van de hele test X verdeeld in twee gelijke (equivalente) helften Y en Y' is gelijk aan twee keer de correlatie tussen de testhelften, gedeeld door 1 + de correlatie van de twee testhelften.

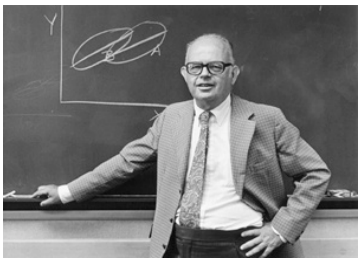
Deze formule kan worden aangepast voor iedere verlenging of verkorting door de 2 in het gepaste getal te veranderen.

Cronbach stelde in 1951 een derde realisering van herhaling, identiteit voor. We kunnen elke test verdelen in alle mogelijke helften. Dat is de bekende alfa van Cronbach die in vrijwel iedere testhandleiding staat.

$$(12)\ \rho Xx' = \frac{n}{n-1}\left(1 - \frac{\Sigma \sigma^2 Yi}{\sigma^2 X}\right) \quad \text{Cronbachs } \alpha$$

In woorden: Een factor iets kleiner dan 1 (n = het aantal items) wordt vermenigvuldigd met 1 minus een breuk. Als die breuk erg klein is dan is de waarde nabij de 1 x 1. Dat is een hoge alfa die een hoge interne consistentie uitdrukt. Dat wordt althans in handleidingen

geschreven. Boven de lijn treffen we de som van alle itemvarianties aan. Dat is de overgebleven variantie nadat de covarianties van de items met alle andere items er van afgetrokken zijn. Onder de lijn staan de varianties + covarianties (de totale variantie).



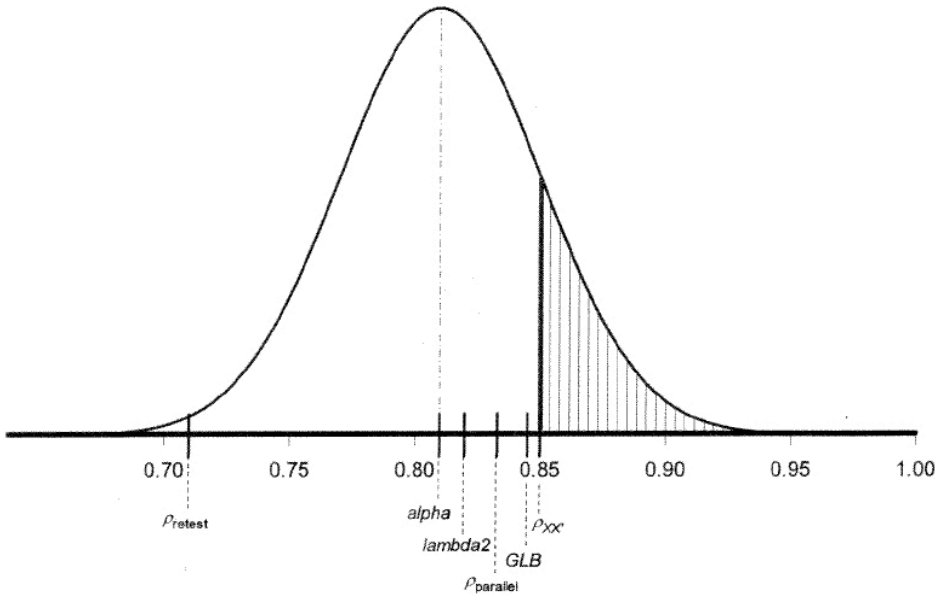
Lee J. Cronbach, de in 2002 overleden Stanford psychometricus en onderwijspsycholoog, begon zijn loopbaan als leraar exacte vakken. Hij verbreedde de KTT in de Generaliseerbaarheidstheorie en probeerde in 1977 samen met Richard Snow geschiktheden *(aptitudes*) te verbinden met behandelingen *(treatments*): ATI: een uitwerking is de huidige diagnose-behandel-combinatie (DBC) die zowel in de geneeskunde als de psychologie en pedagogiek voet aan de grond heeft gekregen. Dit idee was tot heden in de psychologie geen succes. Het 1977 boek staat vol met studies die geen ATIs laten zien. Zijn alfa is de meest gebruikte betrouwbaarheidsindex. Hij had een scherp oog voor wat een betrouwbare observatie was omdat hij het waarnemen van operateurs achter radarschermen bestudeerde. Het onderscheiden van ruis en signaal gaat immers over wat betrouwbare en valide observaties zijn.

Als de som van de covarianties samenvallen met de totale variantie dan zijn de items perfect verbonden. De index rekent het gemiddelde uit van alle mogelijke testhelften van een test. De index is gebaseerd op één test en gegevens over een aparte paralleltest ontbreekt. Het Spearman-Brown idee volgend, is het de ondergrens van de betrouwbaarheid. Er is een alternatief dat dit bezwaar niet kent: Guttmans (1945) lambda2 die in SPSS te vinden is. De waarde van de lambda2 is in de praktijk meestal iets hoger dan de alfa. Een gissing op basis van ervaring: tussen ongeveer .02 en .07. Een berekening met 12 subtests van de Indiase DAT (Differentiële Aanleg Test) *IAM Intelligence-Aptitude-Measurement* bij meer dan 2000 leerlingen tussen 12 en 16 jaar liet kleine verschillen zien: tussen de .02 en .06 (persoonlijke communicatie Dr. Gokhale, Universiteit Pune, februari, 2012). Peterson en Kim (2012) beamen dat de alfa een onderschatting is van de ware betrouwbaarheid. Zij stellen een alternatief voor: een samengestelde betrouwbaarheid, berekend met structurele vergelijkingsmodellen (SEM). Ze analyseerden 2524 paren van alfa's en SEM-schattingen. De alfa was gemiddeld .86 en de SEM-samenstellingen .84. Het verschil is zelfs minder dan onze gissing. De meer dan 60 jaar oude alfa blijft een nuttige index voor betrouwbaarheid.
De alfa wordt een index voor de unidimensionaliteit van een test genoemd. Dat is onjuist want een test waar meer factoren aan ten grondslag liggen kan een hoge alfa hebben als die andere factoren geen verschil maken in de steekproef. De theoretische en technische basis van Cronbachs alfa en andere interne consistentie maten zorgen nog steeds voor publicaties. Wittmann (1988) noemde het de favoriete bezigheid van psychometrici is om een nieuwe coëfficiënt te maken of een oude te nuanceren. Sijtsma (2009) laat zien dat de

Cronbach alfa een onderschatting is en stelt de lambda2 als vervanger voor. Als al de coëfficiënten op hetzelfde neerkomen, waarom dan niet de hoogste gekozen? Dat zou - strikt genomen - tot gevolg hebben dat er geen test-hertest (stabiliteit) coëfficiënten meer bepaald hoeft te worden. Deze zijn feitelijk tussen de .10 en .15 lager dan de interne maten (ervaring met testbeoordelingen en persoonlijke communicatie met Arne Evers, UVA, april, 2009). Iedere diagnosticus weet dat het verstrijken van tijd invloed uitoefent op de cognitie en persoonlijkheid van cliënten. Ieder heeft zo zijn eigen geschiedenis.
De coëfficiënt alfa is het hoogst bij p waarden van rond de .50 want dat levert de hoogste variantie: p x q =.25 op. Waarden > .80 en < .20 dragen weinig bij aan de variantie. Daar moet de alfa het niet van hebben. Toch kan het zo zijn dat we graag items willen hebben, die zuiver aan het einde van de schaal meten bijvoorbeeld bij zelfdoding, criminaliteit, hoge en lage begaafdheid. Als een aantal items door slechts 10% van de steekproef goed beantwoord of beaamd wordt maken ze kans verwijderd te worden. Ze drukken immers de alfa. Het SPSS programma laat het gedrag van de alfa zien bij weglating van achtereenvolgens ieder item.

Sijtsma (2012) werkt de verschillen tussen betrouwbaarheidscoëfficiënten uit door de verschillende theoretische manieren van omgaan met meetfouten bij parallel, hertest en interne consistentie te analyseren. Hij vergeleek Cronbachs alfa, de GLB (Greatest Lower Bond van Bentler & Woodward, 1980) en de lambda2 van Guttman (1945). In zijn toespraak als voorzitter van de internationale club van psychometrici stelt hij eerst vast dat er een kloof is tussen de praktijk van de testconstructeurs en itemschrijvers - bij het Cito 'itembakkers' genoemd - en psychometrici. Als psychometricus behandelt hij eerst een theorema uit de literatuur dat logisch en dus onontkoombaar leidt tot de vaststelling dat de alfa kleiner of gelijk is aan de testscore betrouwbaarheid. Omdat hij het wiskundig bewijs geeft is dat noodzakelijk zo. Het is geen empirisch feit dat ook anders zou kunnen uitpakken. De ondergrens bij de alfa is dus een mathematisch gegeven en geen empirisch feit.
Wat vertelt Cronbachs alfa? Hij toont hoe sterk de items covariëren, homogeen zijn in een steekproef. Als een diagnosticus bijvoorbeeld een waarde van .82 leest in de testhandleiding is hij tevreden want voor zijn doel heeft hij bijvoorbeeld .80 als criterium voor betrouwbaarheid genomen (Nunally en Bernsteins, 1994 vuistregel). Hij denkt niet aan een betrouwbaarheidsinterval rond de alfa, dat wil zeggen rond de ware ρxx. Hoewel dit zelden gedaan wordt is dit geen luxe want de coëfficiënten zijn gevoelig voor kenmerken van het *sample,* bijvoorbeeld het niveau van homogeniteit. De geobserveerde alfa is zo een schatting van de ware alfa.
Sijtsma merkt op dat ten onrechte de indruk gewekt wordt dat de paralleltest, test-hertest en interne consistentie theoretisch verschillende soorten betrouwbaarheid zijn. Alle berusten op tau-equivalentie (Lord & Novick, 1968, p. 50). Theoretisch moeten ze tot dezelfde waarden leiden. Maar tau-equivalentie en parallellisme zijn te strenge eisen voor de praktijk van de scoring van items. De verschillende methoden leveren verschillende

waarden op voor dezelfde test. De diagnosticus ziet in testhandleidingen meestal verschillen tussen test-hertest ten opzichte van interne consistentie, de alfa.
Het beoordelingssysteem voor tests en vragenlijsten van de Cotan stelt *verschillende* eisen aan de twee. Een test-hertest coëfficiënt mag wat lager zijn dan de interne consistentie om als voldoende of goed gekwalificeerd te worden. Figuur 1 geeft een theoretisch voorbeeld van een ware betrouwbaarheid van ρxx = .85 en de ware waarden voor alfa, lambda2 en GLB en voor een parallel test worden afgebeeld. De figuur neemt het theorema als uitgangspunt dat de volgorde moet zijn: alfa ≤ lambda2 ≤ GLB.



Figuur 1 bevat een theoretisch voorbeeld met een ware betrouwbaarheid van .85. De vijf methoden leveren verschillende waarden op. De volgorde is noodzakelijk zo. Er is voor een normale verdeling van de ware coëfficiënten gekozen. (Sijtsma, K. Future of Psychometrics: Ask what Psychometrics can do for Psychology. *Psychometrika, 77,* 1, 4-20; overgenomen na de auteur geïnformeerd, en de formulieren-invul-bureaucratie van het tijdschriftsecretariaat terzijde gelegd te hebben.

Dit *zijstapje* laat zien dat wat theoretisch (mathematisch uitgedrukt en dus geïdealiseerd) gelijk moet zijn het in de praktische uitvoering niet is. Iets verschillende aannamen leiden tot verschillende schattingen voor de ware coëfficiënt. De waarden van de alfa, lambda2 en GLB verschillen feitelijk *nauwelijks*. De test-hertest waarde verschilt wel van de alfa. De diagnosticus heeft naast de drie een schatting van test-hertest nodig. De exercitie laat zien dat het zinvol is om over betrouwbaarheidsintervallen rond een geobserveerde waarde te beschikken. Bij meta-studies wordt een schatting van de ware betrouwbaarheidscoëfficiënt van test-criterium correlaties vermeld. Deze treffen we ook aan bij effectstudies: d-waarden op basis van een aantal empirische studies en bij predictieve validiteitsstudies: r-waarden op basis van vele coëfficiënten. Deze bevatten overigens ook indirect informatie over de betrouwbaarheid: onbetrouwbare tests mogen niet mee doen in een meta-studie of krijgen een gering gewicht.

Veel testhandleidingen benutten de KTT en deze blijft waardevol. Ze was lang dominant en is niet weggedrukt door de IRT (Trafimow, 2013). De IRT is weliswaar eleganter en theoretisch verder uitgewerkt maar voor de predictie maakt het niet veel uit.
Cronbach et al. (1970) hebben een poging gedaan om de herhaalprocedures van de KTT te integreren. De drie worden ondergebracht in de generaliseerbaarheidstheorie: herhaling binnen de test (interne consistentie), over gelegenheden (stabiliteit, test-hertest) en beoordelaars (interbeoordelaarsovereenstemming). In testhandleidingen treft men zelden generaliseerbaarheidscoëfficiënten aan want de bepaling ervan is veel werk. Daarnaast is één coëfficiënt soms genoeg, bijvoorbeeld bij voorspelling over tijd of als men de samenhang tussen de items van een test wil weten en afwijkende wil verwijderen of de overeenstemming tussen beoordelaars van bijvoorbeeld observatiecategorieën.
De KTT maakt duidelijk dat de diagnosticus niet met een somscore of observatie genoegen kan nemen en dat weet hijzelf uiteraard ook. Er kleeft een meet- of waarnemingsfout aan. De testhandleiding levert hem informatie over betrouwbaarheid van de test. Nu kan hij verschillende waarden aantreffen voor dezelfde test. Ze liggen niet ver uit elkaar. Test-hertest waarden zijn doorgaans lager dan de Cronbach alfa, lambda2, GLB of een SEM schatting. De laatste vier verschillen echter te weinig om er in de praktijk veel aandacht aan te besteden. Omdat de waarden gevoelig zijn voor de variantie in de steekproef is het van belang na te gaan of de cliënt past in de steekproef waarin de waarden zijn geschat. Ook bij meta-studies treft de diagnosticus regelmatig aanzienlijk variatie in correlaties en d-waarden aan.
Het idee van de verwachte score is gecanoniseerd in het boek van Lord en Novick (1968). Tot nu toe vermelden de meeste testhandleidingen KTT betrouwbaarheidswaarden: interne consistentie, tussenbeoordelaarsovereenkomst en stabiliteit en veel minder vaak generaliseerbaarheidsindices, lambda2, SEM maten en GLBs. Er is bovendien een testtheorie die veld wint: de Item Respons Theorie (IRT).

Samenvatting en conclusie

Met behulp van de KTT beschrijven we hoe nauwkeurig de ware of verwachte score wordt geschat uit geobserveerde scores. De KTT heeft hiervoor veronderstellingen nodig. De geobserveerde is een som van de ware en de foutenscore. De fouten correleren niet, evenals de fouten en ware scores. Dit maakt het mogelijk de betrouwbaarheid van een test te definiëren als de verhouding tussen totale, geobserveerde variantie van de scores en de ware variantie van de scores. De ware variantie is de ‘gebonden, gemeenschappelijke verklaarde, gedekte, betrouwbare’ variantie bij afname van een paralleltest. Een paralleltest bevat het idee van herhaling. Dit wordt gerealiseerd door tests in twee (equivalentie) of alle mogelijke helften (interne consistentie) te verdelen en de correlatie te berekenen. Ook kan men een test vaker bij eenzelfde steekproef afnemen (stabiliteit). KTT gaat over het schatten van meetfouten en kan op elk soort meting worden toegepast. De theorie is afgerond en geaccepteerd. Lord en Novick hebben de KTT gecanoniseerd in hun boek

Statistical Theories of Mental Test Scores van 1968. De Cronbach alfa, lambda2, GLB en op SEM berustende maten zouden theoretisch gelijke waarden moeten opleveren. Door kleine verschillen in aannames leveren ze verschillende waarden op. Deze zijn meestal te klein voor de diagnosticus om zijn keuze op te baseren. Test-hertest waarden kunnen wel aanzienlijk verschillen van de overige schattingen. Om de bruikbaarheid van een coëfficiënt te kunnen beoordelen is kennis over kenmerken van de steekproef nodig. Homogene steekproeven leiden tot lagere coëfficiënten dan heterogene.
KTT levert betrouwbaarheidscoëfficiënten op van tests. Een integratie van deze coëfficiënten is tot stand gebracht in de generaliseerbaarheidstheorie. Deze treft men zelden in testhandleidingen aan. Er is wel enige informatie over de betrouwbaarheid van complexe diagnostische procedures zoals interviewen, observeren, hypothesen formuleren en de stappen van het diagnostisch proces. Deze moeten per studie bepaald worden omdat ze niet standaard zijn. De psychometricus gaat er gemakkelijk vanuit dat de diagnosticus geen oog heeft voor meetfouten. Dat is naïef. Hij heeft er wel degelijk een idee van. Hij heeft echter geen kwantificering van de meetfout bij iedere cliënt. Daarom kan hij terugvallen op een van de coëfficiënten uit de testhandleidingen. Deze scheren echter iedere cliënt over een kam en dat doet/kan de diagnosticus niet. Over complexe diagnostische procedures is weinig informatie over herhaalbaarheid.

3. Moderne testtheorie

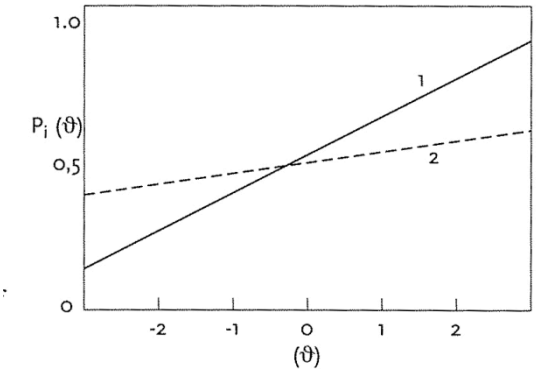
De moderne testtheorie is in eerste instantie niet een model om fouten te schatten. Het is een verzameling functies die een item-responsmodel specificeren. Ze wordt weergegeven met behulp van een itemkarakteristieke functie (IRF). De functie geeft het verband weer tussen de kans op een goed antwoord en de positie van een persoon op een theoretisch attribuut. Het antwoord kan goed/fout gescoord worden of een stelling kan al of niet beaamd worden. Daarnaast worden waarden op een continue schaal gemodelleerd. Het bepalen van de vorm van de functie gaat met behulp van steekproefgegevens en via een computerprogramma wordt de functie bepaald waarbij zowel de subjecten als de items geordend worden langs een unidimensionele schaal. Elke IRF is een minitheorie (item-respons-functie) voor een minigedrag (ja-nee, goed-fout antwoord). Er zijn lineaire en niet-lineaire functies. De laatste worden het meest gebruikt.

Lineair model Dit geeft de relatie weer tussen de positie van een persoon op een latente trek en de kans op een goed antwoord als een lineaire functie. Er wordt verondersteld dat er een ééndimensionele trek ten grondslag ligt aan het antwoord op het item.

$$(13)\ Pi(\theta) = bi + ai\theta$$

Het latent lineair IRT model: Pi (ϑ) geeft de kans p op een goed antwoord weer op een item i door een willekeurig subject met een vaste plaats op de latente variabele ϑ.

Er staan twee parameters in de functie: bi en ai. De rechte lijn in de functie is vastgelegd door de bi (de intercept) en de ai (de helling).



Figuur 2: Lineaire functies van twee items. Item 1 discrimineert beter dan item 2 tussen goede en minder goede studenten omdat de regressie in item 1 steiler is dan in item 2. Voor item 2 is de kans op een goed antwoord ongeveer gelijk voor elke waarde op de latente trekschaal θ.

Parameter bi verwijst naar de moeilijkheid van het item. Voor het goed beantwoorden van een moeilijk item is een hoge positie op de latente trekschaal nodig om het goed te beantwoorden. Item 1 maakt meer onderscheid tussen subjecten met verschillende posities op de latente trek dan item 2. Een item dat even goed beantwoord wordt door zowel personen met een hoge als met een lage positie op de latente trek kan hoge en lage presteerders niet onderscheiden. Het wordt verwijderd want voldoet niet aan de eisen van het model.

Niet-lineaire modellen Een niet-lineaire relatie tussen capaciteit en prestatie is intuïtief aanvaardbaar. Leer- en vergeetcurven laten zien dat wat erbij komt en verdwijnt niet met gelijke stukjes per tijdseenheid plaatsvindt. In het begin gaat het hard met leren en trainen, maar verder op in het proces is het moeilijker en wordt een steeds kleinere winst geboekt. Dit is een realistische interpretatie van de logistische functie. Deze lijkt op de cumulatieve normaalverdeling. In de modellen wordt ervan uit gegaan dat er een ééndimensionele trek ten grondslag ligt aan de antwoorden. In feite gaat het bij het beantwoorden en uitvoeren van welke vraag of taak om meer dan één vaardigheid. Bij het maken van een redactiesom, is rekenen niet het enige. Ook lezen, zelfvertrouwen, aandacht, ervaring met redactiesommen en bekend zijn met de testsituatie spelen een rol. De voorkeur voor een ééndimensionele trek is zinvol. Als je meet moet je één dimensie tegelijk meten anders kun je de uitkomst niet interpreteren. We lezen unidimensioneel hier realistisch als een dominante dimensie. Dat is het geval als verschillen op basis van andere dimensies ontbreken want alle proefpersonen scoren daar gelijk op.
Er is een toets om na te gaan of items unidimensioneel zijn door bepaling van lokaal statistische onafhankelijkheid. Als items dat niet zijn kunnen subjecten met dezelfde vaste positie op de latente schaal verschillende scores op de test hebben. Items zijn lokaal

statistisch onafhankelijk als de kans op een specifiek patroon van antwoorden van subjecten met een vaste positie op de latente gelijk is aan het product van de kansen van de afzonderlijke antwoorden van de subjecten. Stel er is een test met drie items en de persoon heeft een vaste positie op de latente schaal a (θa). Gegeven deze positie heeft hij een kans van .80 (p= .80, q = 1.00- .80 = .20) om item 1 goed te beantwoorden; p = .50 voor item 2 en p = .40 voor item 3. Er zijn 2^3 mogelijk patronen. Onder de aanname van lokale statistische onafhankelijkheid kan men de kans op de patronen uitrekenen (Tabel 1).

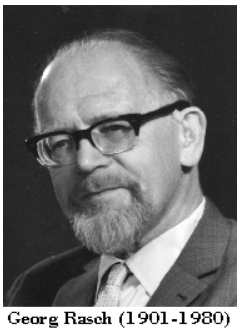
Tabel 1 De kans op 0-1 patronen onder de aanname van lokale statistische onafhankelijkheid van drie items: item 1: p = .80, item 2: p = .50 en item 3: p = .40. De corresponderende q's (1-p) zijn .20, .50 en .60.

patroon 000	q1q2q3 = .2 x .5 x .6 = .06
patroon 100	p1q2q3 = .8 x .5 x .6 = .24
patroon 010	q1p2q3 = .2 x .5 x .6 = .06
patroon 001	q1q2p3 = .2 x .5 x .4 = .04
patroon 110	p1p2q3 = .8 x .5 x .6 = .24
patroon 101	p1q2p3 = .8 x .5 x .4 = .16
patroon 011	q1p2p3 = .2 x .5 x .4 = .04
patroon 111	p1p2p3 = .8 x .5 x .4 = .16

Als de geobserveerde waarden afwijken van de voorspelde dan is er geen steun voor unidimensionaliteit en moet een tweede, derde, enzovoort variabele een rol spelen. Tabel 1 laat het verschil tussen de IRT en de KTT zien. In de KTT doet het er niet toe hoe je aan een score van 1 of 2 komt als je er maar een of twee van de drie goed doet. Bij de IRT wordt het item gewogen op grond van zijn discriminatiewaarde. Lokale statistische onafhankelijkheid houdt niet in dat de items niet gecorreleerd zijn in een steekproef met verschillende trekscores dat wil zeggen een steekproef met variantie. De items zijn niet gecorreleerd als alle subjecten dezelfde waarde hebben op de latente trek: geen variantie en dus geen covariantie en geen correlatie.
Hoe nu het model met woorden weer te geven? Persoon j heeft de kans pi - gegeven zijn vaste positie op trek θ - om het item correct te beantwoorden. In feite doet een persoon het óf goed óf fout dus p = 1.00 of p = .00. Dan moet je het wel zo lezen: de kans om item i goed te beantwoorden, gegeven θ, is de kans dat een willekeurig subject met niveau X op θ item i correct beantwoordt.

Eén en twee parameter logistische modellen Er zijn meer dan vijftig IRT modellen met hun Item-Karakteristieke-Curven (ICCs) of Item-Respons-Functies (IRFs). Een overzicht is te vinden in Hambleton en Swaminathan (1985). Twee bekende zijn het *One Parameter Logistic Model* (OPLM) en het *Two Parameter Logistic Model*. Het eerste is bedacht door de Deense wiskundige Georg Rasch (1960) en het tweede is een uitbreiding. Als er zoveel zijn,

hoe valt er dan te kiezen? Er zijn subtiele verschillen in de vormen van toegestane functies, van een logistische kromme tot monotone stijging. Soms ligt een keuze voor de hand. Bij een meerkeuzevragentest met een kans op goed gokken is een model met een extra parameter voor het gokken op zijn plaats.


Georg Rasch (1901-1980) was een Deense wiskundige. Psychologische tests kwamen hem vaag voor en hij had meer aan het denken van Newton. Hij vernieuwde de testtheorie. IRT modellen heten ook wel Rasch modellen.

Het basismodel is het *One Parameter Logistic Model*: OPLM. Daarbij wordt verondersteld dat het antwoord een functie is van een vaste positie van de geteste op de unidimensionele latente trek. De positie en het antwoord zijn volgens een *logistische regressiefunctie* verbonden. Deze heeft dezelfde vorm als de cumulatieve normaalverdeling. We kunnen de functie als volgt lezen: de kans op succes van een persoon op item i is het product van zijn positie op de latente vaardigheid/trek (weergeven met de Griekse letter thèta (θa) en de moeilijkheidsgraad van het item (bi). De kans van succes op een item i is θa/bi. De kans op succes is hoog bij een persoon met een hoge positie op de latente trek (en bij gemakkelijke items) en laag voor een persoon met een lage positie op de latente trek (en bij moeilijke items). Nu spiegelt volgens Rasch de bi (moeilijkheidsgraad van item i: bi) het niveau van persoon A op de trek (θa). De functie is de logistische en die ziet er zo uit:

(14) f(y) = exp (y)/ 1 + exp (Y)

Exp (y) wil zeggen y vermenigvuldigd met het getal 'e': de natuurlijke logaritme: 2, 71828… Als y zeer hoog is, nadert de functie y, f(y) 1. Als y heel laag is, nadert f(y) 0.

De kans op een goed antwoord wordt in veel formules uitgedrukt als een verhouding tussen de kans dat je iets goed doet: kans Pi, gegeven je vaardigheidsniveau θa, gedeeld door de kans dat je iets fout doet, faalt, gegeven je niveau: 1 – Pi/ gegeven θa).

(15) θa/bi = Pi(θa) / 1 – Pi(θa)

De kans dat persoon a het item i goed beantwoordt of beaamt.

Het kan zo gezegd worden: De kans om een item correct te beantwoorden gegeven het niveau op de vaardigheid a [Pi (θa)] is dat niveau [θa] + bi (de moeilijkheidsgraad van het item). Thèta (θ) is steeds de latente trekschaal, het theoretisch attribuut dat volledig

bepaald is door parameters die een functie beschrijven. Er wordt aangenomen dat een subject een vaste positie op de latente variabele heeft. Een steekproef heeft verschillende posities. Er is variantie, de ware variantie van de steekproef. Als we de moeilijkheidsgraad en het vaardigheidsniveau gelijkstellen (het hangt helemaal van het niveau af hoe moeilijk een item is; dat is geen aparte parameter) om 50% kans te hebben op een goed antwoord dan ziet dat er zo uit:

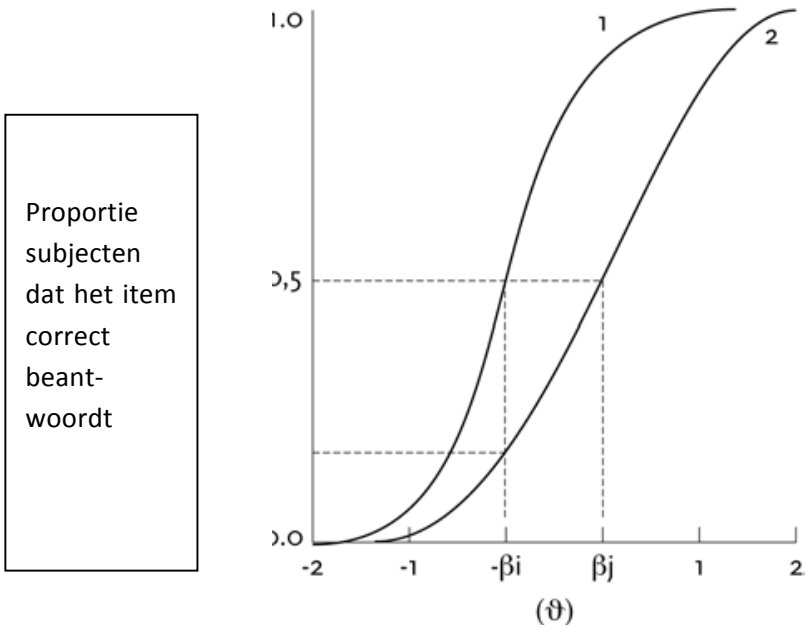
$$(16)\ f_i(b_i) = \frac{\exp(0)}{1+\exp(0)} = \frac{1}{1+1} = 0.50$$

Als een item meer vaardigheid (een hoger niveau op de latente trek) vraagt om een item goed te maken dan is het item moeilijker en volledig gespiegeld: een moeilijker item vraagt een hoger niveau op de latente variabele (zie ook Figuur 4.3).

Het twee-parametermodel voegt een kenmerk toe. Naast de moeilijkheidsgraad (bi) wordt (ai) toegevoegd. Dat is de discriminatieparameter van een item. Deze drukt uit hoe goed een bepaald item de hoge scoorders van de lage scoorders kan onderscheiden. Dat is te zien aan de steilheid (*slope*) van de ICC. Bij het twee-parametermodel hebben de items niet alle dezelfde steilheid maar ze mogen elkaar niet overlappen. In het OPLM voegen de curven als zich als lepeltje-lepeltje, zonder overlap. Een vlakke functie geeft niet veel informatie over de positie op de trek als men dat item goed doet of beaamt. Bij een steile functie is dat het geval (het niet-lineaire model in Figuur 4.3: item 1 is iets steiler dan item 2). Het twee-parametermodel ziet er als volgt uit:

$$(17)\ P_i(\theta) = \exp(a_i(\theta - b_i) / 1 + \exp(a_i(\theta - b_i)$$

De kans dat een willekeurig geselecteerde persoon met vaardigheidsniveau ϑ, item i correct beantwoordt is een functie van zijn niveau op de trek (ϑ), de moeilijkheidsgraad (bi) en de discriminatiewaarde van item i (ai).



Figuur 3: Twee logistische responsfuncties. Ze mogen elkaar niet snijden. Om item 1 goed te beantwoorden bij p = .50 is een lagere positie op θ voldoende dan op 2; item 2 is moeilijker.

Genormaliseerde scores worden gebruikt omdat er goed mee te rekenen valt. De b waarden nemen meestal waarden tussen de + 2 en - 2 aan en de ai waarden lopen meestal tussen 0.00 en 2.00. Hoge waarden op ai leiden tot een steile functie. Dat lijkt op de discriminatiewaarde van een item in de KTT. Die laat zien hoe goed een item bijvoorbeeld de 20% hoogste en 20% laagste scores van elkaar kan scheiden. Er is geen overlap tussen personen, dat wil zeggen op een bepaald item doen alle hoge scoorders het goed en alle lage het fout.

De IRT is bekend sinds de jaren 60 van de 20ste eeuw. Er zijn voorlopers: Thurstones *Case V* voor attitudemeting was er een (zie Edwards, 1957). Nu is het de populaire manier voor het modeleren van items en tests. Cito toetsen voldoen aan de eisen van een Rasch model. De IRT wordt vooral gebruikt voor school- en prestatietoetsen. Het Cito heeft een Rasch-model met *relaxed* assumpties ontwikkeld. Dat verwijst naar het feit dat meer items in het model behouden blijven dan bij het OPLM omdat elke functie niet dezelfde vorm en steilheid hoeft te hebben. Daarna is IRT modeleren op bijna alle persoonskenmerken toegepast. Hier wordt IRT als een minitheorie voor een minigedrag opgevat. Voor atomistisch ingestelde psychologen - als we maar de kleinste bouwstenen van gedrag isoleren, dan komt het goed - is dit een compliment. Je kunt evengoed verdedigen dat grotere eenheden van gedragingen (hogere aggregatiegraden en *Gestalten*) nodig zijn om complexe criteria te voorspellen of ingewikkeld gedrag te beïnvloeden en te beslissen. In sommige tekstboeken is de IRT een onderdeel van itemanalyse (bijvoorbeeld hoofdstuk 6 in Crocker & Algina, 1986; Hogan, 2003).

In de IRT spelen begrippen als latente trek, niet geobserveerde kenmerken, theoretisch attribuut een rol. Het is iets dat we niet direct kunnen waarnemen en het heeft een precieze structuur. Ze zo beschouwd verschillen niet van Piagets onderliggende structuren afgebeeld met wiskundige Kleinse groepen. Klein een wiskundige/logicus heeft die bedacht. En, die zijn nog ingewikkelder dan de basale Rasch structuur en functie. Ze zijn nog moeilijker te operationaliseren, te meten en te toetsen.

Box Bestaat iets, als je het niet kunt zien en nooit zult zien?

Psychologen kijken achter de façade, psychometrici ook. Ze zoeken achter alles iets! Is gedrag niet wat het lijkt? Freud was er een meester in om de donkerte te zoeken: wat steekt er achter ons moreel, religieus en politiek correct gedrag? Hij gebruikte dromen (ze blijven populair, zie Draaisma, 2013), fouten, versprekingen als uitdrukkingen van wat we eigenlijk denken en willen. Nu hebben psychometrici het over *true scores* en *latent traits*. Ze beschrijven met functies onderliggende, onzichtbare dimensies. De ware of verwachte score is virtuele werkelijkheid. Bollen (2002, p. 612) omschrijft de latente variabele zo*: 'A latent or nonrandom variable is a random or nonrandom variable for which there is a realization for at least some observations in a sample'*. Dit lijkt op de

omschrijving van operationaliseren als empirisch realiseren of specificeren. Borsboom et al. (2004) vatten de latente variabele op als een werkelijk bestaande trek, structuur of functie die variatie in gedrag *veroorzaakt*.
De gewoonte om iets te geloven wat je niet ziet kun je afwijzen. Maar het is toch niet handig. Mendel probeerde de paarse kleur van zijn erwten te verklaren door aan te nemen dat er dominante en recessieve genen bestonden. Hij zag ze niet maar het was achteraf een goede gedachte. We hebben het over een zwaartepunt dat we niet zien maar het verklaart wel gedrag van objecten. We zien niet direct elektromagnetische velden maar ze verklaren het gedrag van elektrische stroom. Het is wetenschappelijk gezien geoorloofd iets aan te nemen dat je niet ziet. Tegelijk is er de plicht verschijnselen zichtbaar te maken (gas, straling, donkere energie) bijvoorbeeld door waarneembare of meetbare gevolgen te voorspellen en te toetsen. Dit geldt ook voor gedragingen zoals Freuds onderbewuste, latente variabelen en sociale structuren. Klassieke en moderne testtheorie springen weliswaar in het duister maar specificeren dat wat niet te zien is met behulp van een functie. Dat wordt getoetst door antwoorden op items to confronteren met het model.
De gewoonte te geloven wat je niet ziet heeft een keerzijde. Dat weten schrijvers en dichters goed uit te leggen: George Orwell: '*...it is the obligation of any intellectual to observe the obvious, and to prick through the stinking small dogmas which endanger our soul'*. De dichter T.S. Eliot: *'Humans can bear a few realities, and easily accept the non-real'*. Het blijft uitkijken!

Voor de diagnosticus is een passend model een garantie dat de items aan modeleisen voldoen. Ze vormen een ééndimensionele latente trek. Hij wel kan ontevreden zijn over de inhoud van de items omdat wat hij als relevant gedrag ziet niet blijkt te passen in het model. De psychometricus zet hem toch aan het werk om nieuwe items te schrijven. De testinformatiefunctie vertelt over de kwaliteit van de items. De IRT geeft informatie over de betrouwbaarheid van de items en over wat de test meet, in de zin van het passen in een model. De items worden ook niet zonder meer opgeteld voor de totaalscore. Ze krijgen een gewicht dat afhangt van hun onderscheidend vermogen. De somscore bij een IRT test is een *gewogen* score. De interpretatie van de latente trek of het theoretisch attribuut wordt vaak aan de theoreticus en diagnosticus overgelaten. Bij de constructie van studietoetsen wordt een interpretatie omzeild door de inhoud van het curriculum als uitgangspunt te nemen. Men meet bijvoorbeeld kennis van de Engelse grammatica op havo 4-niveau aan de hand van een aantal verschillende toetsen. IRT-statistieken (testinformatie functies) zijn voor de diagnosticus lastiger te lezen dan betrouwbaarheidscoëfficiënten met waarden tussen 0 en 1. In de testhandleiding moet de informatiefunctie uitgelegd en vorm en steilheid geïnterpreteerd worden.

Samenvatting en conclusie

De moderne testtheorie bevat meer dan vijftig modellen met corresponderende IRFs en ICCs. Het is een minitheorie (een functie uit vele mogelijke) over een minigedrag (antwoord op een item). De functie specificeert de relatie tussen de positie van een persoon op de latente trek en de kans dat hij een item correct beantwoordt of beaamt. Niet-lineaire functies worden het meest gebruikt. En dat is ervan gekomen. Er is geen bewijs dat de

lineaire niet zouden voldoen. Het basismodel is het OPLM van Rasch. Daarvoor is kennis van de positie van de persoon op de latente trek (θ) genoeg om de kans te schatten dat hij het item correct beantwoordt. De positie op de latente trek van de geteste bepaalt in het model meteen de moeilijkheidsgraad van het item (bi). Het twee-parametermodel gaat over de positie van de persoon op een trek (θ), de moeilijkheidsgraad (bi) en de discriminatiewaarde (ai) van het item: het vermogen hoge van lage scoorders te onderscheiden. IRT modellen toetsen of de items passen in een unidimensionele schaal. De functies kunnen er verschillend uitzien wat betreft steilheid, wel of geen overlap strikte en *relaxed* assumpties. Hoe meer je toestaat, hoe meer items in he model passen. Er is informatie over de kwaliteit van de items: de ai en bi. De somscore op de test komt tot stand door items te wegen naar hun onderscheidend vermogen. De IRT wordt vooral gebruikt bij het ontwerpen van prestatietoetsen. Ze wint veld bij de meting van persoonskenmerken. De diagnosticus kan zijn voordeel doen met de IRT modellen. Hij weet dat een IRT model berust op een onderliggende unidimensionele schaal die inhoudelijk geïnterpreteerd moet worden. Hij krijgt informatie over de items: moeilijkheidsgraad en onderscheidend vermogen door middel van iteminformatie functies.

4. Relatie KTT en IRT

Beide theorieën geven informatie over kenmerken van items en tests. De relatie tussen KTT verschilt tussen auteurs. Embretson en Reise (2000) benadrukken overeenkomst. Statistieken van de KTT corresponderen met itemparameters van de IRT. Het percentage getesten dat het item goed beantwoordt is te vergelijken met de bi's van de items van de IRFs. Itemtotaal correlaties in de KTT zijn vergelijkbaar met de ai's van de IRT. Er is meer nadruk op verschillen. Er is kritiek op de KTT.

Bezwaren van IRT tegen KTT In de KTT krijgt elk item hetzelfde *gewicht.* Er wordt geen rekening gehouden met hun discriminatiewaarde (ai). In de KTT wordt verondersteld dat elk item voor elk ander item ingeruild kan worden. In de IRT is de somscore een gewogen score. Men kan niet alle mogelijk manieren een somscore van 20 krijgen uit 40 goed-fout items. Bovendien toetsen IRT-modellen of er één dimensie aan de test ten grondslag ligt. In de KTT wordt dat vooronderstelt dat de items over een vaardigheid of kenmerk, bijvoorbeeld numerieke vaardigheid of gewetensvolheid. Verder is de KTT betrouwbaarheidscoëfficiënt gevoelig voor de variantie van de steekproef. Is de variantie beperkt, bijvoorbeeld in een homogene steekproef dan is de betrouwbaarheid navenant lager. Dit volgt uit de definitie van betrouwbaarheid als de verhouding tussen geobserveerde en ware variantie. Dit betekent ook dat er niet één betrouwbaarheidscoëfficiënt is. Deze is gecontextualiseerd door de steekproeven. Crocker en Algina (1986) vermelden een methode om te corrigeren voor de homogeniteit van de steekproef maar dat is een *deus ex machina*. Interne consistentie coëfficiënten zijn evenzeer gevoelig voor de variantie van de items. Moeilijke of gemakkelijke items hebben geringe variantie en dus is er weinig covariantie en dat leidt op

zijn beurt tot lage interne consistentie. Cronbachs alfa wordt naar verhouding lager als er heel moeilijke of heel gemakkelijke items in de test zijn. De variantie van die items is gering. De maximale variantie = 0.25; p × q = 0.50 × 0.50 = 0.25 bij items die de helft van de proefpersonen goed beantwoorden. De activiteit om te nuanceren in de hoge en lage regionen van een test levert een lage interne consistentie op. Dat kan niet de bedoeling zijn.
Om de betekenis van een KTT-score van een persoon vast te stellen moet zijn score vergeleken worden met een steekproef. Een score van 20 is hoog als 10% deze waarde behaalt maar laag als 90% die behaalt. Normeren is tijdrovend want er moet een representatieve steekproef getrokken worden. Men moet een referendum houden om de verkregen score van een geteste persoon te interpreteren. De IRT heeft de oplossing omdat de itemparameters los van de steekproef geschat worden (zie Hambleton, 1989, p. 151-152). Bovendien kunnen met behulp van IRT prestaties vergeleken worden los van de specifieke test. Als dat met KTT gedaan wordt moet eerst uitgezocht worden of de tests equivalent zijn. Als ze dat niet zijn moeten ze gelijk gemaakt worden. Sommige IRT modellen pretenderen de positie van de persoon (zijn score) op de latente trek te bepalen los van de moeilijkheid van het item. Ook twee tests die dezelfde trek meten maar die verschillen in moeilijkheidsgraad kunnen met IRT vergeleken worden.
Nog een punt: in de KTT kan zich regressie naar het gemiddelde voordoen. Dit is een andere manier om te zeggen dat de test niet perfect betrouwbaar is. Dat heeft gevolgen voor de extreme scores. Als een score fluctueert van meting tot meting dan zullen de hoogst scorenden bij een tweede afname een lagere score hebben en vice versa. De verandering kan een artefact zijn. Er is altijd een controlegroep nodig om te toetsen of er een verschil is.

Basis van KTT - IRT geschil Ten eerste ontbreekt een meetmodel in KTT. Er wordt meestal een intervalschaal voorondersteld. Het is *measurement by fiat* volgens Torgerson (1958). De meeteigenschappen van de schaal worden in de IRT getoetst met behulp van gespecificeerde IRFs. Ten tweede is de standaardmeetfout in de KTT gelijk bij elke score, hoog en midden en laag. De item-informatiefunctie van de IRT geeft elk item zijn eigen meetfout. Het is moeilijk is om op elk niveau even goede items dat wil zeggen met voldoende discriminerende waarden te schrijven. Nuanceren in het midden van een schaal is gemakkelijker dan aan de uiteinden. De vorm van de IRFs geeft daar genuanceerde informatie over. En, als je items wil schrijven die de hele schaal bestrijken is de alfa van Cronbach niet geschikt. Iems met lage covarianties resulteren in lage alfa's. Een dergelijke resultaat wordt gelezen als onbetrouwbaar of onvoldoende intern consistent.
De verschillen worden door psychometrici bestempeld als zwakten van de KTT. Hun argumenten snijden hout. Als de twee vergeleken worden met het oog op de predictieve validiteit maakt het echter niet veel uit. IRT-tests en -vragenlijsten voorspellen criteriumgedrag niet beter dan KTT-instrumenten. IRT items zijn gewogen en die uit de KTT krijgen elk hetzelfde gewicht. De correlaties tussen de IRT-gewogen en KTT-ongewogen

versies van dezelfde tests blijken zeer hoog. Zo kan een diagnosticus zeggen: *much ado about almost nothing.*

Boodschap van KTT-IRT vergelijking De IRT produceert een informatiefunctie die in staat stelt de betrouwbaarheid van items op verschillend niveau van de latente schaal te schatten. De items aan de uiteinden van de schalen zijn minder betrouwbaar dan die in het midden. Dit is mooi maar de diagnosticus heeft geen begrijpelijk getal zoals in de KTT. Hij moet de item-karakteristieke-functie en de item- en testinformatiefunctie interpreteren. Daar zijn wel vuistregels voor, maar de diagnosticus kan niet verwachten dat IRT tests veel winst opleveren bij het voorspellen van criteria. De IRT en KTT versies van dezelfde test correleren hoog: > .90. IRT zorgt ervoor dat er één schaal is. Dat is zinvol voor de constructvaliditeit. Itemschrijvers zullen gemakkelijk meer capaciteiten of dimensies aanspreken zonder er direct bij stil te staan. Het model zal leiden tot verwijdering van items. Intussen blijft de itemschrijver bij zijn standpunt dat het verworpen item wel degelijk bij de inhoud van bijvoorbeeld Engels op Havo IV hoort. Doorgaans wordt dat geen debat want de model-ongevallige items gaan er uit. De IRT biedt de mogelijkheid items los van groep, bijvoorbeeld m/v, leeftijden en etnische groepen te beschouwen. Ze kan laten zien of items in verschillende groepen iets anders betekenen door studie van *Differential Item Functioning* (DIF).
IRT levert - vergeleken met de KTT - iets op voor (a) constructvaliditeit: op zijn minst discussie/debat tussen itemschrijvers en psychometrici over de kwaliteit van de afzonderlijke items (testinformatiefunctie maar die is niet meteen te lezen) en (b) voor het vergelijken van groepen (Itempartijdigheid: DIF). Zo zeggen psychometrici dat IRT alles kan wat KTT kan maar verantwoorder en eleganter is. DIF is winst want er kan mee aangetoond worden dat er zich bij een test itempartijdigheid voordoet. Een voorbeeld: personen die gelijk scoren op het criterium scoren verschillend op de predictoren op basis van de itemkeuze. Voor predictieve validiteit maakt het niet uit of men IRT- of KTT-geleide instrumenten gebruikt.

Implementatie van KTT en IRT Beide minitheorieën worden volop gebruikt. De IRT wint terrein. In Nederland worden (tot nu toe, want Europese regelgeving eist dat er meer testaanbieders moeten/mogen zijn) alle onderwijstoetsen centraal gemaakt. Op het Cito geldt een daar ontworpen IRT model. Persoonlijkheidsvragenlijsten hebben het lang met de KTT gedaan maar dat gaat veranderen. Reise en Waller (2009) merken op dat ...*IRT methods have emigrated to clinical measurement* en ze een *…deeper appreciation of cognitive and clinical constructs* laten zien. Een nogal semantische en begrip verhelderende opvatting van de IRT en de IRFs. Het zijn curven en functies. Ze stellen IQ en cognitief testen min of meer gelijk met testen van grote groepen terwijl men bij klinisch testen te maken heeft met kleine steekproeven, gemengde en gevarieerde groepen patiënten, scheve scoreverdelingen, matig omschreven domeinen, smalle band constructen en hoge correlaties tussen de

trekken die te danken zijn aan de *omnipresent negative affectivity dimension*. Deze opmerkingen zullen ertoe leiden vele van hun items uit *poorly articulated domains* niet in IRT modellen passen. De KTT heeft overigens ook enkele statistieken die items uit de verzameling stoten. De opmerkingen lijken vooralsnog lipservice aan IRT. De auteurs leggen de nadruk op processen die tot de antwoorden leiden en op pathologische persoonskenmerken. De IRT zal ook terrein winnen bij selectie en loopbaanbegeleiding. Er zullen ook IRT tests voor kwaliteit van leven, pijnbeleving, bewegen, geheugentraining, enzovoort komen.

En de practicus en diagnosticus? Zij zijn vertrouwd met betrouwbaarheids- en validiteitscoëfficiënten. Ze mogen van psychometrici en testconstructeurs verwachten dat ze een goed product bieden. Hoewel diagnostici op de hoogte zijn van de superioriteit van de IRT beïnvloedt dat de praktijk nauwelijks. Het gebruik van IRT raakt overigens wel leraren die Cito-eindtoetsen afnemen en leerlingvolgsystemen bijhouden. Heeft nu de introductie van de IRT de kwaliteit van tests verhoogd? Ondanks alle tegenwerpingen in de pers en uit het onderwijs heeft het Cito bijgedragen aan psychometrisch verantwoorde tests. Omdat ze op grote schaal afgenomen zijn heeft het ons veel geleerd toetsontwikkeling. Dit zal het commentaar niet stoppen want het raakt ons immers als er over onze capaciteiten en die van onze kinderen geoordeeld wordt, denk aan de *well to do* Havo-is-geen optie-ouders.
Evers et al. (2010) hebben laten zien dat de kwaliteit van tests in Nederland in 30 jaar volgens de Cotan vereisten nauwelijks verbeterd is. Voor kwaliteitsbevordering zijn een professioneel instituut en goed aantal omschreven domeinen nodig. Volgens sommigen doen de theoretici en professionals dat onvoldoende. Sijtsma (2012) beweert dat er te winnen valt door betere theorie: *much terrain remains to be conquered* (p. 5). De psychometrici zijn er met hun modellen en functies al klaar voor.

Samenvatting en conclusie

De KTT biedt informatie over de betrouwbaarheid van tests en vragenlijsten. Er zijn verschillende indices die op hetzelfde idee berusten. Ze leiden wel tot verschillende waarden. De parallelcoëfficiënt, de alfa van Cronbach, de lambda2 van Guttman en de GLB van Woodward en Bentler verschillen weinig: tussen .01 en .05 is mijn schatting. De test-hertest waarden liggen meestal lager dan deze vier: tussen .05 en .15. Deze schatting berust op ervaring met testbeoordelingen en collega's. De diagnosticus kan kiezen tussen de eerste vier als het gaat om de samenhang tussen de items maar als hij op termijn moet voorspellen is de test-hertest waarde adequater. De coëfficiënten zijn gevoelig voor de variantie in de steekproef. Er is dus niet één coëfficiënt voor een test of vragenlijst.
Een enkele auteur legt de nadruk op de vergelijkbaarheid en gelijkwaardigheid van de KTT en IRT modellen. De meeste benadrukken verschillen en bekritiseren de KTT. Er wordt geen rekening gehouden met de discriminatiewaarde van de items, betrouwbaarheid hangt af van de variantie van de steekproef, er is een omslachtige normering nodig om de betekenis

van de score te bepalen en wordt een intervalschaal is voorondersteld. De voorspellende waarde van KTT en IRT tests verschilt echter nauwelijks. IRT en KTT versies van dezelfde test correleren hoog zodat ze inwisselbaar zijn. Een diagnosticus kan zeggen dat in dat opzicht de IRT niet *het verschil maakt*. De IRT heeft als voordeel dat er per item bepaald kan worden hoe nauwkeurig die de latente trek en het niveau meet. Bovendien wordt bepaald of items hetzelfde betekenen in verschillende groepen door studie van DIF.

5. Reflectie en evaluatie

Testtheorie gaat over het scoren van items. Items zijn psychometrisch te beschrijven met behulp van statistieken. De psychometrische theorie is (nog) niet op subjectieve procedures toegepast. Testtheorie heeft betrekking op individuele verschillen op unidimensionele latente trekken/ theoretische attributen. Ze is minder benut voor het afbeelden van ontwikkeling van gedrag en het meten van (effecten van) de sociale context. Het bepalen van itemmoeilijkheidsgraden, item-test correlaties, betrouwbaarheidsindices en IRFs leidt tot uitspraken over de betrouwbaarheid en unidimensionaliteit van een instrument. Ze staan los van andere modellen die vragen over items en tests kunnen beantwoorden, zoals factoranalyse en technieken voor attitude schaling.

Diagnosticeren is een n = 1 studie. De standaardmeetfout van de KTT is een populatie/ steekproefkenmerk. De veronderstelling van een persoon-specifieke meetfout is wellicht zinvol (Baird et al., 2006). Wordt de stap van de steekproef naar de specifieke cliënt niet wat snel gezet? Cliënten verschillen niet alleen in ware scores maar ook in de intervallen rond hun ware scores. De kritiek vanuit de IRT op de KTT leidt tot verfijning van de bepaling van betrouwbaarheid en een analyse van constructvaliditeit. Predictieve validiteit overheerst in de diagnostiek en de IRT draagt niet bij aan een toename van die validiteit.

Klassieke en moderne testtheorie worden *beide* bekritiseerd omdat zij statistische formules en functies leveren voor items en tests. Ze bevatten geen theorie over gedrag (Levy, 1974; Guttman, 1978; Goldstein & Wood, 1989; Blinkhorn, 1997; Sijtsma, 2012; Wilson, 2013). Ze zeggen dat de testtheorieën

gaan over abstract meten en inhoud veronachtzamen zoals in Guttmans facetanalyse wordt gedaan,
gericht zijn op itemanalyse zonder na te gaan waaróm een persoon een item moeilijk of gemakkelijk vindt,
geen theorieën zijn, maar een verzameling functies en
een voorkeur bevatten voor testen van grote groepen te testen, bijvoorbeeld in onderwijs en bedrijfsleven. Ze houden geen rekening met het werk van clinici die individuele cliënten onderzoeken. Er is geen lokale statistische onafhankelijkheid bij klinisch testgebruik. Ze gaan voorbij aan de zin van inhoudelijke theorievorming of zeggen dat die gebrekkig is.

De opmerkingen wijzen in dezelfde richting: KTT en IRT zijn statistische modellen en bevatten geen inhoudelijke theorie. De verhouding tussen theorie en meten is in de psychologie een blijvend heet hangijzer. Hoe moet je bijvoorbeeld de G-factor van

intelligentie interpreteren? Als een latente trek die individuele verschillen veroorzaakt in wat we intelligentie noemen? Hoe komen we aan een algemene factor? Is dat een evolutionair uitgeselecteerd geheel van gedragingen, berust deze op snelheid van geleiding in ons zenuwstelsel, op biochemische processen of is het de uitkomst van een statistische techniek, factoranalyse dus, die meestal een sterke eerste factor aanboort? Of nog anders, is het een maatschappelijk gecreëerd verschijnsel om groepen te onderscheiden en binnen groepen verschillen te maken met het oog op een belang?

Voor de diagnosticus levert de testtheorie kwaliteitsindicatoren van instrumenten op. De coëfficiënten van de KTT spreken voor zich. De IRT parameters moeten in de handleiding uitgelegd worden. In de *Standards for Educational en Psychological Tests* van de APA van 1999 en de voorlopige nieuwe editie van 2014 spelen beide testtheorieën een rol, waarbij de IRT terrein wint Ze kunnen helpen om bij meer tests die te kiezen met de beste psychometrische kwaliteiten. Vragen over de enig juiste interpretatie van een latente trek of theoretisch attribuut blijven onbeantwoord. Pragmatici vermijden dit soort vragen en werken met semantische interpretatie van een reeks items als een meting van bijvoorbeeld algemene intelligentie en numerieke vaardigheid.

Testtheorieën bevatten eisen waaraan een diagnosticus moet voldoen. Ze zijn vooral beperkt tot test- en vragenlijstgebruik. Diagnostici houden zich meer en meer aan de eisen die aan de KTT ontleend zijn. Testhandleidingen verstrekken informatie over betrouwbaarheid en validiteit. De Cotan gebruikt een uitvoerig beoordelingssysteem dat bestaat uit zeven rubrieken en ongeveer 70 te scoren test- en itemkenmerken. Het berust vooral op de KTT. De IRT zal terrein gaan winnen al zijn haar statistieken en item- en testkenmerken moeilijker te interpreteren. De winst van de IRT is voor de diagnosticus vooralsnog beperkt. Dit neemt niet weg dat psychometrici kritiek uitoefenen op het testgebruik. Het is suboptimaal gegeven de nieuwe ontwikkelingen in de testtheorie. Het construeren van tests en vragenlijsten komt echter steeds meer in handen van gespecialiseerde instituten, denk aan het Cito dat een monopolypositie heeft op grond van omvang, breedte en expertise. De diagnosticus wordt testafnemer, consument die goede waar kan kopen. De kou is zo uit de lucht. De psychometricus heeft de diagnosticus niet veel meer te verwijten. Als consument ligt hij niet meer onder vuur.

De psychometricus werpt de diagnosticus regelmatig voor de voeten dat zijn semantische theorieën te vaag zijn. Hij kan zijn expertise niet kan toepassen om ze te toetsen. Dat is voor een deel een terecht verwijt. Begrippen kunnen altijd scherper. Voor een belangrijker deel is het een fundamenteel verschil over hoe je kennis over een cliënt opbouwt. De psychometrie is formeel en atomistisch. Het item is de welbewust gekozen kleinste eenheid en het ja-nee antwoord het te bestuderen gedrag. De diagnosticus werkt met gedragsbegrippen die dichter bij het alledaagse taalgebruik liggen. Die begrippen vormen *Gestalten* en geen atomistische kleine eenheden. Daarmee brengt hij individuele verschillen, ontwikkeling en de sociale context van het denken, voelen en doen van de cliënt in kaart. Dat zijn geen itemparameters. Door gericht te zijn op die kleine eenheden is er vanuit de psychometrie

weinig aandacht voor de diagnose als geheel en het diagnostisch proces. Dat zijn *Gestalten* waar psychometrici zich doorgaans niet mee bezighouden. Zo beschouwd liggen ze zover uiteen dat er van onder vuur liggen weinig sprake hoeft te zijn.

Onderwerpen en namen Hoofdstuk IV

Objectief scoren van items en tests
Klassieke testtheorie (KTT/)
Moderne testtheorie (IRT)
Ware score, error score
Populatie
Steekproef
Meetfouten
    systematische fouten
    toevalsfouten
Ware waarde (score),
Verwachte waarde (score)
Paralleltest
Betrouwbaarheid
    parallel
    equivalentie
    stabiliteit
    interne consistentie
    test-hertest
Betrouwbaarheidsinterval
Cronbachs alfa (α)
Spearman-Brown verlengingsformule
Testvariantie, itemvariantie
Wegen van items in IRT
Latente trek of theoretisch attribuut
Itemkarakteristieke functie (ICC)
Item respons functie (IRF)
Logistische functie
One Parameter Logistic Model (OPLM)
Moeilijkheidsgraad en discriminatiewaarde van een item
Relatie tussen KTT en IRT
Testnormen

V Kritiek op diagnosticeren door wetenschappers, professionals en het publiek

De diagnose en het diagnostisch proces liggen regelmatig onder vuur. Vraag niet om medelijden, maar het lijkt op wat de psalmdichter (no. 3, vers 1) zegt: 'Heer, hoe talrijk zijn mijn belagers, velen vallen mij aan, velen zeggen van mij: God zal hem niet redden'. En in psalm no. 22 vers 5: 'Een troep stieren staat om mij heen, buffels van Basam omsingelen mij, roofzuchtige, brullende leeuwen sperren hun muil naar mij open'. Belagers zijn van alle tijden en in dit geval statistische predictie, diagnose-behandel-combinatie aanhangers, betrouwbaarheid en validiteit en testtheorie. Ze zijn hiervoor besproken en gerelativeerd. Behalve de kritiek van professionals (DBCs) komen ze voort uit het beoefenen van wetenschap zelf. Dat is volgens sommigen een gesloten en gefragmenteerd bedrijf. Tillie een UVA hoogleraar Electorale Politiek zegt in de NRC: 'Wetenschappelijke kwaliteit is het publiceren in een Engelstalig tijdschrift dat bijna niemand leest'. 'Het gemiddeld aantal lezers per artikel is ongeveer drie en het biedt geen inzicht, maar informatie'. De interne kritiek is recent aangezwollen en gaat van een kritische blik op theorievorming, operationalisatie, meten en tests tot de replicatiecrisis en fraude. Naast juichende stukken in de pers over wat de medische, exacte en soms zelfs historische wetenschappen vermogen, kritiseren polemisten graag de sociale wetenschappen. Filosofie en vooral de levensleren komen er nog beter van af. Het publiek, de consument is mondig en mengt zich in ook de discussie. Vooral testen en toetsen liggen onder vuur.

1. Interne kritiek

De theorievorming is niet scherp onder meer door vage gedragsbegrippen en zwakke relaties met andere door de bescheiden voorspellende waarde van de scores. Er is geen wet in de zin van Newton te vinden en die zullen er ook niet komen. De operationalisatie van de begrippen, constructen zijn nooit dekkend. Kernconstructen formeel afbeelden is een vlucht naar voren. De afbeelding in een formeel systeem moet immers semantisch geïnterpreteerd worden. Meten en theorie staan op gespannen voet en volgens bijvoorbeeld Trendler is meten onmogelijk. Zij die verzoening voorstellen (bijvoorbeeld Guttman) hebben geen invloed gehad. Er wordt gezegd dat er geen theorietoetsing plaats vindt en zelfs kan vinden. Onderzoek verstrekt slechts empirische informatie die geldt voor een steekproef in een specifieke situatie. De kritiek beperkt zich in de praktijk op het instrumentarium door zowel professionals als leken en consumentenorganisaties.

Kritiek op tests en vragenlijsten Naar verhouding wordt de meeste kritiek gegeven op het instrumentarium. Tests en vragenlijsten staan centraal in de diagnostiek en daarop is de meeste kritiek uitgeoefend. Eerder heb ik opgemerkt dat de complexe procedures minder gekritiseerd worden omdat empirisch onderzoek daar tijdrovend en lastig is. Tests worden in Nederland beoordeeld met behulp van een uitvoerig systeem dat bestaat uit zeven rubrieken. Het is een hybride van eisen aan theorie (theoretische uitgangspunten), uitvoering van testmateriaal, handleiding, normering, betrouwbaarheid en begrips- en criteriumvaliditeit. Een scoringssysteem met ongeveer 70 vragen stelt in staat elke test, vragenlijst of procedure een keurmerk te geven. De overheid neemt dit oordeel serieus en vraagt een voldoende waardering op elke rubriek alvorens een instrument toe te laten voor gebruik bijvoorbeeld in het onderwijs . Evers et al. (2007) hebben de resultaten van de beoordeling op meer dan 500 tests en vragenlijsten bij elkaar gezet. De waardering per rubriek is ‘goed’, ‘voldoende’, ‘onvoldoende’. Van de meer dan 500 instrumenten waren de theoretische uitgangspunten in 62% goed, 25% voldoende en 13% onvoldoende. Dit is een hoge score gegeven de opmerkingen over de zwakten van psychologische theorievorming en het aantal ‘wetten’ die ze oplevert. De criteria zijn kennelijk niet streng. Het testmateriaal van de instrumenten wordt in 72% goed, 21% voldoende en 8% onvoldoende beoordeeld. Dit is een gevolg van het bekend zijn van de eisen aan het materiaal en testuitgevers kunnen ermee uit de voeten. De handleiding is in 49% goed, in 28% voldoende en in 23% onvoldoende. Voor de handleiding is het format bekend. De meeste testconstructeurs proberen zich daar zoveel mogelijk aan te houden. Normering is in 13% goed, in 29% voldoende en in 58% onvoldoende. Dat is een matige score. Het weerspiegelt de hoge eisen die aan normering gesteld worden. Het is misschien goed om lokale normen te waarderen en niet meer landelijk representatieve steekproeven te eisen waar iedere provincie, iedere de sociaal economische klasse representatief vertegenwoordigd zijn. Betrouwbaarheid met een onderscheid tussen interne consistentie en test-hertest maten levert in 28% een goed, in 40% een voldoende en in 32% een onvoldoende resultaat.

Begripsvaliditeit is goed in 19%, voldoende in 46% en onvoldoende in 35% van de beoordeelde instrumenten. Criteriumvaliditeit is goed in 8%, voldoende in 22% en onvoldoende in 56%. Bij 15% was het criterium niet van toepassing. Dit zijn bescheiden resultaten.

Kritiek op de DSM De DSM-IV TR en DSM-5 zijn het bekendste psychiatrisch classificatiesysteem. Het ligt steeds weer onder vuur. Batstra en Thoutenhoofd (2013) hebben de geschiedenis van de DSM op een rij gezet: De DSM-I (1952) en de DSM-II (1968) bevatten 106 en 182 stoornissen. Deze waren niet alle beschreven met een afzonderlijke etiologie. De DSM-III (1980) en de DSM-III-R (1987) vermeldden 265 en 292 stoornissen. De laatste twee werden herzien om de betrouwbaarheid van de categorietoekenning te vergroten. De derde versie bevatte criteria die ervoor zorgden dat een persoon gemakkelijker aan een categorie kon worden toegewezen. Het aantal afwijkingen nam met 32% toe in de jaren 80 en met 48% in de jaren 90 (Kessler et al., 2005). In 1994 werd de DSM-IV gepubliceerd en deze zorgde voor een toename van diagnoses van autisme en ADHD bij kinderen (CDC, 2012). De tijdgeest zorgde ervoor dat er aandacht besteed werd aan neurobiologische oorzaken van stoornissen. Kupfer et al. rapporteerden al in 2002 dat er geen duidelijke neurobiologische oorzaken aangewezen konden worden voor de stoornissen. Kupfer werd een van de voorzitters van een groep experts die de DSM-5 uitbracht. Spitzer, de voorzitter van de DSM-III Task Force vertelde in 2007 in een BBC televisieserie dat zijn werk de medische kijk op stoornissen had versterkt. Frances zei hetzelfde over de DSM-5 in de Los Angeles Times van 7 maart, 2010. De toename van het aantal stoornissen vloeide *niet* voort uit nieuwe theoretische inzichten. Het was ook niet het resultaat uit toetsend empirisch onderzoek. De oorzaak van de toename was niet wetenschappelijke vooruitgang, maar marketing: de verkooptechnieken van farmaceutische industrieën. Psychiaters hielden lezingen voor geïnteresseerd publiek in exotische oorden. De industrie sponsorde medische opleidingen en patiëntenverenigingen, adverteerde in wetenschappelijke tijdschriften en legde contact met DSM-experts (*ISM Health Reports*). De pogingen hadden succes. Er werd voor ongeveer 25 miljard US dollars aan medicijnen tegen psychosen en depressies verkocht. De syndromen van de DSM kregen de status van biomedische entiteiten en sommige cliënten dwongen de behandelaars om hen de medicijnen voor te schrijven. De diagnoses zijn immers soms een opluchting voor cliënten, een erkenning van hun lijden. Dit maakte de weg vrij om zich tegen aandoeningen in de DSM te verzekeren. Volgens Batstra en Thoutenhoofd zal de DSM-5 niet zorgen voor een afname van stoornisdiagnoses. Integendeel, er worden weer nieuwe stoornissen toegevoegd, bijvoorbeeld *disruptive mood dysregulation, hoarding disorder* (kan alleen als men veel spullen heeft; is onwaarschijnlijk bij armoede), *minor cognitive disorder, bing eating disorder, skin picking disorder'* om enkele te noemen. Ook onderzoekers verwachten een toename in de diagnoses van stoornissen (Andrews & Hobbs, 2010; Mewton, 2011; Timpano et al. (2011). De auteurs concluderen bovendien dat er veel valse positieven en

negatieven met de DSM-5 zullen voorkomen omdat de epidemiologische studies vooral door kort opgeleide leken-interviewers gedaan worden en niet door ervaren psychiaters, klinisch psychologen en pedagogen. Als gevolg worden cliënten behandeld die een *minor disturbance* of voorbijgaande afwijking vertonen. En, sommigen die wel behandeling nodig hebben vanwege de ernst van de stoornis, worden niet als zodanig gediagnosticeerd. Ze twijfelen ook aan de wetenschappelijke kwaliteit. Er zijn te veel economische en professionele belangen, zodat onafhankelijk onderzoek schaars is.
De DSM-5 kan opgevat worden als een eerste stap, als een screeningsinstrument dat gevolgd wordt door vragenlijsten voor specifieke stoornissen, zoals psychopathie en narcisme. Er zijn verschillende psychopathie vragenlijsten. Walters et al. (2011) vroegen zich af of samenstellingen van factorscores van de *Psychopathy Checklist* (gewelddadig) recidivisme even goed voorspelde als de afzonderlijke scores. En, niet zo verrassend, bleek, dat het verschil klein was. Omdat het toekennen van categorieën al veel werk is, zal het systeem niet als een eerste screening ingezet worden.

De kritiek op het instrumentarium is weliswaar genuanceerd - niet alles is fout - maar ze liegt er niet om. Het is een vuurtje dat steeds weer oplaait. Ik kan eraan toevoegen dat over andere diagnostische procedures, zoals het diagnostisch interviews en het diagnostisch proces nauwelijks studies over de betrouwbaarheid en validiteit verricht zijn.

Kritische reflectie door diagnostici zelf Het nadenken door practici over hun gewoonten en interpretaties is als onderzoek van impliciet diagnosticeren op te vatten. Voorbeelden:

(1) Evers onderzocht samen met Europese collega's de frequentie van testgebruik van psychologen en pedagogen die bij landelijke beroepsverenigingen waren aangesloten. Zij vroegen naar problemen die ze bij testgebruik ervoeren en welke drie tests ze het meest gebruikten. De respons van de Nederlandse psychologen was 21%, België 14%, VK 30%, Kroatië 18%, Slovenië 40% en Spanje 12%. De respondenten waren voor twee derde vrouwen; 58% was werkzaam als klinisch psycholoog en 19% als arbeid- of organisatiepsycholoog. Allen beaamden de noodzaak van het gebruik van objectieve tests en uitten kritiek op enkele gewoonten en vaardigheden van testgebruikers, zoals

diagnosticeren op basis van ongeschikte tests,
niet bijhouden van ontwikkelingen in het vakgebied,
niet toetsen van eigen interpretaties bij ervaren collega's,
geen rekening houden met de standaardmeetfout van een test,
tests laten afnemen door ongekwalificeerd personeel,
geen rekening houden met omstandigheden die eventueel eerder aangetoonde validiteit kunnen aantasten: p-c correlaties zijn gecontextualiseerd en
het trekken van onverantwoorde conclusies, gelet op de beperkingen van de test.

(2) Brenner (2003) bestudeerde gewoonten bij het schrijven *psychologische rapporten* en beval een cliëntgericht rapport aan om de relevantie te verhogen. De bezwaren luidden: practici gebruiken te

veel jargon, leggen weinig nadruk op sterke kanten van de cliënt, individualiseren het rapport onvoldoende en geven abstracte adviezen.
(3) Moreland et al. (1995) bepaalden empirisch *test user competencies*. Dit zijn kwaliteitseisen die diagnostici zichzelf opleggen. Er werden twaalf minimumvaardigheden genoemd voor juist testgebruik, bijvoorbeeld '*refraining from helping a favored person to get a good score', avoiding errors in scoring and recording, willingness to give interpretation and guidance to test takers in counseling situations'* (p.16). Ze deden een factoranalyse op de items van testgebruik en trokken zeven factoren met eigenwaarden > 1.00 (p. 17)

I nauwkeurig verslag leggen van wat er tijdens het onderzoek gebeurt
II adequaat testgebruik
III psychometrische kennis: vooral inzicht in *error* en de standaardmeetfout
IV inzicht in testresultaten; bij cutoff scores rekening houden met de standaardmeetfout
V nauwkeurig scoren: denk aan de opmerking van Schmidt en Hunter dat bijna alle fouten van administratieve aard zijn
VI gebruik van juiste normen bij het interpreteren van de scores
VII steun bij het interpreteren in *counseling* situaties.

Kritiek van consumentenorganisaties op tests, selectie en plaatsing Er is een nummer van het tijdschrift van de Nederlandse consumentenbond geweest (rond 1985) dat op basis van het Cotan oordeel bijna alle tests en vragenlijsten onvoldoende voor de consument vond. Het waren volgens de Consumentenbond onvolwaardige producten. Die kritiek is overgewaaid. Recent gaat het om specifieke tests voor specifieke doeleinden die het publiek raken. Vooral de Cito toets voor groep 8 komt jaarlijks in het nieuws. Deze toets voldoet echter op alle rubrieken met een goed of voldoende waardering.
De Amerikaanse consumentenbond onder leiding van Ralph Nader heeft bezwaren evenzeer tegen tests geuit. Dat heeft meer effect gehad op de testpraktijk dan het oordeel van de Nederlandse consumentenbond. In de VS worden klachten ingediend en rechtszaken aangespannen tegen het label dat een persoon toegewezen krijgt op basis van testonderzoek. Er is aangevochten dat een IQ testscore een label als *mildly retarded* zou rechtvaardigen. Ouders hebben een dergelijk label soms via een gerechtelijke procedure laten verwijderen uit het dossier van hun kind. Testuitslagen en certificatiebeslissingen hebben tot gerechtelijke procedures geleid. Ralph Nader leidde in de VS een consumentenorganisatie die geprobeerd heeft intelligentie (IQ) en schoolprestaties (SATs) te laten verbieden omdat minderheidsgroepen er door benadeeld zouden worden. Denk aan de observatie van Jensen (1923-2012) dat het IQ-verschil tussen blanke en niet-blanke Amerikanen meer dan één SD bedroeg (15-19 IQ punten). Dit verschil is niet afgenomen tussen 2000 tot 2010. Hij concludeerde daaruit dat inspanningen om het verschil te verkleinen, nutteloos zijn geweest. De diagnosticus zal opmerken dat de tests de verschillen weliswaar vaststellen maar ze niet veroorzaken. Kenmerken van personen, gezinnen, samenleving en het schoolsysteem dragen er naast andere factoren toe bij dat de verschillen gecreëerd worden en behouden blijven. Peilingsonderzoek *(Minimal Competency*

Testing) van lezen en rekenen van basisschoolleerlingen en van toekomstige leraren kan eenzelfde lot treffen. Telkens blijkt dat kennis niet gelijk verdeeld is over de sociale klassen. In Nederland komen gerechtelijke processen zelden voor maar dat kan verkeren.

Samenvatting en conclusie

De kritiek van de beoefenaars zelf op hun werk kwam aan de orde bij de kritiek op de vaagheid van psychologische constructen en hun relaties en de beperkte toetsbaarheid. Operationalisatie is nooit dekkend en sommige psychometrici willen van de inhoud van items af (interpretatie is *in the eye of the beholder)* af en stellen als eenheid een model voor een item of een latente trek voor. Meten is soms niet meer dan antwoorden optellen, sommigen - van Thurstone tot Rasch - zien kans psychologische kenmerken te schalen. Trendler vindt meten in de Newtoniaanse zin onmogelijk in de psychologie.

Je leest echter het meest over kritiek op meetinstrumenten. In Nederland is een uitvoerig systeem ontworpen dat berust op de *APA Standards* dat vooral klassiek testtheoretische vereisten aan tests en vragenlijsten bevat. In 2007 zijn meer dan 500 instrumenten beoordeeld op zeven rubrieken. Dat beeld is niet onverdeeld gunstig en het is niet verbeterd in de laatste 20 jaar. Misschien is er een limiet aan wat tests en vragenlijsten kunnen. De kritiek op het psychiatrisch classificatie systeem voor stoornissen de DSM gaat door. Er komt geen ander dimensioneel systeem voor in de plaats op basis van multivariaat en modern testtheoretisch onderzoek. Dat is een *tour de force* en die is niet te verwachten bij de gewoonte onderzoek op te splitsen in *least publishable units.*

Reflectie van diagnostici op hun omgang testen en met gegevens uit testhandleidingen laat zien dat ze de klassiek testtheoretische regels willen volgen. Ze zijn zich bewust van de Schmidt en Hunter opmerkingen dat er administratieve fouten gemaakt worden, te snel geïnterpreteerd wordt en onvoldoende rekening gehouden wordt met het feit dat betrouwbaarheidscoëfficiënten gecontextualiseerd zijn door de steekproef waarop ze berekend zijn.

Consumentenorganisaties gebruiken de testbeoordelingen om kritiek op tests voor selectie en plaatsing uit te oefenen. In de VS zijn op basis daarvan tests verboden bij het toekennen van labels aan cliënten. In Nederland is er af en toe protest. Dat is zeer beperkt gegeven het grote aantal getesten.

2. Fraude door professionals en wetenschappers

Professionals kunnen over een cliënt of opdrachtgever/niet-cliënt rapporteren zoals de eerste of de laatsten dat wensen. Hij kan onzorgvuldig zijn. Hij kan zich laten beïnvloeden door vooroordelen en eigenbelang. Er zijn weinig gegevens over dit gedrag bij diagnostici en therapeuten. In Nederland zijn er maar enkele aangemelde klachten. Jaarlijks volgen op 10 tot 15 klachten reacties van de ethische commissie van de beroepsvereniging: het NIP. Dat is zeer beperkt gegeven het grote aantal personen dat met een diagnosticus of therapeut te maken krijgt: een schatting rond de 500.000 personen worden gediagnosticeerd. Klinische

experts noemen dit het topje van de ijsberg en het verschilt waarschijnlijk niet van klachten over medici en verpleegkundigen, die ook BIG geregistreerd zijn.
Onderzoekers zijn lang beschouwd als vrijgestelden die het in harmonieuze samenwerking om de waarheid en niets dan de waarheid te doen is. Het zijn mensen die bij veeleisende instellingen zoals universiteiten, hogescholen en onderzoeksinstituten werken. Bestuurders vatten deze soms op als bedrijven met als doel productie en de medewerkers worden op apenrotsen en in slangenkuilen geplaatst. Er is kinnesinne en bijgevolg fraude en plagiaat. Onderzoeksresultaten worden gunstig voorgesteld voor opdrachtgevers. Om te publiceren worden gegevens selectief benut, gewenste resultaten getoond en zelfs data verzonnen. Van der Meulen (2008) zegt in zijn afscheidsrede dat alle vormen van empirisch onderzoek uiteindelijk op ethische principes berusten. Hij noemt:

(a) eerlijkheid, bijvoorbeeld geen onwelkome feiten wegdrukken
(b) onpartijdigheid: het buiten boord houden van niet-wetenschappelijke belangen
(c) burgermoed, de moed tot openheid, duidelijkheid en uitnodigen van tegenstanders: het zich kwetsbaar durven op te stellen en
(d) onafhankelijkheid: respect voor redelijke argumenten en weerstand tegen druk van machtigen of meerderheden.

Deze principes worden af en toe geschonden: Voorbeelden:

(1) John et al. (2012) deden verslag van praktijken zoals het weglaten van variabelen die de kernhypothese niet ondersteunen en selectief rapporteren zodat het onderzoek succesvol lijkt. Daarnaast wijzen ze op de cultuur om onderzoek niet te repliceren, focus op significantietoetsing, de 'publiceer of ik schiet' cultuur en gebrekkige methodologische kennis van onderzoekers. Een *educated guess* is dat fraude in de een of andere vorm voorkomt bij 3% (± 2%) van de onderzoekers. Dit getal treft men vaak aan bij fraude in allerlei instellingen en bij transacties. Het varieert van liegen en contraproductiviteit tot financiële fraude.
(2) De journalisten Berkhout en Rosenberg (2012) deden verslag van 110 onderzoeken met een luchtje aan twaalf Nederlandse universiteiten tussen 2005 en 2011. Zevenentwintig waren aanleiding voor maatregelen, bijvoorbeeld een medisch specialist van de UVA produceerde 112 artikelen in één jaar (NRC, 2013). Deze getallen berusten op zelfrapportages. Het werkelijke aantal is waarschijnlijk hoger. In een enquête bij 800 basisartsen en medische specialisten rapporteerden 36% dat coauteurs waren toegevoegd die geen bijdrage hadden geleverd aan het onderzoek. Volgens de *Vancouver rules* is elke auteur verantwoordelijk voor de studie. Als die waren toegepast had de helft van de auteurs verwijderd moeten worden.
(3) Redacteuren van het *British Medical Journa*l vroegen 2700 wetenschappers of ze wisten dat collega's gegevens hadden veranderd of gefabriceerd. Het resultaat was dat 13% meedeelde enige vorm van fraude te hebben waargenomen.
(4) Fanelli deed in het *Open Access* tijdschrift *PloS One* van mei (2009) verslag van haar meta-studie van 18 reviews met 2.434 artikelen uit verschillende wetenschapsgebieden. Ze vond dat 2% van de onderzoekers de resultaten hadden gefabriceerd of veranderd en 33.7% deelde mee dat hun

praktijken twijfelachtig waren zoals gegevens weglaten die de hypothese niet ondersteunen of hypothesen weglaten die niet bevestigd werden.
(5) Voorbrood en Van Luijn (2010) onderzochten of psychologische gegevens geregistreerd en beschikbaar waren voor andere onderzoekers. Veel gegevens worden immers gefinancierd door de belastingbetaler of door een goede doelen fonds zoals KWF en de Hartstichting. Van Luijn (2012) ondervroeg uitvoerig 19 UHDs en hoogleraren en verzamelde gegevens met behulp van enquêtes bij 173 onderzoekers. Zij concludeerde dat er nauwelijks systematische archivering plaats vindt en er geen regels zijn voor het delen van gegevens. Hoewel velen wel gegevens wilden delen merkten ze op dat het moeizaam is om data te verkrijgen en dat anderen daar niet goede sier mee mogen maken. Om fraude te voorkomen stelt ze voor om het verzamelen en archiveren van betrouwbare gegevens te belonen als publicaties. Dit moedigt replicatie- en meta-studies aan.

Hoeveel fraude? Er is enig zicht op soort en omvang van fraude. Je weet het nooit precies want men loopt er als instelling en persoon niet mee te koop. 'Wie zonder zonde is, werpe de eerste steen'. De 'publiceer of ik schiet' cultuur wordt soms als boosdoener aangewezen. Wicherts et al. (2006) en Wicherts et al. (2011) stellen voor om de ruwe gegevens vrij te geven. Dat zou fraude voor een deel voorkomen. Ze suggereren dat het niet bereid zijn om ruwe data ter beschikking te stellen te maken heeft met twijfel aan de robuustheid van de gegevens en resultaten want als belangrijk en robuust omschreven resultaten blijken bij passende toetsing soms niet significant. Wicherts et al. stellen dat één op zeven onderzoekers die hun gegevens niet wilden delen niet-significante resultaten gevonden zouden hebben als ze streng en goed getoetst zouden hebben. Ze gaan ervanuit dat niet-significante resultaten niet bijdragen tot kennis. Waarom eigenlijk niet? Het laat zien dat een veelbelovende variabele, interventie of procedure niet effectief is. Dat leert dat je het niet daar maar elders moet zoeken.
De onderzoeksjournalist Kolfschoten (2012) heeft een boek over de bekendste fraudegevallen in Nederland op een rij gezet. In 1993 (2de druk) had hij al de *Valse vooruitgang: Bedrog in de Nederlandse Wetenschap* geschreven. De KNAW, de Nederlandse Organisatie voor Wetenschappelijk Onderzoek en de Vereniging van Nederlandse Universiteiten hebben in 2001 al een *Notitie Wetenschappelijke Integriteit* het licht doen zien met als ondertitel: *Over normen van wetenschappelijk onderzoek en een landelijk orgaan voor wetenschappelijke integriteit*. Fraude is niet van vandaag of gisteren en de discussie laait weer op na nieuwe gevallen van fraude. Staatsecretaris van Onderwijs, Dekker heeft in november (2013) aangekondigd dat binnen 10 jaar alle gesubsidieerde onderzoeksartikelen vrij toegankelijk moeten zijn. Amerikaanse psychologen die de ethische code onderschrijven moeten data vrijgeven volgens Artikel 8.14 dat luidt:

After research results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims though reanalysis and who intend to use such data only for that purpose, provided

that the confidentiality of the participants can be protected and unless legal rights concerning proprietary data preclude their release.

Iets voorschrijven is gemakkelijk. Het gaat erom of er gevolg aan wordt gegeven. De weg naar de hel is immers geplaveid met goede voornemens. Het is zaak bij te houden of voornemens uitgevoerd worden. Principes opleggen is gemakkelijker dan ze in gedrag omzetten. Na fraude is er steeds een commissie die zegt hoe het beter moet. De tijd moet uitwijzen of dat tot verandering leidt. Zolang de klassieke prikkels blijven zal verandering beperkt zijn en zullen nieuwe fraudes voorkomen.

Box Wetenschapsfraude

Een populaire en hogelijk gewaardeerde hoogleraar sociale psychologie aan de Universiteit van Tilburg (UT) moest erkennen dat hij gedurende jaren gegevens verzonnen had. Dit werd bewezen in 55 van de meer dan 130 van zijn *peer-reviewed* artikelen. Zelfs zijn promovendi mochten geen gegevens verzamelen en analyseren. Toen enkele vroegen om de data kregen ze wat ingevulde vragenlijsten en berekenden Cronbachs alfa. Deze was zo laag dat ze het idee kregen dat de lijsten lukraak ingevuld waren. Hij had gepubliceerd met 70 coauteurs en slechts een enkele voelde nattigheid. Op die ene enkeling na lichtte niemand de decaan in over de te mooie resultaten. Het was een schok voor de Nederlandse sociaalpsychologen. Het is breed uitgemeten in de pers. Psycholoog-journalisten zijn vaak als sociaalpsycholoog opgeleid. Er kwam een commissie onder leiding van Levelt (oud voorzitter van de KNAW) die een rapport maakte over de affaire. Het is te downloaden van www.knaw.nl of te verkrijgen via div@bureau.knaw.nl. Het rapport bevat een grondig onderzoek van een groot aantal artikelen en geeft de kwalificatie *sloppy science* aan sociaalpsychologisch onderzoek vanwege methodologische en conceptueel-theoretische zwaktes. De sociaal psycholoog De Dreu (2012) reageerde op het rapport met enkele opmerkingen en aanbevelingen:

(a) Psychologisch onderzoek is niet langer het werk van één persoon: waardeer de onderzoeksgroep voor publicaties en niet individuele onderzoekers om ze af te rekenen.

(b) Waardeer replicatiestudies en maak tijd en geld vrij voor replicaties.

(c) Versnel het rondgaan van onderzoeksbevindingen: het *review* proces neemt soms meer dan een jaar in beslag. *Reviewers* moeten waardering krijgen voor hun werk. Er wordt in het rapport gesuggereerd dat zij mee verantwoordelijk zijn voor *sloppy science* door slordige te reviewen.

(d) Leg meer nadruk op integriteit in psychologiecursussen en op overleg tussen onderzoekers: ieder moet mogen meekijken.

(e) De richtlijnen voor integriteit moeten onderschreven worden door verenigingen van psychologen.

Na de fraude hebben de sociaalpsychologen in Nederland dit onderschreven en voegden toe: maak de gegevens beschikbaar voor alle medewerkers van een vakgroep; vermeld het gewicht van de bijdrage van elk auteur en coauteur; geef pas de bevindingen door aan de media als het artikel beoordeeld en geaccepteerd is; *reviewers* moeten meer aandacht besteden aan de methodologie: samenstelling steekproef, metingen en data analyses.

Redacteuren van tijdschriften moeten replicatiestudies en niet-significante bevindingen publiceren, desnoods in een aparte rubriek.

Samenvatting en conclusie

Elke professionele en wetenschappelijke activiteit kan beter. Er worden echter ook bewust fouten gemaakt om een belang veilig te stellen, bijvoorbeeld de reputatie van een instelling of het aanzien van een professional of onderzoeker. Naast een enkel voorbeeld van flagrante fraude in de vorm van plagiaat en het verzinnen van data komen pogingen voor om de resultaten mooier voor te stellen dan ze zijn. Door recente duidelijke gevallen van fraude heeft men de noodzaak gevoeld de omvang vast te stellen en maatregelen te nemen om het te voorkomen.

3. Kritiek van publiek en cliënten

Publiek en de pers Het publiek is overwegend sceptisch over psychologie en diagnostiek. Sommige columnisten beschouwen haar niet als wetenschap. Ze worden in de kaart gespeeld door fraude en onenigheid van en tussen wetenschapsbeoefenaars. De polemist, econoom, neerlandicus en schrijver van een brievenroman samen met fraudeur Stapel, Dautzenberg (2013, p. 12) is daar een voorbeeld van:

Vijfentwintig jaar geleden kwam ik voor het eerst in aanraking met de wetenschap (als student), daarna beroepsmatig (als adviseur, onderzoeker, journalist en schrijver). Sindsdien zie ik de universiteiten als hoeders van fantastische constructies, als saaie en bovenal bureaucratische toverkastelen waarin niettemin lekker geheimzinnig gedaan wordt over weinig tot niets. De wetenschap heeft de samenleving absoluut verrijkt - met illusies, raadsels en enkele praktische handvatten.

De pers is voor een deel juichend over spraakmakende goed in het gehoor liggende bevindingen. Daarnaast meet ze fraude en gebrek aan replicaties van belangrijke studies breed uit.

De psycholoog Lilienfeld (2011) somt de kritiek van leken en de pers op. Het is gezond verstand, of dat nog niet eens, de methoden zijn niet wetenschappelijk, het levert de samenleving niet veel op, de resultaten zijn te mager: zwakke predicties, controles en hulp bij beslissingen. Hij verdedigt zijn vak en vermeldt gerechtvaardigde redenen voor de kritiek, zoals gebrek aan kwaliteitscontrole en toezicht bij klinische en onderwijsdiagnostiek en het uitoefenen van betwistbare praktijken. Ferguson (2015) erkent de zinnigheid en relevantie van Lilienfelds kritiek. Optimist als hij is, doet hij suggesties om de kritiek te ontzenuwen. Hij pleit voor een cultuuromslag in onderwijs en onderzoek. Met MA en Phd studenten (universitaire docenten noemt hij niet) zijn open discussies nodig over de beperkte resultaten uit onderzoek en praktijk, over de eenzijdigheid van methoden en de reductie van gedrag en omgeving tot een simpel waarneembare respons en stimulus. Hij legt nadruk

op de resultaten, meer nog dan op theorie: terug naar de feiten in al hun beperktheid. Hij heeft oog voor de complexiteit en gelaagdheid van gedragingen. De menselijk conditie (de *condition humaine* van de (katholieke) Franse filosofen) is geen recht gedaan met een paar reflexen en observeerbare reponsen. Transparantie, replicatie van resultaten en falsificeren vat hij op als ononderhandelbare kenmerken van welke wetenschap dan ook. Je kunt je afvragen of een wetenschapper volledig transparant kán zijn. Hij heeft immers zijn ongearticuleerde onwrikbare regels voor wat een wetenschappelijk geldige uitspraak over gedrag is. En falsificeren? Kan en gebeurt dat feitelijk in psychologisch onderzoek? Ik heb er geen voorbeelden voor gevonden. De ene hypothese/theorie wordt niet tegen een andere weggestreept op basis van onderzoek. Kennis is in de psychologie eerder een olievlek die zich uitbreidt, terwijl het oude wel eens vergeten wordt, maar altijd een nieuwe kans kan grijpen.
De rationalistische attitude volgens welke leken en practici irrationele brokkenmakers zijn is hiervoor gerelativeerd. Het is een uitdaging om in eigen huis orde op zaken te stellen en het debat te blijven voeren over wat diagnostiek waard is. De enige remedie tegen scepsis is openheid en realisme over wat onderzoek vertelt en wat toepassing oplevert. Bescheidenheid is realistisch, maar een valse is niet zinvol.

Cliënten Naast de houding van het publiek en pers is er de ervaring van getesten. Er is weinig bekend over tevredenheid van cliënten. Commerciële onderzoeksinstituten zullen daar een beeld van hebben. Ze prijzen zich op grond van positieve uitingen aan. Negatieve berichten worden zelden gepubliceerd. Drenth is in 1967 nagegaan of er nu veel 'protesten tegen testen' waren bij betrokkenen. Dat viel toen erg mee. Het beeld nu is minder duidelijk. Het Nederlands Instituut van Psychologen (NIP) heeft een commissie die klachten behandelt. Het gaat jaarlijks om minder dan honderd klachten: een betrekkelijk klein aantal gegeven de omvang van diagnostische en therapeutisch werk.
In de VS worden gemakkelijk klachten ingediend en rechtszaken aangespannen. Er is met succes aangevochten dat een IQ testscore een label als *mildly retarded* zou rechtvaardigen. Ouders hebben een dergelijk label soms via een gerechtelijke procedure laten verwijderen uit het dossier van hun kind. Testuitslagen en certificatiebeslissingen hebben tot gerechtelijke procedures geleid. Ralph Nader leidde in de VS een consumentenorganisatie die geprobeerd heeft intelligentie (IQ) en schoolprestaties (SATs) te laten verbieden. De diagnosticus zal opmerken dat de tests de verschillen weliswaar vaststellen maar ze niet veroorzaken. Kenmerken van personen, gezinnen, samenleving en het schoolsysteem dragen er naast andere factoren toe bij dat de verschillen gecreëerd worden en behouden blijven. In Nederland komen gerechtelijke processen zelden voor, maar dat kan verkeren.
Scholen betrekken testresultaten (Cito-scores) in hun toelatingsbeleid. Daarbij mag een instrument volgens onderwijs-autoriteiten alleen gebruikt worden als een commissie van deskundigen (de Cotan) het instrument op een aantal onderdelen minimaal een voldoende geeft. Het is een belang van testconstructeurs en -uitgevers dat de test het

‘voldoende/goed’ keurmerk krijgt. Diagnostici zijn betrokken bij de selectie van personeel. Als iemand afgewezen wordt is dat frustrerend. Deze stemming wordt door sommige psychologen gebruikt om tests en interviews te devalueren. Ze bieden een training aan die zou wapenen tegen ‘lepe selecteurs’ en hun tests en geven ‘handvatten’ om beter te scoren en ‘uitgekookte vragen’ van interviewers te omzeilen. Er zijn ook toets- en examentrainingen bij toelating tot scholen en tot universiteiten. Dit betekent dat de ene groep psychologen tegen de andere opstaat.
Deze kritiek is zinvol als het reflectie is op wat de diagnosticus en de wetenschap kunnen waarmaken. Een diagnosticus kan niet anders dan transparant zijn mogelijkheden en beperkingen uitleggen. Diagnostiek sluit naast kennis en kunde ook ethiek in. Cates (1999) noemde diagnostiek een hybride van wetenschap, kunst en kunde. De cliënt heeft het recht dat te weten.

Samenvatting en conclusie

Leken zijn sceptisch over de diagnose en het diagnostisch proces. Over hulp in de vorm van trainingen en therapieën zijn ze minder kritisch. De pers is dubbelhartig: ze brengen graag nieuws over spectaculaire vondsten. Daarvoor selecteren uit de grote pool aan onderzoeken dat wat hen uitkomt en over fraude rapporteren is *fun.* Schattingen ontbreken vaak. Polemisten leggen soms een vinger op de zere plek en wijzen op het triviale en beperkte van bepaald onderzoek en de reputatie die instellingen en wetenschappers zich menen te kunnen toe-eigenen op basis van hun prestaties. De cliënt uit, gegeven de omvang van testonderzoek en hulpverlening, weinig kritiek. De diagnosticus kan tegen publiek, pers en de cliënt niets anders doen als open zijn over zijn werkwijze en wijzen op de bescheiden resultaten van diagnostiek als beschrijving, voorspelling en controle van het (on)gewenst gedrag. Hij kan geen zekerheid bieden.

4. *Faken* door cliënten

Bij twee partijen, cliënt en diagnosticus zijn er twee die ‘frauduleus’ kunnen zijn. De nadruk ligt vooral op de diagnosticus. Ik ga hier ook in op de cliënt. Een sollicitant kan zich anders voordoen dan hij is. Intentionele oneerlijkheid van een persoon noemen we faken. De belangen van getesten kunnen van invloed zijn op de hoogte van de scores op sociaal wenselijk geachte persoonskenmerken. Het is onwaarschijnlijk dat cliënten faken op prestatietests ofschoon dat in het verre verleden gebeurde om onder militaire dienst uit te komen. De kans op faken is aanwezig bij sollicitaties en toelating tot prestigieuze opleidingen. Psychologen hebben integriteitstest ontworpen bij selectie van personeel dat op fraudegevoelige posten terechtkomt, bijvoorbeeld accountants, politiefunctionarissen, personeel van geheime diensten en financiële beheerders. Er zijn veel studies over faken van werknemers in bedrijven. Want werkgevers willen betrouwbaar personeel. Voorbeelden:

(1) Onderzoek naar het effect van instructies, bijvoorbeeld een vragenlijst invullen bij een sollicitatie of als deelnemer aan wetenschappelijk onderzoek of naar het effect van plaatsing van de vragen, bijvoorbeeld alle items over extraversie bij elkaar of verspreid over de hele vragenlijst, laten verschillen zien. Een sollicitatieconditie leidt doorgaans tot hogere scores voor Emotionele Stabiliteit en Gewetensvolheid dan de conditie: 'Ga af op je eerste indruk'. Resultaten bij Nederlandse psychologiestudenten lieten *geen* verschillen zien in gemiddelden op de *Big Five* onder de instructie: 'U solliciteert naar een zeer gewilde stageplaats' of: 'Antwoordt u s.v.p. snel en spontaan'. Bij een kleine groep samengesteld uit studenten en hoogopgeleide ongeveer 40-jarigen is een significant effect gevonden van de instructie 'Je antwoord op de vragenlijst waar moeten maken' versus 'Spontaan en snel je antwoord geven'. In de 'waarmaken' conditie werd significant hoger gescoord op Gewetensvolheid en Emotionele Stabiliteit (ter Laak et al., 2000).
(2) Een meta-studie in de VS van Ones en Viswesvaran (1993) laat zien dat het verstandig is om rekening te houden met *fakability*. De studie is gebaseerd op 655 validiteitscoëfficiënten en toont dat integriteitstests significant werkprestaties en contraproductief gedrag voorspellen. Bij het laatste moet je denken aan stelen, disciplinaire problemen en absentie. De geschatte gemiddelde predictieve validiteit van integriteitstests voor de beoordelingen van supervisoren van werkprestaties van personeel bedroeg r = .41. Men kan toevoegen dat antwoorden op integriteitstests ook *gefaked* kunnen worden. De waarde van .41 is hoog. Er zijn waarschijnlijk andere demografische en persoonskenmerken bij betrokken die met integriteit samenhangen.
(3) Viswesvaran en Ones (1999) vonden dat alle Big Five factoren te faken *waren*, maar Birkeland et al. (2006) troffen in een meta-studie, berustend op 33 onderzoeken kleinere verschillen aan: over allerlei functies een gemiddeld verschil tussen sollicitanten en niet-sollicitanten voor Extraversie d = 0,11, Emotionele Stabiliteit d = 0,44, Gewetensvolheid d = 0,45 en Openheid d = 0,13. Dit beeld werd eerder aangetroffen door Griffin et al. (2004) bij Australische en door McFarland et al. (2000) bij Amerikaanse subjecten. Ellington et al. (2007) gebruikten de *California Personality Inventory* (CPI) bij 713 subjecten en wilden het nut van zo'n instrument aannemelijk maken. Ze zeggen dat de analyses bij verschillende instructies *'...a limited degree of distortion'* lieten zien. De Cattell lijst 16PF is ook geanalyseerd onder verschillende condities. Stark et al. (2001) noemden *faken* met de 16PF een zorgelijk verschijnsel. Men heeft dat proberen tegen te gaan door de instructie: 'zo snel mogelijk spontaan te antwoorden'. Holden et al. (2001) deden dat maar het bleek dat beperken van de responstijd het effect van faken op de validiteit niet verhoogde.
(4) Tett et al. (2011) analyseerden faken op Emotionele Intelligentie (EI) en de (Big Six: BF + 1 bij 150 eerstejaars studenten. Als predictoren gebruikten ze cognitieve capaciteitsscores, de mogelijkheid tot faken en de relevantie van de trekken voor verpleegkundigen, marketeers en computerprogrammeurs. Studenten met de naar verhouding hogere cognitieve capaciteiten die naar verhouding lager scoorden onder de 'eerlijk conditie' scoorden hoger op faken en het was hoger voor job-relevante trekken. Het zich mooier voordoen op EI was bijvoorbeeld hoger bij verpleegkundigen dan bij computerprogrammeurs. De gemiddelde effecten over subjecten en trekken hadden een omvang van 0,83 SD. Dat is substantieel. Faking nam 60% toe bij laagscoorders onder de 'eerlijk conditie', 26% op de job-relevante trekken en 20% bij de subjecten met hoge cognitieve capaciteiten. Het hoogste niveau van faken kan verwacht worden bij hen die laag scoren op de job-relevante *traits* maar slim genoeg zijn om zich voor te doen als geschikt voor de job (p. 201). Dit was al gevonden in een studie van Johnson et al. (2009). Zij rapporteerden dat verkopers voor de farmaceutische industrie extreme antwoorden gaven op de vragen die wenselijk waren voor

hun succes als verkoper. Deze *bias* verlaagde de betrouwbaarheid en validiteit en liet een eerste sterke factor zien met veel *job-desirable* items.
(5) Scherbaum et al. (2013) benutten IRT om te laten zien dat de antwoorden (de IRFs) op de afzonderlijke items veranderden bij twee instructies. De eerste luidde: *Please answer this inventory as honest as you can; the results will be anonymous; describe yourself as you really are.* De tweede was *Answer as if you were applying for a job you really want; respond in ways that make you attractive for the organization; but not too obviously*. Beide condities werden voorgelegd aan 1001 studenten (60% vrouwen), gemiddelde leeftijd 20.37 jaar (SD 6.45 jaar) die deelnamen aan een eerstejaars psychologiecursus. Er waren significante verschillen tussen condities: Neuroticisme (N) was lager in de fake conditie: t (998) = -21,94; d = - 0,61 en Gewetensvolheid (C) hoger: t (998) = 20.77; d = 0,48. Dit zijn gemiddelde tot substantiële effectgroottes. Het IRT model paste op de schalen N en C van de *International Personality Item Poll* van Goldberg (1999). De IRT itemparameters lieten zien dat er één C item een vooral verschil maakte: *Fear for the worst* (zie Scherbaum et al. 2013, p. 213 voor de itemparameters).

Deze studies laten zien dat *faken* op persoonlijkheidsvragenlijsten regelmatig en met een zekere omvang voorkomt. De sollicitatiesituatie leidt tot gemiddelde of hoge d-waarden dan een onderzoeksituatie. Er zijn culturele verschillen. Het faken komt vooral voor als er belangen van respondenten op het spel staan. Het verschijnsel is niet gelijk voor elke persoonlijkheidstrek: vooral Neuroticisme en Gewetensvolheid zijn gevoelig. Het beperken van de responstijd maakt niet uit. Met behulp van IRT itemparameters kan per item getoond worden hoe faken uitwerkt. Vecchoine et al. (2012) laten nog zien dat de factorstructuur van de *Big Five* niet wijzigt onder de twee condities. Die is meet-equivalent bij sollicitanten en niet-sollicitanten maar de gemiddelden verschillen wel (zie Scherbaum et al., 2013, hierboven).

'Malignering' Een speciaal geval van faken is het simuleren van psychische stoornissen. Dit is doelbewust liegen over symptomen om er juridisch of financieel voordeel uit te halen. Een simulant wil bijvoorbeeld in aanmerking komen voor een uitkering, een verblijfsvergunning verkrijgen, of een gevangenisstraf ontlopen. Het gaat over klachten die lastig hard te maken zijn, zoals moeheid, somberheid, slaap- en concentratiestoornissen. Het is moeilijk om de incidentie te bepalen: een goede simulant betrap je immers niet. Niettemin zijn er schattingen: Schmand et al. (1998) zeggen dat 60% van de whiplash patiënten, die door een verzekeringsmaatschappij verwezen zijn, simuleren. PTTS klachten van ongeveer 25% van de VS veteranen zouden op simulatie berusten. Volgens Singh et al. (2007) worden PTSS, psychose, cognitieve stoornissen en zwakbegaafdheid relatief het meest gesimuleerd. IND medewerkers, klinisch psychologen en psychiaters krijgen met simulanten te maken. Ze zijn daarbij aangewezen op hun kennis van de dossiers van betrokkenen, van symptomen bij stoornissen en van hun vaardigheden als ondervrager. De diagnose vereist het schatten hoe 'echt' de symptomen zijn. Het is zaak om subtiel door te vragen en inconsistenties proberen op te sporen. Ze kunnen gebruik maken van enkele

hulpmiddelen, bijvoorbeeld de *Structured Inventory of Malignered Symptomatology'* (SIMS: rapportage van te zeldzame symptomen: Schmand et al., 1999) en de Amsterdamse Korte Termijn Geheugen Test (AKTG) om doelgericht onderpresteren op te sporen (Schmand et al., 1999). Simuleren, *malignering* kan bijna niet worden *bewezen* (Van Staveren, 2014).
De diagnosticus is niet naïef. Hij weet dat mensen zaken soms mooier voorstellen dan ze zijn, vooral als dat in hun belang is. Hij moet bij de cliënt zien te achterhalen of dit verschijnsel zich voordoet en of het ertoe doet. Het is niet meteen duidelijk of dit de predictieve validiteit verlaagt, ook al vonden Johnson et al. dit in de studie bij verkopers van farmaceutische producten. Het kan product- en beroepsspecifiek zijn.

Samenvatting en conclusie
Bij fraude onderzoek ligt de nadruk op de onderzoeker en de professional. De cliënt kan ook frauduleus zijn, dat wil zeggen dat hij intentioneel voor zijn eigen belang een mooiere voorstelling van zaken geeft. Deze 'fraude' is vooral uitgezocht met persoonskenmerken en prestaties op het werk. Steeds blijkt dat Gewetensvolheid en Emotionele stabiliteit significant tot substantieel hogere waarden krijgen in situaties waar dat in het belang van de cliënt is. Integriteit kun je opvatten als het ontbreken van 'fraude'. Dit kenmerk hangt gemiddelde tot substantieel samen met werkprestaties. Een bijzondere vorm van fraude is *malignering*. Een cliënt simuleert daarbij een klacht bijvoorbeeld met het oog op het verkrijgen van een financiële regeling of verblijfsvergunning. Het is geconstateerd bij whiplash en PTSS.

5. Omgaan met kritiek
Geen professional ontkomt aan kritiek Dat geldt voor een wetenschapper die een artikel wil publiceren en voor de professional met taken in dienst van de overheid of een instelling. De wetenschapper weet hoe kritisch *reviewers* zijn zowel in het eigen taalgebied als in Engelstalige tijdschriften. De inhoud van die kritiek is niet empirisch geanalyseerd. Waar vallen reviewers over? Dat weten goede publicisten impliciet wel. Het is hun *tacit knowledge.* Er zijn zelfs cursussen waarin de kennis gedeeld wordt om de kans op publicaties van aankomende wetenschappers te verhogen. Dit laat zien dat kritiek niet alleen van objectief wetenschappelijke aard is.

Wat zeggen ervaren publicisten? Gigerenzer (2011) vertelde over zijn ervaring op een congresdiner. Hij zat met drie collega's en vier studenten aan een tafel met Chinees eten. De studenten wilden uit de eerste hand horen hoe hun PhD thesis af te ronden en daarna een goede onderzoeker worden. De heren - het zijn meestal mannen - waren bereid advies te geven. Met autoriteit zei een: 'Doe je vijf experimenten, sla er een nietje door en leg het op het bureau van de promotor. De PhD studenten knikten tevreden. Gigerenzer gooide roet in het Chinese eten: *Don't follow this advice unless you are mediocre or unimaginative; try to think in a deep, bold and precise way; take risks and be courageous*. De PhD studenten

bleven knikken. Ze hadden als de beroemde rabbi kunnen zeggen bij het vaststellen van een tegenstrijdigheid: U hebt beiden gelijk en u hebt ook gelijk dat het tegenstrijdig is. Zo voorkom je de klap die soldaat Tersides kreeg toen hij voorstelde aan koning Odysseus om de belegering van Troje te stoppen na tien jaar uitzichtloos voor de muren gelegen te hebben.

Gigerenzer wees op analyses die laten zien dat de helft van de studies in twee bekende sociaalpsychologische tijdschriften bijna tautologisch waren. Denk aan Smedslunds (2009) opmerkingen over pseudo-empirisch, tautologisch, semantisch onderzoek. Hij vertelde dat hij enkele artikelen in het *Journal of Experimental Psychology* van de jaren 20 tot 30 van de vorige eeuw had gelezen. Hij zei dat er zijns inziens veel verloren is gegaan: verschillende (statistische) methoden, precieze rapportage van individuele gevallen, zorgvuldige selectie van proefpersonen, nagaan of de sekse van de proefpersoon of proefleider van invloed was op het resultaat van de studie en aandacht voor *unobtrusive* metingen. Hij zag nu een dominantie van het verzamelen van data en ze publiceren in de *smallest publishable units without substantive theory.* Een citaat: '*…data without theory are like babies without parents: their life expectancy is low'* (p. 296). Maar, hij is dubbelhartig (Goethes *zwei Seelen in meiner Brust*). Op het Max Planck instituut in Berlijn wordt een publicatie in een Amerikaans hoog impact tijdschrift steeds gevierd.

Ervaren publicisten verschillen van mening: wees pragmatisch, doe wat de goegemeente van wetenschappers wil en voorkom en anticipeer op de kritiek. Anderen zeggen: wees origineel, volg je droom. Voorbeelden:

(1) *Trek je niets van kritiek aan volg je droom* Een toespraak van Mr. Steve Jobs van Apple voor Stanford studenten (2005) zes jaar voor zijn dood staat dichter bij Gigerenzer (zijn ene ziel van de twee in zijn borst) dan bij Kruglanski. Ik heb dit op de TV gezien en aantekeningen gemaakt van de (ondertitelde) toespraak van deze frêle man. U kijkt het maar terug; het filmpje staat vast op *You tube*. Jobs vertelde uit eigen ervaring. Het was geen lezing over Apple. Het is best mogelijk dat de toespraak voor hem geschreven is maar zijn boodschap is duidelijk. Hieronder een parafrase van zijn toespraak.

Zijn eerste ervaring was zijn adoptie. Zijn academische biologische moeder wilde een gezin voor hem vinden dat uit universitair opgeleiden bestond. Zijn adoptieouders waren geen academici. Men kan dit zo lezen: aanleg en omgeving zijn van belang, maar geen is doorslaggevend. De tweede ervaring had te maken met zijn opleiding. Hij behaalde geen graad want de studie verveelde hem. Hij voelde zich schuldig over het geld dat hij verspilde. Hij stopte niettemin met de studie en deed wat hij interessant vond: kalligrafie. De Apple is gekenmerkt door een mooie vormgeving. Zijn derde ervaring: hij werd door zijn eigen bedrijf op straat gezet, toen het niet goed ging. Hij begon een nieuw bedrijf, dat zo vernieuwend was dat Apple het kocht. Hij keerde terug naar huis. Men kan het lezen als: maak dingen die je zelf mooi en de moeite waard vindt, let niet alleen op verkoopcijfers, aandeelhouderswinst. Laat je niet uit het veld slaan door anderen die je dromen onzin vinden, bewandel zijpaden en serendipiteit zal je overkomen en vergezellen. De vierde ervaring was de mededeling dat hij kanker had. Na de boodschap dat hij niet lang meer zou leven, kon hij geopereerd worden en kreeg weer perspectief. Hij toonde letterlijk het *Wunderkind* verhaal (Thompson, 1982)

ook wel het *Golden Boy Narrative* genoemd. Een tragisch verhaal lag meer voor de hand lag. Hij eindigde met *…follow your heart, be foolish, never give up, create beautiful things: Stay hungry, stay foolish!*

(2) *Gezond van geest of ziek van ziel* In James' analyse (1902, 1982) van de religieuze ervaring is een tegenstelling beschreven tussen de eerstgeborenen: de gezonden van geest en de tweemaal geborenen: de zieken van ziel. De eersten hebben het gevoel dat het goed gaat met de wereld en dat ze aan de rechterhand van God zitten. Iets dergelijks propageert de positieve psychologie. De twee maal geborenen zien kwaad, pijn en verlies in de wereld. Ze zijn melancholisch - volgens Aristoteles is dat de (vruchtbare) gemoedstoestand van onderzoekers - met als dieptepunt dat ze het gevoel verliezen voor wat verloren is gegaan en het kwaad ontdekken. Deze tegenstelling bestaat ook in de Islam, waar de Soefi beweging de zwaarmoedige, bevindelijke kant benadrukt, terwijl de shar'ia het wettische en zelfingenomene benadrukt. De Griekse goden Apollo en Dionysos gaan over een georganiseerde versus de anarchistische wereld. In de Chinese cultuur gaat het over het Confucianisme versus het Taoïsme, waarbij de huidige regering terugvalt op Confucius, nu het Maoïsme verlaten is. Onder Mao waren beide verboden.

Welke houding is vruchtbaar, aan welke lessen heeft de geïnteresseerde student het meest? Die van Kruglanski, die van de gezonden van geest (James), de shar'ia, het Apollinische, het Confucianisme? Of de andere pool? Hebben de gezonden van geest niet greep op de besturen, de politiek, het bedrijfsleven, de banken en zelf op de media? In de Nederlandse media is vrijwel iedere spreker vertegenwoordiger van een belangengroep. Hij vertelt je waarschijnlijk niet wat interessant of goed is voor jou, maar voor hem en zijn groep. De nadruk op de *context of justification* van de wetenschap is er gelukkig ook nog; niet alleen de *context of selling.* Nu nog hernieuwde aandacht voor de *context of discovery* (Reichenbach, 1938)) en ontdekken welke kanten van gedrag nog niet of te weinig beschreven zijn.

Omgaan met kritiek van collega-wetenschappers Er is meer bekend over hoe om te gaan met kritiek op wetenschappelijk werk (artikelen) dan op het professionele. Ik beschrijf voorbeelden uit de VS en Nederland. De houding is in ons land dubbel: volg de succesvolle Amerikaan, maar wees niet te volgzaam; ga er een enkele keer tegenin.

(1) Voorkomen en anticiperen De Amerikaanse sociaalpsycholoog Arie Kruglanski onderwijst AIOs, PhD studenten en medewerkers - ook nederlandse - hoe artikelen te schrijven dat ze gepubliceerd worden in Amerikaanse tijdschriften: *'… don't think big, just do your five experiments, clip them together, and hand them in'* (Kruglanksi & Higgins, 2004). De APA heeft een boek gemaakt waarin precies staat hoe een artikel opgeschreven moet worden, wil een beoordelaar er naar kijken. Ze merken op dat sociale psychologie en persoonsleer meer en meer *data-driven* en minder *theory-driven* (p. 96) zijn geworden. Ze beweren dat het vooruitgang tegenhoudt en verhindert om bruggen te slaan tussen verwante wetenschappen. Ze wijzen als oorzaak het BAMA onderwijs aan. Dat is geschikt om studenten methoden en data analyses te leren, maar het schiet te kort om interesse op te wekken voor theorieontwikkeling. De *publish or perish* attitude lokt eenvoudige snel te begrijpen artikelen uit. En professoren en UHDs halen de vereiste hoeveelheid artikelen alleen PhD studenten aan het werk gaan en het onderzoek doen volgens het boekje van de APA richtlijnen.

Kruglanski heeft in Nederland aanhangers. Het *Utrecht Centre for Child and Adolescent Studies* (CAS, z.j. verzonden: November, 2014) biedt als onderdeel van de PhD training - alles in het Amerikaans/Engels: 'Specialized courses offered by CAS staff'aan: # 1 -Publishing in Social Science en # 6 -Tacit Academic Knowledge: Hidden rules for Academic Success in Times of the Replicability Crisis'. Daarin worden antwoorden gegeven op vragen zoals: 'How do I write up my research findings in a way that interests and convinces others'. 'How can I decide on the best potential outlets for my manuscripts'. 'How do I wisely create opportunities for myself on the job market'. In zes twee- uur sessies gaat het om hoe onderzoek gepubliceerd te krijgen, hoe subsidies binnen te halen, hoe productie te verhogen, hoe om te gaan met afwijzingen en hoe de indruk van *sloppy science* - een term van Levelt, de voorzitter van de commissie 'Fraude Stapel' - te vermijden. De houding van Kruglanski en aanhangers is die van de sofisten in Plato's tijd. Die waagden zich evenmin *Theorein* (niet belangeloos schouwen maar overtuigen en je gelijk halen) en *Sooidzein* (het object van onderzoek recht doen, het laten zien zoals het is).

(2) De kritiek de wind uit de zeilen nemen Vier Nederlandse bèta wetenschappers deden (NRC, 20 april, 2013) onder meer suggesties hoe je doel verwezenlijkt te publiceren door de kritiek de wind uit de zeilen te nemen. Ten *eerste* moet onderzoek origineel zijn. Je verricht het bijvoorbeeld op plekken waar iemand anders nooit geweest is: Noord- of Zuidpool, de binnenlanden van Afrika en Zuid-Amerika. Of, je pak een onderwerp waar een ander nog niet aan gedacht heeft. De primatenonderzoeker De Waal zei dat hij coöperatief gedrag bij apen is gaan onderzoeken omdat deze dieren meestal als agressieve overlevers worden beschouwd. Hij zei dat hij dit deed om eruit te springen en dat is gelukt. Wellicht maakt zijn vermogen het gedrag van primaten te observeren meer indruk dan zijn interpretaties. De titel van zijn laatste boek luidt: *Bonobo en de tien geboden*. Observatie en interpretatie probeer je uit elkaar te houden. De vragen mogen diffuus en onduidelijk, niet Cartesiaans klaar en distinct zijn, zeggen onze bèta's. Dit is in strijd met nulhypothesetoetsing, logisch positivisme en de publicatiegewoonten in de psychologie. Ten *tweede* moeten onderzoekers het tijdschrift en de *reviewers* kennen. Vaak is een auteur zelf ook *reviewer*. Samenwerken met Amerikaanse auteurs vergroot de kans van publicatie. Soms kun je vóóraf vragen of het tijdschrift geïnteresseerd is in je onderzoek. Het bezoek van congressen kan de kans vergroten op publicatie. Dit zijn geen intrinsiek wetenschappelijke, maar wetenschapssociologische regels. Ten *derde* moet het artikel helder geschreven zijn. De titel is belangrijk: niet te aarzelend, niet te brutaal en zonder jargon. De referentiebrief (de brief, die met het artikel meegestuurd wordt) moet precies zijn. De Groot had moeite met deze eis van helderheid, want op grond van dit criterium zouden veel filosofische geschriften nooit gepubliceerd zijn van Plato, Selz, tot Heidegger en Popper. De Groot was bekend met het werk van Popper en de filosoof/denkpsycholoog Otto Selz. Beiden spelen een rol in zijn dissertatie: *Het denken van de schaker* (1946) en in zijn boek: *Methodologie* (1961). Filosofen werken een grondintuïtie uit die niet rationeel en logisch te verantwoorden is. Als het lukt, brengt die intuïtie een nieuw facet van denken/gedrag of een verrassend perspectief op de mens, zijn gedrag en de samenleving aan het licht. Deze zijn niet van meet af aan 'klaar en distinct'. Betrekkelijk eenvoudige intuïties zijn vruchtbaar geweest voor onderzoek en inzicht, denk aan de evolutieleer, de genetica (dominante en recessieve genen) de zwaartekracht, Freuds driedeling van de psyche in Es, Ego en Superego en het gesprek als therapie. Ten *vierde* wordt aanbevolen open te staan voor onverwachte resultaten en te accepteren dat projecten niks opleveren. Veel hypothesen blijken onjuist. Deze bewering staat in contrast met wat er in psychologische tijdschriften gebeurt. Fanelli (2010, 2011) liet zien dat rond de 90% van de onderzoekshypothesen aanvaard worden. Zou

dit te maken hebben met selectief publiceren, met vermijden van riskant onderzoek met onzekere uitkomsten? De logica van het lineair geschreven verhaal correspondeert niet met het werkelijk verloop van onderzoek. Dat is immers niet logisch, serieel, eerder maar grillig. Ten *vijfde* wordt aanbevolen om slim met *reviewers* om te gaan en bijvoorbeeld een experiment al vast te doen waar zij om *zullen gaan* vragen. De auteurs moeten accepteren dat er vertraging is. De eerste aanbieding van een artikel en de publicatie kunnen soms meer dan een jaar uit elkaar liggen. Ook al is de verbetering klein, als er om gevraagd wordt, moet je het doen. Als je er toch tegenin gaat, doe het voorzichtig en diplomatiek. Ten zesde moet je er voor zorgen zichtbaar te zijn. Tijdschriftredacteuren zien je gemakkelijk over het hoofd en *reviewers* zitten niet op concurrenten te wachten. Je moet met *breaking news* komen om je in de kijker te spelen. Dat is ook riskant: denk aan de Eindhovense chemicus die het middel tegen AIDS dacht gevonden te hebben.

Onze bèta's hadden een formule kunnen maken die de aantrekkingskracht/zwaartekracht van een publicatie of verhaal meet. Waarom niet als denkexperiment en provocatie de omgekeerde kwadraatwet van Newton gebruikt. Als de afstand tussen jouw publicatie en die van een Harvard-, Berkeley-medewerker verdubbelt, wordt de aantrekkingskracht van jouw publicatie vier keer zo zwak. Hoe verder van een Harvard-, Berkeley-, Princeton- medeauteur, hoe geringer de kans op publicatie. Binnen de VS is de kans het grootst en bij het groter worden van de afstand neemt de kans kwadratisch af. Dat klopt waarschijnlijk niet en je kunt empirisch een predictieformule opstellen met cues die voorspellen hoe groot de kans is dat je artikel/onderwerp/onderzoeksaanvraag gepubliceerd wordt. Ik ben benieuwd aan welke cues onze bèta's denken. Psychologen willen wel langs empirische weg een predictieformule maken. Daar zijn ze goed in.

(3) En creativiteit dan... De bèta's hebben het niet over creativiteit. Misschien wordt dat overschat. Nieuwe inzichten, baanbrekend onderzoek kunnen voortkomen uit creativiteit en hard werken? Roskes et al. (2012) bestrijden dat deze een tweedeling vormen. Ze stellen een *dual pathway* voor: intuïtief, flexibel en moeiteloos naast volhardend, rationeel, stap voor stap. Beide kunnen tot een creatief, nieuw resultaat leiden. Ze concluderen dat rationele vermijdingsgemotiveerde deelnemers evengoed tot een creatieve oplossing kunnen komen als creatieve, associatieve op hun doel afgaande deelnemers. De strikte scheiding tussen willekeurig, rationeel en intuïtief is niet houdbaar. Maar creativiteit en hard werken zijn niet genoeg. Er wordt door onze bèta wetenschappers immers ook op gewezen, dat je moet weten hoe het wetenschapsbedrijf werkt. Er staat geen aanduiding van interessante thema's of van de *context of discovery* in. Een antropologische studie van het wetenschapsbedrijf, zoals Luyendijk (2014) heeft gedaan met de banken in Londen, zou hier niet misstaan. Ook nog een mooie predictieformule erbij, en je weet weer iets meer van de predictoren van de productie van het gemiddelde wetenschapsbedrijf. De aanvulling van cultureel antropologische *case studies* is er al, maar er zijn andere aanvullingen te bedenken. Welke cases zouden interessant zijn om onder de loep te nemen? Neem, om te beginnen succesvolle én mislukte, bijvoorbeeld het CERN in Genève, de chipbouwer ASML uit Veldhoven, De Spijker van Muller, enzovoort.

(4 Je niet op VS successen richten Wordt onderzoek kwalitatief beter door oriëntatie op VS tijdschriften? Moet ieder met een loopbaan in de psychologie voor ogen naar Harvard, Stanford, enzovoort? Moet ieder artikel in een Amerikaans tijdschrift? Feitelijk wordt erin geloofd en naar gehandeld maar er zijn tegengeluiden. De natuurkundige Ad Lagendijk (NRC, 10 november, 2014, p. 17) gelooft er niet in: *'...we richten ons op de VS. Alles zou daar beter zijn. Ik ben wel dertig keer in de VS geweest voor kortere en langere perioden. Ik heb daar nooit iets aangetroffen wat beter is dan op*

het vasteland van Europa'. Hij vergelijkt de VS werkwijze met de Franse. *'...Fransen zijn meer van de inhoud dan de presentatie'. '...jezelf verlagen om je bij buitenlanders voortdurend aan te prijzen als individuele onderzoeker of als academische instelling past niet bij de Franse cultuur...'*.
Dit geldt de facto niet voor de Nederlandse cultuur en handelsgeest. Psychologie is meer cultuurgebonden dan natuur- en wiskunde en daarom is lokaal onderzoek relevant. Europa is overigens onderling zo verdeeld dat dezelfde kritische houding tegenover een Europese buitenstaander ten toon gespreid wordt als Amerikanen ten opzichte van Europees psychologisch onderzoek. Daardoor loopt lokaal onderzoek achterstand op. Nationalisme doet zich met andere woorden niet alleen in de politiek voor. Daarnaast worden lezers soms behaagd in onderzoek, bijvoorbeeld adoptie en echtscheidingsonderzoek laat ons steeds weer zien dat adoptie ouders altijd 'beter dan gemiddeld' zijn en echtscheiding nooit goed. Je kunt ook onplezierige informatie weglaten, bijvoorbeeld het percentage Amerikaanse kinderen dat onder de armoedegrens, je benadrukt posttraumatisch *groei* of prijst kleine verschillen (p< . 01, <.05) aan als belangrijke bevindingen.
(5) De tegenaanval kiezen Een enkele keer is er de tegenaanval tegen de kritiek uit de VS, bijvoorbeeld zin er naar aanleiding van de aanval op *priming* experimenten 28 commentaren geschreven op een kritisch artikel van Newell en Shanks (2014). Eén is van Dijksterhuis et al.: *It* - het artikel van Newell en Shanks - *is neither a theoretical article, as it lacks a theory, nor it is a good review article, because it is biased and selective* en verder *the degree of cherry picking is too extreme* (p. 25). De auteurs blijven de rol van *Unconscious Thought* (UT) benadrukken en vinden steun in studies waar *Unconscious Thought Theory* (UTT) verbonden wordt met de *fuzzy trace theory* en met evidentie voor onbewust denken uit MRI scans. Ze besluiten met: *Although we surely agree that the road to progress in this field is rocky, focusing on consciousness without understanding its unconscious precursors is a dead end.* De buren in de wetenschap bijten elkaar af en toe. Niet slechts volgen om het succes, ook je eigen weg gaan en ertegenin gaan komen voor.

Samenvatting en conclusie

In je onderwijs is geleerd dat de kritische dialoog tussen gelijken vruchtbaar is in de wetenschap. Gelijken moet je maken en een gelijke moet je worden door een opleiding te geven enen te volgen waarbij ervaring en informatie loyaal gedeeld wordt. Op papier is dat zo. Als wetenschappelijke kritiek op je werk duidelijk is dan je zit er naast. Je accepteert het en betert je leven. Geen probleem. Zo gaat het niet in de psychologie. Ervaren publicisten zijn soms pragmatisch: doe wat de *reviewers* van tijdschriften willen. Maak het ze niet te moeilijk, geen lange verhalen en strijk ze niet tegen de haren in. De romantische eenling wil er nog wel tegenin gaan. Mr. Steve Jobs van Apple was daar een voorbeeld van. Je weet dat er weinigen zijn zoals Mr. Jobs, dus toch maar eieren voor je geld kiezen. William James was als filosoof pragmatisch: onderzoek zaken waar de mensen iets mee kunnen. Hij had ook een andere kant waarin hij het risico van je eigen weg gaan neemt en oproept tot originaliteit over thema's waar de grootste denkers al het hunne (m/v) over gezegd hebben. Nederlandse wetenschappers geven tips hoe door je aan te passen een grotere kans te maken op publicaties, maar gaan er ook tegen regels in. Soms in de vorm van een echte tegenaanval. Dan is er geen dialoog meer maar schofferen van de tegenstander. De

voorbeelden geven de (aankomende) wetenschapper bijna een dubbele boodschap. Hij moet met die ambiguïteit kunnen dealen. Dat wijkt niet af van veel andere activiteiten.
De professional heeft van doen met de kritiek vanuit de wetenschap. Dat is hiervoor aan de orde geweest. De praktijksetting waarin hij werkt kent ook allerlei criteria. Afwijking daarvan leidt tot kritiek op werkprestatie en houding. Criteria betreffen de productie, de kwantiteit, de kwaliteit, het volgen van protocollen enzovoort. Daar gelden uiteraard ook pragmatische kenmerken. Als je loyaal bent, niet overdreven veel initiatief neemt en de leiding niet bedreigt, voldoe je als werknemer.

6. Reflectie en evaluatie

Zolang psychologie en diagnostiek doen alsof ze een natuurwetenschap zijn zal haar het verwijt van het ontbreken van gedragswetten worden gemaakt. Theorievorming, operationalisatie en meten verlopen anders als in de natuurwetenschap. Er is daarop dan ook kritiek zoals in de voorafgaande hoofdstukken aan de orde is gekomen. Binnen de psychologie is vooral aandacht besteed aan de kwaliteit van het diagnostisch instrumentarium. Tests en vragenlijsten zijn het gemakkelijkst aan te pakken. Complexe procedures blijven buiten schot. Er is een beoordelingssysteem voor tests en vragenlijsten. Als je de beoordelingscriteria aanvaardt - en waarom niet, er is vooralsnog geen beter systeem - verkrijg je een beeld van de kwaliteit van het instrumentarium. Er zijn beslist wel veel onvoldoendes maar enkele veelgebruikte toetsen tests en vragenlijsten voldoen redelijk. De kwaliteit is de laatste 20 jaar niet toegenomen. Het is mogelijk dat testontwikkelaars en psychometrici hun limiet bereikt hebben of dat er onvoldoende middelen zijn om het voor verbetering nodige onderzoek te doen. Ik neem aan dat de grens bereikt is. De DSM is een systeem apart. Ondanks kritiek wordt het veel toegepast. Het blijft overwegend een categoriesysteem dat ervoor zorgt dat de gebruikers eenzelfde taal spreken. Professionals tonen en zijn bereid hun zwakke punten te erkennen. Bij onderzoekers treft je dat zelden aan maar ook zij kunnen zich er niet aan onttrekken. Die wordt heftig als er sprake is van fraude. Plagiaat is daarbij een kleinere zonde dan het fabriceren van gegevens. Het laatste is misschien wel het geval omdat je kennelijk gemakkelijk ‘aannemelijke resultaten’ kunt verzinnen zonder onderzoek doen. Is wat uit onderzoek komt een open deur? Dat is ernstige kritiek. Misschien moet de onderzoeker toegeven dat hij bijna nooit spectaculaire vondsten doet. Hij kan meningen, opvattingen van het weldenkend publiek misschien slechts nuanceren.
Statistici willen graag op methodologische gebreken van onderzoek wijzen. Dat is oplosbaar en in beginsel wil ieder onderzoeker zich aan methodologische regels houden. Als je de recentste *sophisticated* methoden niet hanteert wil dat niet zeggen dat je fouten maakt. Recht toe recht aan analyses en zelfs *‘eye balling’* leveren meestal geen resultaten die afwijken van resultaten uit moderne complexe designs en technieken.
Wat doe je aan wetenschapsfraude? Het stil houden kan niet meer en klokkenluiden is een gevaarlijke activiteit. Er zijn nu meldpunten. Studenten mogen anoniem hun grieven uiten

over docenten. Dat heeft ook zo zijn bezwaren. Gaat dat ook zo bij onderzoek van medewerkers? Waarschijnlijk niet. Sommige universiteiten hebben een meldpunt (een fraude ombudsman) aangesteld. Zijn rapportage leert wellicht over wat er aan fraude voorkomt en wat de aard is.

Het publiek is mondig en kritisch en reageert op de verwachtingen die wetenschappers en professionals scheppen. De laatsten maken reclame voor hun werk; neo-liberaal allemaal keurig maar ook met risico's. Het enige dat de onderzoeker en professional ertegenover kan stellen is open zijn over zijn werkwijze en de resultaten precies weergeven. Resultaat en succes zijn ook zonder complexe analyses te verduidelijken.

De cliënt staat centraal in de diagnostiek. In onderzoek is dat meestal niet het geval: *Scientia non est individuorum.* Die cliënt is geen heilige. De diagnosticus wil objectief recht doen aan zijn individueel verschillend profiel, stand van ontwikkeling en inbedding in zijn sociale context. Onderzoek dat laat zien dat sociaal wenselijke persoonskenmerken hoger uitpakken als een belang van de cliënt in het spel is. *Malignering* is een verschijnsel op zich. Het is niet zo gemakkelijk om dat aan te tonen.

Hoe gaan de professionals en onderzoekers om met de kritiek? Alle denkbare houdingen zijn wel eens getoond. Trek je er niets van aan als onderzoeker/professional, ga je eigen weg. Dat is voor weinigen weggelegd omdat de beroepsuitoefening plaatsvindt in instellingen met regels en belangen. Slechts een enkel bedrijf laat een groep of individu zijn gang gaan. Pas je zoveel mogelijk aan aan wat de instelling of het forum - vooral het Amerikaanse - van je vraagt. Er zijn 'sollicitatiecursussen' voor jonge onderzoekers en professionals (nog niet bij het UWV). Ga er met het gestrekte been in als je aangevallen wordt. Dat is een riskante aanpak, vooral omdat het moeilijk is tegen een instelling of bedrijf in te gaan. Wetenschappers nemen elkaar de maat. Omdat ze zichzelf meestal intelligent vinden - anders doceer je niet op de universiteit en doe je geen onderzoek toch - wordt de tegenstander beticht van bezig zijn met triviale zaken of saai en dom te zijn. Creativiteit wordt overschat zowel in onderzoek als de toepassing in de diagnostiek.

Het zou interessant zijn om na te gaan of in andere takken van wetenschap zo'n habitus heerst. Filosofen kunnen elkaar hard aanvallen. Historici kunnen strijd voeren om de 'juiste' interpretatie van historische gebeurtenissen. Bij natuurwetenschappers zie je naast intense samenwerking ook strijd.

Belangeloos de waarheid zoeken en succesvol zijn staan op gespannen voet. Bij Plato was er al de tegenstelling tussen het *theorein* en *sooidsein* (het object van onderzoek recht doen) van de filosofen en de pragmatiek van de sofisten. De laatsten verdienden hun geld met het publiek te leren hoe je debatten en onderhandelingen kunt winnen. Er zijn zelfs reclamespotjes in *prime time* over iemand die je leert anderen te overtuigen om daar beter van te worden. Er wordt niet bij verteld van wat je en wie er overtuigd moet worden. Waarheidsvinding is een doel van diagnostiek. Niettemin is onderzoek soms retorisch: de onderzoeker/diagnosticus wil iets bereiken bij de lezer/cliënt dat hem en de lezers goed uitkomt. We volgen de leider graag. Onderzoekers doen hun best Amerikanen te imiteren

en volgen nauwgezet het *Publication Manual of the American Psychological Association* (2010). Deze praktijk van onderzoeken en publiceren toont dat deze overzeese psychologie regeert en dat houdt volgens een citaat van P.J. Proudhon (een 19de -eeuwse anarchist; uit het Engels vertaald, 1923) in: 'bewaakt, geïnspecteerd, gereglementeerd, in een vakje gestopt, geïndoctrineerd, de les gelezen, gecontroleerd, beoordeeld, geteld, geformaliseerd, getarifeerd, geannoteerd, vermaand, berispt en verbeterd te worden'. Dit is onderzoekers bekend die een artikel naar een Amerikaans tijdschrift sturen. Wordt onderzoek kwalitatief beter door oriëntatie op VS tijdschriften? Moet ieder met een loopbaan in de psychologie voor ogen naar Harvard, Stanford, enzovoort? Moet ieder artikel in een amerikaans tijdschrift? Nee zegt Lagendijk. Psychologie is meer cultuurgebonden dan natuurkunde en daarom is lokaal onderzoek relevant. Europa is overigens onderling zo verdeeld dat dezelfde kritische houding tegenover een Europese buitenstaander ten toon gespreid wordt als Amerikanen ten opzichte van Europees psychologisch onderzoek, waardoor lokaal onderzoek ook daar achterstand oploopt. Nationalisme doet zich met andere woorden niet alleen in de politiek voor. Daarnaast worden lezers soms behaagd in onderzoek, bijvoorbeeld adoptieonderzoek laat ons steeds weer zien dat adoptie ouders altijd 'beter dan gemiddeld' zijn, of men laat onplezierige informatie weg, bijvoorbeeld het percentage Amerikaanse kinderen dat onder de armoedegrens leeft of men prijst kleine verschillen (p< . 01, <.05) aan als belangrijke bevindingen.

Onderwerpen en namen Hoofdstuk V

Kritiek op tests en vragenlijsten

Zeven rubrieken om een test te beoordelen

Kritiek op de DSM versies

Reflectie van diagnostici op hun werk

Wetenschapsfraude: plagiaat en gegevens verzinnen

Kritiek van cliënten op testen

Faken door cliënten

Malignering

Wijzen van omgaan met kritiek

Slotbeschouwing

Als je - zoals de psychologische diagnostiek - geen eigen inhoud en methode hebt, staan belendende vakken klaar om die aan je op te leggen. De methodologie legt het nodige op omdat onze spontane diagnoses zouden vertekenen. Vier thema's keren terug als het om het kritiseren - positief gezegd: in goede banen leiden - van de diagnose en het proces gaat:

(1) Het gaat over fouten bij voorspellen van (on)gewenst gedrag van de cliënt. De conclusie is dat de statistische voorspelling als winnaar uit de bus komt. 'Winnaar' is overdreven want de klinische predictie is bijna in de helft van de gevallen gelijkwaardig en de voorspelling geldt de steekproef en niet de cliënt. Het is niet eenvoudig het een of het ander.

(2) Het bezwaar is dat de diagnose onvoldoende biedt om tussen behandelingen te beslissen. Er is op gewezen dat diagnosen valse positieven en negatieven kennen; zo ongeveer in de helft van de gevallen. Als je dat tegen kansniveau afzet is het een behoorlijke score bij voorspellen en controleren van dat grillige menselijk gedrag. De helft van de variantie in criteriumgedrag binden, verklaren lijkt de limiet. En, de d-waarden van behandelingen komen zelden boven de 0,50. Beide delen (D en B) van de DBCs hebben zo hun beperkingen (halve waarheden en halve successen). Dan klinkt de eis van behandelaars om meerwaarde wat voorbarig. Eerst orde op eigen zaken stellen, kijk naar jezelf, maar dat is 'jij-bakken'. Het is de vraag of de cliënt (verzekeraar) dat een goede score vindt en geld voor diagnose en behandeling over heeft. De *evidence-based* eis voor van *treatments* is terecht en laat zien dat ook daar evenals bij diagnosen het glas zowel halfleeg als halfvol is.

(3) Instrumenten dienen betrouwbaarheid en valide te zijn De meeste instrumenten voldoen goed. Er is weinig bekend over betrouwbaarheid van andere ingewikkelder procedures en het diagnostisch proces. Daar is werk te doen. Testtheorie levert een kader om betrouwbaarheid en validiteit van items en tests te duiden. Ze wordt als een voorschrift opgevat en niet als een theorie om gedrag van een cliënt te verklaren of te interpreteren.

(4) Er is in algemene zin kritiek op de diagnose en het diagnostisch proces en op de manier waarop de diagnosticus daarmee omgaat. Dat varieert van meegaan met de critici en *reviewers* tot er met gestrekt been tegenin gaan. Dat verschilt niet van conflicten in het dagelijks leven. Het beeld van de slecht betaalde wetenschapper die op zijn kamertje of in het laboratorium aan iets nieuws werkt is valse romantiek. Het gaat om reputatie, invloed en macht. Ik werk de onderwerpen van dit boekje nog eens uit:

De controverse statistische versus klinische predictie kent als basis het verschil tussen nomothetisch en idiografisch. Alle verschijnselen kunnen met beide benaderingen bestudeerd worden volgens Windelband. De psychologie nam evenwel een plaats in de ene of andere categorie, bijvoorbeeld eerst nomothetisch van Freud tot Watson en Skinner maar ook idiografisch (Allport) en zoals in casusbeschrijvingen. Binnen de psychologie werd het een tegenstelling waarbij de clinicus de statisticus een neerbuigende houding verweet en de clinicus de statisticus pedant noemde. De tegenstelling is er ook nu nog door de

pikorde binnen de psychologie. De functiepsycholoog zegt dat psychonomie verhoudt zich tot psychologie als astronomie tot astrologie.
Het klinisch oordeel kwam onder druk te staan door Meehls (1954) *disturbing little book*. De zwakte van het klinisch oordeel werd experimenteel onderzocht door Tversky en Kahneman. Zij starten de *Heuristics* en *Biases* traditie. Het (on)gewapend oordeel leidt tot veel fouten. Daar kwamen studies bij die lieten zien dat statistisch prediceren minder fouten maakte. Het was de *bedoeling* te laten zien hoe slecht ons ongewapend oordeel het doet. Denk aan Van Dams monografie *Fixatie op fouten.* In meta-studies werd *Clinicia* met *Academia* vergeleken. De boodschap luidt: clinici en diagnostici zijn brokkenmakers, vertrouw op de formule. De verdeling was niet 0 voor de clinicus en 10 voor de statisticus. Er was wel een lichte winst de statistische predictie op het niveau van de steekproef. Dit lokte verzet uit bij clinici maar dat hielp niet.
Een bijvangst van de controverse was de studie van het klinisch oordeel. Brunswiks lensmodel gaat uit van een adaptieve oordelaar. Hij kan ernaast zitten maar door informatie van anderen en door eigen ervaring leert hij: het probabilistisch functionalisme. Kahneman werd later ook iets milder: er zijn *wicked environments* waarin je er met je voorspelling naast zit, bijvoorbeeld speculeren op de beurs. Bovendien is er de studie van *Natural Decision Making* door experts, bijvoorbeeld brandweerofficieren of artsen bij ongelukken. Ze kunnen niet anders als onmiddellijk handelen op grond van de *cues* die ze tot hun beschikking hebben. Gigerenzer zocht en vond voorbeelden waar het ongewapende oordeel zelfs beter voorspelde dan het modelmatige. Het onderzoek doet vergezocht aan maar het is wellicht bedoeld als Popperiaanse tegenvoorbeeld. Daar moet een aanhanger van het cruciaal experiment - en dat is Kahneman *after all* - iets mee.
Er kwam door het werk van Brunswik en Gigerenzer weer *lucht voor het klinisch oordeel*. Het werd als een adaptief verschijnsel onderzocht en er kwam ruimte voor oordelen in de alledaagse werkelijkheid buiten het laboratorium. De tijd was rijp voor iets anders. Meestal is dat in de psychologie iets ouds: de intuïtie. Heidegger zei al in zijn baanbrekende werk *Sein und Zeit* (1928/1962, 6^{e} druk) dat zijn boek niet nieuw was maar dat hij blij zou zijn *wenn es alt genug währe.* Intuïtie kwam terug als onderwerp door de Hogarth, een voormalig aanhanger van de HB traditie. Daarmee is het normatieve rationalistische model meestal in de vorm van een lineaire optelling van valide *cues* wat teruggedrongen. De controverse is complexer geworden. Het is te eenvoudig om de deelnemers als dom versus slim tegenover elkaar te zetten. Er is oog voor het adaptieve proces van de clinicus en hij is door het HB onderzoek een gewaarschuwd man (m/v) geworden en die telt voor twee. Het klinisch oordeel komt (een beetje) terug, bijvoorbeeld in de DSM-5 waar de ernst van een stoornis klinisch beoordeeld moet worden. Overigens zou daar de psychometrische schaling weer een functie kunnen hebben.
De diagnosticus maakt gebruik van zijn opgebouwde *intuïtie,* verkeert in een *context van discovery* over wat nu de vraag of het probleem is, hanteert soms snelle heuristieken: ‘als A aan de hand is dan meteen X doen’, gaat rationeel stapsgewijs te werk en doordenkt de

stappen, doet aan patroonherkenning (zo *gereframed* om van het brokkenmakers-imago van gevoelsmatige in zijn hoofd combinerende beslisser af te komen) en combineert bij de cliënt covariaties tussen individuele kenmerken en zijn ontwikkeling en sociale context. Dit houdt in dat hij streeft naar maatwerk voor zijn cliënt.

Diagnose-Behandel-Combinaties beloven ons maatwerk. Weer een controversiële thema. Er zijn congressen met de belofte goede DBCs te laten zien. U kunt er in de winter van 2015/2016 naar toe in mooie Zwitserse oorden. Een titel als 'MAATWERK, de weg naar *precision therapy* voor uw patiënten' laat zien hoezeer de DBC - in dit geval met de nadruk of de B - in de belangstelling staat. Ja, de ene patiënt is de andere niet. Wanneer psychologen het initiatief nemen tot een dergelijk congres ligt de nadruk op de D zonder de behandeling uit het oog te verliezen. Je kunt vaststellen dat voor het uitzoeken en onderzoeken van het effect van DBCs nog werk te doen is. De praktijk loopt voorop in de toepassing. Bestuurders zien DBCs wel zitten want ze rekenen zich rijk met goede diagnoses en passende behandelingen. Geen overdiagnose en overbehandeling meer! Het ligt intuïtief voor de hand. Niettemin is maatwerk op basis van ATI onderzoek en integratie van theoretische oriëntaties niet serieus van de grond gekomen. De dominantie van het experiment is gebleven zoals blijkt uit de nadruk op *evidence-based* behandelingen. Antipestprogramma's zijn daar een voorbeeld van. De kinderombudsman wil wildgroei in deze programma's beperken. Er zijn er meer dan 100 van in Nederland. Er moeten een paar bewezen effectieve overblijven. Hij zegt daarbij te steunen op wat de wetenschap - zijn adviseurs - hem verteld heeft. Dat is één element van de DBC: de kwaliteit van behandelingen. Om effecten van de Bn te bewijzen is er in de empirisch-analytische traditie maar één weg: het *true experiment*. Fisher heeft dat uitgelegd met veldjes groente. De binnenvariantie moet in de hand gehouden worden en de tussenvariantie moet op basis van de experimentele variabele het verschil maken. De binnenvariantie wordt groter naarmate de steekproef personen op allerlei variabelen verschilt. Ke kunt ze proberen weg te filteren met covariaten maar in werkelijkheid kun je de groepen niet zo kiezen dat ze covariaatvrij zijn. En, je doel is meestal niet iets te ontwerpen voor een zeer specifieke homogene groep. Je maakt het programma voor een grote gemêleerde groep en voor de specifieke cliënt. Je sluit geen meisjes of jongens, ouderen of jongeren, lage-hoge SES groepen uit. Dat zijn de bekende covariaten. Het experiment is geschikt om utilitaristisch vast te stellen wat het beste programma is voor een zo groot mogelijke groep. We kennen de d-waarden en dat leidt tot bescheidenheid over de effecten van de Bn. En, dan weten we nog niet bij welke kenmerken van de cliënt en zijn context die effectieve B past.

Is er verbetering mogelijk? Misschien worden DBCs te complex opgezet: te veel covariaten, te veel risico en beschermende factoren, en qua duur en omvang te complexe behandelingen. Er kan onderweg zo veel uit de hand lopen dat je door de bomen het bos niet meer ziet. Lukt het met bescheidener opzetten om iets meer over een DBC te weten te komen? Misschien is een variant op adaptief testen een mogelijkheid om greep en zicht te krijgen op DBCs. Het wordt gedaan met *Rekentuin* en *Taalzee* van de UVA. Ze zijn te vinden

op: http//:www.oefenweb.nl. De afhankelijke variabele kan geschaald worden als een individuele verschillen dimensie van een groep of als een ontwikkelingsdimensie van een leerling. De behandeling bestaat uit veel precies gekozen en uitgeteste opdrachten die passen bij de plaats op de latente trek of de ontwikkelingsschaal van de individuele leerling. Je hebt er een grote voorraad aan precies metende taken en opdrachten voor nodig. Als dit slaagt, kun je langzaam naar meer covariaten (diagnose) en samengestelde programma's (behandelingen) gaan.

Misschien is het mogelijk dat zorgvuldig opgezette en gemonitorde *case studies* stukje bij beetje zicht geven op DBCs. Deze kunnen gegeneraliseerd worden naar soortgelijke gevallen. Je komt ze in tijdschriften niet tegen. Gigerenzer (2009) noemt dit een verlies. Hij vertelde dat hij enkele artikelen in het *Journal of Experimental Psychology* van de jaren 20 tot 30 van de vorige eeuw had gelezen. Hij merkte op dat er iets verloren is gegaan: verschillende (statistische) methoden, precieze rapportage van individuele gevallen, zorgvuldige selectie van proefpersonen en nagaan of de sekse van de proefpersoon of proefleider van invloed was. Dit zijn precies de covariaten die er bij voorbaat uitgehaald worden om succes van een behandeling te bepalen. Ondertussen kun je practici, diagnostici en behandelaars aanmoedigen door te gaan met zoeken naar de succescombinatie bij individuele cliënten. Er valt van hen iets te leren als ze *vastleggen* hoe en waarom ze de DBC in een concreet geval zo vormgeven hebben en wat het resultaat was op korte en middellange termijn. De werkgever/organisatie zou hen kunnen toestaan om één casus per maand zo volledig mogelijk uit te werken, te bespreken en te rapporteren. Daar horen ook *niet* succesvolle DBCs bij. Misschien is er zelfs een tijdschrift te maken dat ze wil publiceren. Je moet de moed hebben om je niets aan te trekken van de slagzin: *Scientia non est individuorum*. Aan de hand van DBCs van individuele cliënten kan een kennisbestand opgebouwd worden. Natuurlijk is dit eerder geprobeerd en het resultaat was beperkt, maar er was ook een andere gestemdheid: de diagnostici en behandelaars doen maar wat. Toch kun je pleiten om het met de kennis van nu nog eens te doen.

De overtuiging en de intuïtie van bestaan van synergie tussen diagnose en behandeling blijven, ook al heb je geen bewijs. Het verdient uitgezocht te worden. Je kunt daarbij buiten het *true experiment* gaan. DCBs zijn in de medische wetenschap aanvaard en daar spiegelt een diagnosticus zich aan. DBCs zijn ook aantrekkelijk om de bijna marxistische gedachte, dat we wel veel weten, maar er weinig mee doen. Is kennis er om de wereld te veranderen, je gedrag te verbeteren? Ja, een beetje en het lukt misschien ook wel een beetje.

De diagnosticus moet zijn plaats bepalen en steeds meer veroveren in het DBC debat. De DBC gaat over de *relatie* tussen Ds en Bn en over de relatie tussen diagnostici en behandelaars. Beide(n) hebben te maken met hun kennisbestanden. Deze dwingen tot bescheidenheid over de resultaten en tot twijfel. Bij *daten* moet je echter niet te veel twijfelen om op stoom te komen. Je moet de ander de ruimte geven. Als de behandelaar geen spoor van twijfel vertoont over zijn behandeling (denk aan psychiaters) en verwacht dat de diagnosticus voedsel verstrekt om hun behandeling te verbeteren, hoef je geen

synergie te verwachten. Dit was de houding van de psychiater/behandelaar Kuiper ten opzichte van Kouwer, de diagnosticus/twijfelaar naar aanleiding van zijn boek *Het Spel van de Persoonlijkheid*. Behandelaars willen wellicht te snel tot actie overgaan. Het is bijna een reflex: probleem dan behandeling. Als de diagnosticus zelf te zeer twijfelt over zijn diagnose, kan hij de *date* vergeten. Je gaat niet met een potentiële *loser* in zee. Als de behandelaar nauwelijks gelooft in zijn interventie, voert hij die niet of kwansuis uit. Hij heeft nochtans enig - niet al te naïef - geloof nodig om op gang te komen en te blijven. Hij doet er van alles aan: interventies maken en benutten, zoeken naar effectieve ingrediënten, uitkijken voor iatrogene effecten en ontsporing en *true experiments* en klassieke n = 1 studies uitvoeren om meer aan de weet te komen.

Om met de metafoor van relatie te besluiten: D en B beoefenaars hebben kans van slagen als elk erkent niet perfect te zijn, elk niet de dominante partij wil zijn op basis van schoonheid en *good breeding* (pikorde in de wetenschap). Elk streeft synergie na zonder volkomen in elkaar op te willen gaan, maar ook niet strikt hun eigen weg willen gaan. Ze kunnen af en toe door één deur gaan. Ze hoeven zich niet mooier voor te doen dan ze zijn.

Betrouwbaarheids - en validiteitstheorie zijn grote woorden voor herhaalbaarheid van scores op verzamelingen items en de dekking die ze bieden voor een gedragsbegrip en de voorspellende waarde. Het valt maar voor een beperkt deel samen met wat er gebeurt en nodig is voor een diagnose. Die gaat als *ultimate criterion* om een ware, geldige uitspraak over het gedrag van een cliënt. Psychologen maken zich niet druk over hun wijsgerig verleden. Ze hebben de logisch positivistisch epistemologische stijl aanvaard met inbegrip van de empirisch-analytische wijze van verwerven van kennis over gedrag. Dat heeft de psychologie tot een serieus vak gemaakt. Dé waarheid bestaat echter niet. In de wetenschap kan dat om verschillende zaken gaan, bijvoorbeeld om een logisch consistente structuur die in eerste instantie de werkelijkheid niet hoeft te beschrijven: je mag van je eigen ogen het ontzaglijk glanzen beschouwen en als de spin uit je eigen lijf de prachtigste webben (theorieën) weven. Empirische wetenschappers moeten daarentegen een verband verzinnen en bewijzen tussen de constructen, formules, structuren en de werkelijkheid van verschijnselen, objecten en vooral gedrag. Je hoeft maar een stap buiten de psychologie te zetten of er is een andere opvatting. Waarheidsvinding betekent bijvoorbeeld in de Nederlandse rechtspraak het leveren van het wettig en overtuigend bewijs, dat X het delict gepleegd heeft. X kan ontkennen, maar als dit bewijs geacht wordt geleverd te zijn, is hij schuldig. De rechter is vooral beducht voor de fout dat hij iemand veroordeelt die onschuldig is. Hij laat X liever lopen ook al vermoedt hij dat X het delict gepleegd heeft. Hij krijgt het dan niet voor elkaar dat wettig en overtuigend bewijs te leveren. In de diagnostiek wegen valse positieven en valse negatieven even zwaar. In China behoort de bekentenis van de verdachte tot de bewijsvoering. De politie zegt dat 99,5% van de verdachten bekent. Hij vertrouwt op zijn verhoortechniek, inclusief de leugendetector. In de VS gaat het weer anders. Je kunt daar bekennen om strafvermindering te krijgen, ook als je onschuldig bent.

De verschillende procedures en de uiteenlopende strafmaten berusten op gedachten en verwachtingen hoe je de waarheid over een misdaad boven water krijgt.
Omdat er niet één waarheid is, houdt elke tak van wetenschap en zelfs iedere persoon er een op na. De bewering van communisten dat er één waarheid (*Pravda*) is of géén waarheid is te begrijpen als wens en dictaat van de macht maar ze is precies daarom onjuist en onethisch. Ook met Duijkers klassieke opmerking: 'Er is één psychologie of er is geen psychologie', is mijn inziens onhandig. Zijn quote spoort met het eerste scholastische logische axioma van non-contradictie: 1.1: *'It is impossible that the same thing be and not be at the same time and in the same respect'*. (de axioma's zijn van het internet te downloaden: *A scholastic list of philosophical axioms*). Feitelijk houden psychologen er heel verschillende opvattingen op na, denk aan de verschillende wijsgerige stromingen, die ze impliciet volgen. Ze bestrijden elkaar wel degelijk en ze hoeven het eigen ongelijk niet te erkennen. Wetenschappers, denk aan de alfa's en de bèta's betwisten elkaars waarheidsclaims. Ieder heeft zijn waarheid, geldigheid. Dat mag je zeggen zonder een radicale relativist te zijn.
Kun je niet beter als de pragmatist zeggen dat er geen antwoord op die vragen is. Ga over tot de orde van de dag! Zoek dingen uit die iets betekenen, waar de mensen iets aan hebben. Of, kies na een beetje nadenken en overleg een werkdefinitie en zeur niet. Of, doe empirisch onderzoek, dan kom je jezelf tegen en zuiver je de begrippen uit. Of, is het gesteld als met het *kritische* en *praktische Vernunft*: de dingen *an sich* kunnen we niet kennen, maar dank zij ons *praktische Vernunft* kunnen we er aardig mee uit de voeten?
Leken, professionals en onderzoekers kunnen niet zonder hun waarheidsopvatting. Eerst is dat een emotioneel belang. Leken, diagnostici en professionals willen daarnaast ook objectief handelen en de waarheid dienen. Niemand, of bijna niemand, wil constant bedriegen en in ieder geval niet doorlopend bedrogen worden. Dat houdt voor leken in: de waarheid spreken, hoe vaag dat ook klinkt. Wetenschappers willen onderzoek doen dat iets zinvols aan het licht brengt. Ze houden daarbij methodologisch hun emoties op afstand maar zijn wel degelijk betrokken bij het onderwerp en bij de geldigheid van hun onderzoeksresultaten. Diagnostici willen rapporten schrijven met diagnoses die de cliënt recht doen. We willen geen reclame horen maar zinvolle en juiste informatie. Waarheid is naast de directe emotie aan waarden gebonden. Opvattingen van leken en professionals dienen een kennisbelang. Het belang kan beheersing/controle, het verwerven van een sociale positie, het verkrijgen van begrip of de emancipatie van de ander zijn. Er is ook een waarheid van de economische en sociale jungle. Dat belang ligt op steeds de loer in een neo-liberale utilitaristische, hedonistische samenleving. Als ik zeg: het ligt op de loer, impliceer ik, dat je dat belang moet vermijden. Dat beweer ik, als ik - dat zeggende - weggezet worden als moraalridder of als *sanft*, *soft*. Ik mag me afvragen: Moet ik U als wetenschapper serieus nemen, als u reclame maakt voor uw eigen winkel en uw werk in dienst staat van uw reputatie? Moet ik geld geven aan de Maag-darm-lever Stichting ('het maag-lever-darm-*alarm*') omdat geld uw noodzakelijk onderzoek mogelijk maakt naar

oorzaken en effectieve medicijnen? En, aan de Hart- en Hersenstichting? Noodzaak wordt in radiospotjes letterlijk genoemd. Voelt de bestuurlijke gezondheidszorg-elite die noodzaak dan niet? Moet het Koningin Wilhelmina Fonds mensen een berg laten beklimmen om 'noodzakelijk onderzoek' te doen? Zonder die tocht geen genezing? Of gaat het om belangen, zoals het overleven van onderzoeksinstellingen, werkgelegenheid en het in stand houden van een structuur met een door de samenleving gewaardeerde en een goedbetaalde directie?
De moraalfilosoof Williams (2002) ziet sporen van belangen bij sociale wetenschappers: Macht en reputatie zijn:

'... *one of the reasons, why, at the present time, the study of humanities runs a risk of sliding from professional seriousness, through professionalization to a finally disenchanted* (betovering wegnemend, ontluisterend, ontgoochelend) *careerism'.*

Hij is bekend vanwege zijn ironische blik op het utilitarisme dat volgens hem uitgaat van een onpartijdige waarnemer die alwetend, belangeloos en zonder emotie is, maar verder een 'normaal mens' is. Williams onderscheidt naast *accuracy:* de objectieve, valide beschrijving, *sincerity:* oprechtheid. Dat houdt onder meer in het afwijzen van *free riders*, het anderen niet willen bejegenen vanuit eigenbelang en in staat zijn tot geven en ontvangen. Deze begrippen impliceren weerstand tegen *wishful thinking*, zelfbedrog en fantasie. *Sincerity* is het nastreven van objectiviteit.
De diagnosticus is geïnteresseerd in de waarheid, in een geldige, valide diagnose. Hij heeft het kennisbestand van de psychologie als ruggensteun bij zijn streven recht te doen aan het probleem, de vraag, het gedrag van de cliënt. Dat bestand is vooral binnen het logisch-positivistisch, empirisch-analytisch kader verworven en uitgebouwd. Dat is omvangrijk, maar hij kan er niet om heen: hij moet 'tegen beter weten in', 'tegen de feiten in' alles weten. Aan het object van onderzoek recht doen, vraagt ook dat de diagnosticus open staat voor betekenissen van gedrag. Hij moet bij de cliënt achterhalen waar de schoen wringt.
Objectiviteit en waarheid hebben vele vaders, moeders, stromingen en draagvlakken nodig. Ze zijn gekenmerkt door kennis van zaken, redelijkheid, zuiverheid, deugdelijkheid, gerechtvaardigdheid en houdbaarheid van beweringen over de sociale werkelijkheid en het gedrag van de cliënt. Als je kiest voor objectiviteit als 'aan het object van studie recht doen', kun je je niet beperken tot de methodologische empirisch analytische kenhouding. Diagnostiek en ook psychologie kunnen zich niet ontdoen van hermeneutiek. Dat houdt in dat ook de diagnosticus de idiosyncratische betekenisverlening van de cliënt serieus neemt. Een minder prominente, meer gewantrouwde stroming is de kritische. Deze staat een andere opvatting van de cliënt toe: recht doen is hem emanciperen. Er hoort een andere methode bij: het verfoeide actie onderzoek; maar wees voorzichtig: er zijn nogal wat handelingswetenschappen met serieuze en gerespecteerde beoefenaars.

Testtheorie valt onder de afdeling M & T. Sommigen zien deze afdeling als dienstverlening, anderen als een discipline met een eigen domein en methode. Hier is testtheorie opgevat als producent van een enkele mini-theorie over een mini-gedrag. Volgens atomistisch ingestelde onderzoekers is dit de enige weg. De KTT en IRT hebben dat veld afgebakend en geclaimd. Ze ontwikkelen zich los van andere technieken die ons iets leren over itemkenmerken en samenhangen. Fischer, de opvolger van Rasch in Wenen heeft de IRT verbreed en langdurig onderwezen. Hij zag in dat de theorie het wel deed op de universiteit maar in de praktijk in zijn land weinig voet aan de grond kreeg. De moderne testtheorie is bijzonder en geeft inzicht in kenmerken van items - het Cito aanvaardt alleen test en toetsen die aan een eigen IRT model voldoen - maar voor de doel van predictie levert ze om een cliché te gebruiken geen meerwaarde boven klassiek geconstrueerde tests. De IRT staat midden in het debat over wat een theorie moet zijn: een formeel model of een semantisch netwerk van hypothetische constructen. Het verwijt is dat KTT en IRT niet over inhoud gaan. Ze zijn geen theorie maar een verzameling functies (IRFs, ICCs)'.

Voor de diagnosticus bevat testtheorie kwaliteitsindicatoren van instrumenten. Daartoe blijft het beperkt en zij biedt geen nieuwe inzichten in gedragingen. KTT en IRT zijn ook buiten de psychologie nuttig, bijvoorbeeld voor meten in geneeskundig en farmacologisch onderzoek.

Kritiek op psychologie en diagnostiek is methodologisch en inhoudelijk en komt van buiten (methodologen, statistici) van andere wetenschappen (de hoger in de pikorde staande natuurwetenschappen) maar ook van binnen. Er zijn studies naar reflecties van diagnostici op hun werk. Hoe met de kritiek om te gaan? Het lijkt dat machtsverhoudingen en reputatie een centrale rol spelen. Onderzoekers volgen wat Engelstalige tijdschriftredacties willen en er is ook een enkele keer verzet. Dit wijkt niet af van het verloop van andere conflicten. Het lijkt eerder op een politiek conflict of een strijd tussen bedrijven in een gevecht om de markt dan op een huiselijk conflict waar verzoening of compromis meestal het doel is.

Referenties en geraadpleegde literatuur

Hoofdstuk I

Aegisdótttir, S., White, M.J., Spengler, P.M., Maugherman, A.S., Anderson, LA., Cook, R.S., Nichols, C.N., Langopoulos, G.K., Walker, B.S., Cohen, G. & Rush, J.D. (2006). The meta-analysis of the Clinical Judgment Project: The fifty-six years of accumulated research of clinical versus statistical prediction. *The Counseling Psychologist, 34*, 3, 341-382.

All port, G.W. (1937). *Personality, a psychological interpretation*. New York: Holt, Rinehart & Winston.

All port, G.W. (1942). The use of personal documents in psychological science. New York: *Social Science Research Council*.

Baker, J. (2010, 3de druk). *Vijftig inzichten in de natuurkunde: Onmisbare kennis.* Diemen: Veen Magazines BV.

Berkel, H.J.M. van (1984). *De diagnose van toetsvragen.* Academisch proefschrift: Universiteit van Amsterdam.

Bloem, J.C. (1982) *Verzamelde gedichten.* Amsterdam: Athenaeum: Polak & Van Gennep.

Bolstad, W.M. (2007). *Introduction to Bayesian Statistics.* New York: J. Wiley.

Brinkmann, S. (2009). Facts, values, and the naturalistic fallacy in psychology. *New Ideas in Psychology, 27*, 1-17.

Cautin, R.L. (2011). Invoking history to teach about the scientist-practitioner gap. *History of Psychology, 14,* 2, 197-203.

Chassy, Ph. & Gobet, F. (2011). A hypothesis about the biological basis of expert intuition. *Review of General Psychology, 15,* 3, 198-212.

Cohen, J. (1988). *Statistical power analysis for the social sciences (revised edition*). New York: Academic Press.

Costello, F. & Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological Bulletin, 121,* 3, 463-480.

Dam, K. van (1991). *Fixatie op fouten*. Lisse: Swets & Zeitlinger.

Dawes, R.M., Faust, D. & Meehl, P.E. (1989). Clinical versus actuarial judgment. *Science, 243*, 1668-1674.

Dawes, R.M., Faust, D. & Meehl, P.E. (1993). Statistical Prediction versus Clinical prediction: Improving what works. ln: G. Keren &C Lewis (Eds.). *A handbook for data analysis in the behavioral sciences.* Hillsdale NJ: Erlbaum.

Dawes, R.M. (1994). *House of cards, Psychology and Psychotherapy built on myth*. New York: The Free Press.

Dawes, R.M., (2000). *Irrationality: Theory and Practice*. New York: The Free Press.

Dawes, R.M. (2005). The ethical implications of Paul Meehl's work on comparing clinical versus statistical prediction methods. *Journal of Clinical Psychology, 61,* 10, 1245-1255.

De Bruyn, E.E.J., Ruijssenaars, A.J.J.M., Pameijer, N.K. & van Aarle, E.J.M. (2015). *De diagnostische cyclus: een praktijkleer.* Leuven/Den Haag: Acco.

Dhami, M.K. & Harries (2001). Fast and frugal versus regression models of human judgment. *Thinking and Reasoning, 7*, 5-27.

De Groot, A.D. (1961). Via clinical to statistical prediction. *Acta Psychologica, 18*, 274-284.

De Groot, A.D. (1946, 1978). *Thought and choice in chess.* The Hague: Mouton.

Dougherty, M.R. & Thomas, R.P. (2012). Robust decision making in a nonlinear world. *Psychological Review.* Advance online publication. Doi: 10.1037/a0027039.
Draaisma, D. (2013). *De Dromenwever.* Groningen: Historische uitgeverij.
Diagnostic and Statistical Manual of Mental Disorders. DSM-5™ (5th Ed.). American Psychiatric Association (2013). Washington DC: APA.

Einhorn, H.J., Kleinmuntz, D.N. & Kleinmuntz, B. (1979). Linear regression and process-tracing models of judgment. *Psychological Review, 86*, 4, 465-485.
Einhorn, H.J. (1986). Accepting error to make less error. *Journal of Personality Assessment, 50,* 3, 387-395.

Faust, D.W (1997). Of science, meta-science, and clinical practice: the generalization of a generalization to a particular. *Journal of Personality Assessment, 68*, 331-354.
Freeman, J.B. & Ambady, N. (2011). A dynamic interactive theory of personal construal. *Psychological Review, 118,* 2, 247-279.
Frijda, N. (1986). *The Emotions*. Cambridge: Cambridge University Press.

Ganzach, Y (1995). Non-linear models of clinical judgment: Meehl's data revisited. *Psychological Bulletin, 108*,3, 422-429.
Garb, H.N. (1998). *Studying the Clinician. .Judgment Research and Psychological Assessment*. Washington, DC: American Psychological Association.
Garb, H.N. (1995). Using computers to make judgments: correlations among predictors and the comparison of linear and configural models. *Computers in Human Behavior, 11*, 2, 313-324.
Garb, H.N. (1989). Clinical judgment, clinical training, and professional experience. *Psychological Bulletin, 105,* 3, 387-396.
Garb, H.N. (2005). Clinical judgment and decision making. *Annual Review Clinical Psychology*, pp. 67-89.
Gigerenzer, G. (2008). *Rationality for mortals. How people cope with uncertainty*. Oxford: Oxford University Press.
Gigerenzer, G. & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology, 62*, 451-482.
Goldberg, LR. (1971). Five models of clinical judgment: an empirical comparison between linear and nonlinear representations of the human inference process. *Organizational Behavior and Human Performance, 6*, 458-479.
Goldstein, D.G. & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review, 109*, 1, 75-90.
Gore, J. & Sadler-Smith, E. (2011). Unpacking intuition: a process and outcome framework. *Review of General Psychology, 15,* 4, 304-316.
Grove, W.M. & Meehl, P.E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: the Clinical-Statistical Controversy. *Psychology, Public, Policy, and Law, 2,* 2, 293-323.
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*, 1, 19-30.

Hampton, J.A. (2012). Thinking intuitively: The rich (and at times illogical) world of concepts. *Current Directions in Psychological Science, 21,* 6, 398-402.
Harris, G.T., Rice, M.E. & Cormier, C.A. (2002). Prospective replication of the violence risk appraisal guide in predicting violent recidivism among forensic patients. *Law and Human Behavior, 26,* 4, 377-394.
Hastie, R. (2001). Problems for judgment and decision making. *Annual Review of Psychology, 52,* 653-683.
Hermans, H.J.M. (1988). On the integration of nomothetic and idiographic research methods in the study of personal meaning. *Journal of Personality, 56,* 785-812.
Hilbert, M. (2012). Toward a synthesis of cognitive biases: how noisy information processing can bias human decision. *Psychological Bulletin, 138,* 2, 211-237.
Hogarth, R. (1987, 2nd edition). *Judgment and Choice*. New York: J. Wiley and Sons.
Hogarth, R. (2001). *Educating intuition*. Chicago, Chicago University Press.
Hogarth, R.M. (2010). Intuition: A challenge for psychological research on decision making. *Psychological Inquiry, 21*, 338-353. (zie ook *Psychological Inquiry*: whole volume *21*, 4, 2010 about Intuition).
Hoffman, P.J. & Wiggins, J. (1968). Cue-consistency and configurality in human judgment. In: B. Kleinmuntz, (Ed.). *Formal representation of human judgment*. New York: J. Wiley.
Holt, R.R. (1970). Yet another look at clinical and statistical prediction: or is clinical psychology worthwhile? *American Psychologist, 25,* 3, 337-354.
Holt, R.R. (1986). Clinical and Statistical prediction. A retrospective and would-be integrative perspective. *Journal of Personality Assessment, 50,* 3, 376-386.
Honderich, T. (2005). *The Oxford Companion to Philosophy (New Edition).* Oxford: Oxford University Press.
Hurlburt, R.T. & Knapp, T.J. (2006). Münsterberg in 1898, not Allport in 1937 introduced the terms idiographic and nomothetic to American psychology. *Theory and Psychology, 16,* 2, 287-293.

Inbar, Y. Cone, J. & Gilovich, Th. (2010). People's intuitions about intuitive insight and intuitive choice. *Journal of Personality and Social Psychology, 99*, 2, 232-247.

Kahneman, D. (2003). A perspective on Judgment and Choice; mapping bounded rationality. *American Psychologist, 58,* 9, 697-720.
Kahneman D. & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 2, 237-251.
Karelaia, N. & Hogarth, R.M. (2008). Determinants of linear judgment: a meta-analysis of lens model studies. *Psychological Bulletin, 134,* 3, 404-426.
Katsikopoulos, K.V. (2009). The conceptual connection between lens models and fast and frugal heuristics. *Theory and Psychology, 19,* 5, 688-697.
Katsikopoulos, K.V., Pachur, T, Machery, E. & Wallin, A. (2008). From Meehl to fast and frugal heuristics (and back). *Theory and Psychology, 18*, 4, 443-464.
Kirmayer, L.J. (1994). Improvisation and authority in illness meaning. *Culture, Medicine and Psychiatry, 18,*183- 214.
Klein, G. A. (1998). *Sources of Power: How people make decisions*. Cambridge. MA: MIT Press.
Klein, G.A. (2003). *Intuition at work*. New York, NY: Currency & Doubleday.
Kleinmuntz, B. (1963). (Ed.). MMPI decision rules for the identification of college maladjustment: A digital computer approach. [Special issue]. *Psychological Monographs, 77,* 577.

Kleinmuntz, B. (1990). Why we still use our heads instead of the formulas? Towards and integrative approach. *Psychological Bulletin, 107*, 296-310.
Korman, A.K. (1968). The prediction of managerial performance. *Personnel Psychology, 21,* 295-322.
Krol, G. (1978, 2e druk). *Het gemillimeterde hoofd.* Amsterdam: E. Querido's Uitgeverij BV.
Kruglanski, A.W. & Gigerenzer, G. (2011). Theoretical Note: Intuitive and deliberate judgments are based on common principles. *Psychological Review, 118,* 1, 97-109.

Laak, J.J.F. (2011). *Elementair begrip van de psychologisch diagnostiek.* Amsterdam: Pearson.
Laak, J.J.F. ter (2015).*Psychologische diagnostiek is diagnostiek van de psychologie.* Utrecht: Zuidam Drukkerijen.
Larkin, J.H., McDermott, J., Simon, D.P. & Simon, H.A. (1980). Models of competence in solving physics problems. *Science, 208,* 1335-1342.
Luan, S., Schooler, L.J. & Gigerenzer, G. (2011). A signal detection analysis of fast-and frugal trees. *Psychological Review, 118,* 2, 316-338.

McGrath, R.E. (2010). Prescriptive authority for psychologists. *Annual Review of Clinical Psychology, 6,* 21-47.
Meder, B., Mayrhofer, R. & Waldmann, M.R. (2014). Structure induction in causal reasoning. *Psychological Review, 121,* 3, 277-301.
Meehl, P.E. (1954). *Clinical versus Statistical Prediction*. Minneapolis: University of Minnesota Press.
Meehl, P.E. (1959). A comparison of clinicians with five statistical methods of identifying psychotic MMPI profiles. *Journal of Counseling Psychology, 6,* 102-109.
Meehl, P.E. (1965). Seer over sign: the first good example. *Journal of Experimental Research in Personality, 1*, 27-35.
Meehl, P.E. (1986). Causes and effects of my disturbing little book. *Journal of Personality Assessment, 50,* 370-375.
Meyer, G.J., Finn, S.E., Eyde, LD., Kay, G.G., Moreland, K.L., Dies, R.R., Eisman, E.J., Kubiszin, T.W & Reed, G.M. (2001). Psychological testing and psychological assessment. A review of evidence and issues. *American Psychologist, 56,* 2, 128-165.
Molenaar, P.C.M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research & Perspective, 2,* 4, 201-218.
Moore, G.E. (1903). *Principia Ethica.* Cambridge: Cambridge University Press.
Moore, D.A. & Healy, P.J. (2008). The trouble with overconfidence. *Psychological Review, 115,* 2, 502-517.,
Moxley, J.H., Ericsson, K.A., Charness, N. & Krampe, R.T. (2012). The role of intuition and deliberative thinking in experts'superior tactical decision-making. *Cognition, 124,* 1, 72-78.
Münsterberg, H. (1913). *Psychology and industrial efficiency*. Boston: Houghton Mifflin.

Nestler, S., Egloff, B., Küfner, A.C.P. & Back, M.D. (2012). An integrative lens model approach to bias and accuracy in human inferences: Hindsight effects and knowledge updating in personality judgments. *Journal of Personality and Social Psychology, DOI 10.1037/a0029461.*
Nisbett, R. & Ross, L (1980). *Human Inference: Strategies and shortcomings of social judgments.* New Jersey: Prentice Hall.

Osman, M. (2010). Controlling uncertainty: a review of human behavior in complex dynamic environments. *Psychological Bulletin, 136,* 1*,* 65-86.

Sagiv, L., Amit, A., Ein-Gar, D. & Arieli, S. (2014). Not all great minds think alike: Systematic and Intuitive Cognitive Styles. *Journal of personality, 82,* 5, DOI: 10.111/jopy.12071.
Salvatore, S. & Valsiner, J. (2010). Between the general and the unique: Overcoming the Nomothetic versus Idiographic Opposition. *Theory and Psychology, 20,* 6, 817-833.
Sarbin, Th.R. (1944). The logic of prediction in psychology. *Psychological Review,* 57, 2, 210-228.
Sarbin, Th.R. (1986). Prediction and Clinical Inference: Forty Years later*. Personality Assessment, 50,* 3, 362-369.
Sawyer, J. (1965). Measurement and prediction, clinical and statistical. *Psychological Bulletin, 66,* 2, 178-200.
Simonton, D.K. (2004). Psychology's status as a scientific discipline: Its empirical placement within a hierarchy of the sciences*. Review of General Psychology, 8,* 59-67.
Stern, W. (1911). *Die differentielle Psychologie*. (The Differential Psychology). Leipzig: Barth.
Stigler, S.M. (1999). *Statistics on the table: The history of statistical concepts and methods*. Cambridge, MA: Harvard University Press.

Thomae, H. (1968). *Das Individuum und seine Welt.* (The individual and his world). Göttingen (Germany): Hogrefe.

Weber, E.U. & Johnson, E.J. (2009). Mindful judgment and decision making. *Annual Review of Psychology, 60,* 53-85.
West, T.V. & Kenny, D.A. (2011). The truth and bias model of human judgment. *Psychological Review, 118,* 2, 357-378.
Westen, D. & Weinberger, J. (2004). When clinical description becomes statistical prediction. *American Psychologist, 59,* 7, 595-613.
Wierzbicki, M. (1993). Clinical versus statistical prediction. In: *Issues in Clinical Psychology*. Allyn & Bacon, ISBN 0205139728. (pp. 131-153).
Wiggins, J.S. (1981). Clinical and statistical prediction: where are we and where do we go from here? *Clinical Psychology Review, 1,* 3-18.
Windelband, W. (1904). *Geschichte und Naturwissenschaft*. (History and Science). Strassburg: Heitz & Mündel.
Wood, J.M., Garb, H.N., Lilienfeld, S.O. & Nezworski, M.T. (2002). Clinical Assessment. *Annual Review of Psychology, 53,* 519-543.

Hoofdstuk II

Barendregt, J. (1974). De relatie van diagnostiek en therapie binnen het fobieën project. *De Psycholoog, 9*, 295-308.
Barlow, D.H., Bullis, J.R., Comer, J.S. & Ametay, A. (2013). Evidence-Based Psychological Treatments: An update and a way forward. *Annual Review Clinical Psychology, 9,* 1-27.
Barth, J., Munder, T., Gerger, H., Nüesch, E., Trelle, S. et al. (2014). Comparative efficacy of seven therapeutic interventions for patients with depression. A network meta-analysis. *Plos Med. 10* (5) e 1001454 doi 10.1371 Journal pmed 1001454.

Boelema, S. (2014). *Alcohol use in adolescence. A longitudinal study of its effect on cognitive functioning.* Academisch proefschrift Universiteit Utrecht.
Braet, C. & Bögels, S.M. (2014). *Protocollaire behandelingen voor kinderen en adolescenten met psychische problemen.* Amsterdam: Boom.

Chambles, D.L. & Ollendick, T.H. (2001). Empirical supported psychological interventions: Controversies and evidence. *Annual Review of Psychology, 52*, 685-761.
Cronbach L.J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12,* 671-684.
Cronbach, L.J. & Snow, R.E. (1977). *Aptitudes and Instructional Methods.* New York: Wiley.
Cuijpers, P., Donker, T., Van Straten, A. Yuan, L. & Anderson. G. (2010). Is guided self-help as effective as face-to-face therapy? A systematic review and meta-analysis of comparative outcome studies. *Psychological Medicine, 21,* 1-15.
Cuijpers, P., Van Straaten, A. & Andersson, G. (2008). Psychotherapy for depression in adults: A meta-analysis of omparative outcome studies. *Journal of Consulting and Clinical Psychology, 76,* 6, 909-922.

Daniels, A.M. (pseudoniem: A. Dalrymple) 2012: *Leven aan de onderkant.* Houten: Spectrum.
De Bruyn, E.E.J., Ruijssenaars, A.J.J.M., Pameijer, N.K. & Van Aarle, E.J.M. (2015). *De diagnostische cyclus. Een praktijkleer.* Leuven/Den Haag Acco.
De Groot, A.D. (1967). *Vijven en zessen*. Groningen: Wolters.
Drenth, P.J.D. (1967). *Testtheorie: Inleiding in de theorie van psychologische tests en zijn toepassingen.* Houten: Bohn, Stafleu van Loghum.
Dumont, J.J. & Kok, J.F. (1973). *Curriculum Schoolrijpheid Deel I.* Den Bosch: Malmberg.

Fisher, R.A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society, 17*, 1, 69-78.

Gigerenzer, G. (2008). *Rationality for mortals. How people cope with uncertainty*. Oxford: Oxford University Press.
Gigerenzer, G. & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology, 62*, 451-482.
Hattie, J. *Visible learning: a synthesis of over 800 meta-analyses relating to achievement'. (*Van internet te downloaden).

Kouwer, B.J. (1963.) *Het spel van de persoonlijkheid*. Bijleveld: Utrecht.
Kuiper, P.C.: Barendregt, J. (1965). 'Ingezonden' en P.C. Kuiper, 'Ingezonden'. *Nederlands Tijdschrift voor Geneeskunde, 109*, 1736.

Lilienfeld, S.O. (2011). Public scepticism of psychology. Why many people perceive the study of human behavior as unscientific. *American Psychologist, DOI: 10.1037/a0023963.* Pp. 1-19.
Loevinger, J. (1997). Stages of Personality Development. In: R. Hogan, J. Johnson & S. Briggs (Eds.). *Handbook of Personality Psychology*. New York: Academic Press (pp. 199-208).

McHugh, R.K. & Barlow, D. (2010). The dissemination and implementation of Evidence-Based Psychological Treatments. *American Psychologist, 65,* 2, 73-84.
Merton, R. (1968). The Matthew effect in science. *Science, 159,* 56-63.

Poston, J.M. & Hanson, W.E. (2010). Meta-analysis of psychological assessment as therapeutic treatment. *Psychological Assessment, 22,* 203-212.

Radler, J. (2015). Bringing the environment in. Early Central European contributions to an ecologically oriented psychology of perception. *History of Psychology,* dx.doi.org/10.1037/a0039059.

Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for generalized causal inference.* New York: Houghton Mifflin Cy.
Shirk, S.R. & Karver, M. (2003) Prediction of treatment outcome from relationship variables in child and adolescent therapy: a meta-analytic review. *Journal of Consulting and Clinical Psychology, 71,* 3*,* 452-464.

Verschueren.K., & Koomen, H. (2007) (redactie). *Handboek diagnostiek in de leerlingbegeleiding. Antwerpen-Apeldoorn: Garant.*

Weisz, J.R., Weiss, B., Han, S.S., Granger, D.A. & Morton, T. (1995). Effects of psychotherapy with children and adolescents revisited; a meta-analysis of treatment outcome studies. *Psychological Bulletin, 117,* 3*,* 450-468.
Weisz, J.R., Doss, A.J. & Hawley, K.M. (2005). Youth Psychotherapy outcome research: A review a critique of the Evidence Base. *Annual Review of Psychology, 56*, 2, 337-363.
Weisz, J.R., McCarty, C.A. & Valeri, S.M. (2006). Effects of psychotherapy for depression in children and adolescents: a meta-analysis. *Psychological Bulletin, 132,* 1*,* 132-149.
Wissler (1901). 'The correlation of mental and physical tests'. *The Psychological Review, Monograph Supplement, 3,* p. 6.
Witteman, C., Van der Heijden, P & Claes, L. (2014). *Psychodiagnostiek: het onderzoeksproces in de praktijk.* Zoetermeer: De Tijdstroom.

Hoofdstuk III

Alleva, L. (2006). Taking time to savour the rewards of slow science. *Nature, 443,*271. Doi:10.1038/443271e.
Amis, K. (1953). *Lucky Jim.* London: Victor Gollancz Ltd.
Aussems, M. C.E., Boomsma, A. & Snijders, T.A.B. (2011). The use of Quasi-Experiments in the social sciences: a content analysis. *Quality and Quantity, 45,* 21-42.

Bakeman, R. & Gottman, J.M. (1997). *Observing interaction: an introduction to sequential analysis*. Cambridge: Press syndicate of the University of Cambridge.
Bauer, D.J. (2011). Evaluating individual differences in psychological processes. *Current Directions in Psychological Science, 20,* 2, 115-118.
Berkel, H.J.M. van (1984). *De diagnose van toetsvragen*. (Assessment of educational test items). Amsterdam: PhD Thesis University of Amsterdam.

Bickman, L (2000). (Ed.). Validity and social experimentation: *Donald Campbell's legacy.* London: Sage Publications. Inc.
Black, M. (1962). *Models and metaphors: Studies in language and philosophy*. Ithaca, NY: Cornell University Press.
Blinkhorn, S.F. (1997). Past imperfect, future conditional: Fifty years of test theory. *British Journal of Mathematical and Statistical Psychology, 50,* 175-185.
Boelema, S. (2014). *Alcohol use in adolescence. A longitudinal study of its effect on cognitive functioning.* Academisch proefschrift Universiteit Utrecht.
Boring, E.G. (1919). Mathematical versus scientific significance. *Psychological Bulletin, 16,* 335-338.
Bornstein, R.F. (2011). Toward a process-focused model of test score validity: Improving psychological assessment in science and practice. *Psychological Assessment, 23,* 2, 532-544.
Bornstein, M.H., Jager, J. & Putnick, D.L. (2013). Sampling in developmental science: situations, shortcomings, solutions, and standards. *Developmental Review, 33,* 357-370.
Braet, C. & Bögels, S.M. (2014). *Protocollaire behandelingen voor kinderen en adolescenten met psychische problemen.* Amsterdam: Boom.
Broers, N. & Roskam, Edw. E. (1991). Facet design and item response theory for appraisive judgments: Application to the study of lonesomeness. *Paper presented at the third international conference on facet theory, Jerusalem*, June, 1991.

Campbell, D.T. (1986). Re-labeling internal and external validity for applied social sciences. In: Trochim, W.M.K. (Ed.). Advances in quasi-experimental design analysis: *New directions for program evaluation*. Newbury Park, CA: Sage.
Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant validation by the multi-trait-multi-method matrix. *Psychological Bulletin, 56*, 1, 81-105.
Campbell, D.T. & Stanley, J.C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: RandMcNally.
Cheng, Y., Yuan, K.H., Cheng, L. (2011). *Educational and Psychological Measurement* (XX(X) 1-16. *DOI: 10.1177/00/3164411407315.*
Cohen, J. (1960). A coefficient for agreement for nominal scales. *Educational and Psychological Measurement, 20, 1,* 37-46.
Cohen, J. (1968). Weighted Kappa: nominal scale agreement with provision for scaled disagreement of partial credit. *Psychological Bulletin, 70,* 4, 213-220.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (Revised edition*). New York: Academic Press.
Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155-159.
Cohen, J. (1994). The earth is round (p <.05). *American Psychologist, 49*, 7, 997-1003.
Comenius, J. A. (1997, published originally, 1639). *Der Weg des Lichts, Via Lucis.* (Introduced and translated by U. Voigt). Hamburg, BRD. (The road of the light, Via lucis).
Cook, T.D. & Campbell, D.T. (1976). The Design and Conduct of Quasi-Experiments and true Experiments in Field settings. In: M.D. Dunette (Ed.). *Handbook of Industrial and Organizational Psychology*. Chicago: Rand McNally College Publishing Cy.
Cook, T.D. & Campbell, D.T. (1981). *Quasi- experimentation: Designs and Analysis Issues for field settings.* Newbury Park, CA: Sage.

Cook, T.D. & Shadish, W.R. (1994). Social Experiments: Some developments over the past fifteen years. *Annual Review of Psychology, 45,* 545-580.
Cook, T.D., Shadish, W.R. & Wong, V.C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-studies. *Journal of Policy Analysis and Management, 27,* 2, 724-750.
Courtis, S.A. (1921). Report of the standardization committee. *Journal of Educational Research, 4,* 78-80.
Crocker, L & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart& Winston.
Cronbach, LJ. (1988). Five perspectives on the validity argument. In: H. Wainer & H.L Braun (1988). *Test Validity.* Hillsdale NJ: Erlbaum (pp. 8-19).
Cronbach, LJ. & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.
Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles.* New York: J. Wiley & Sons.

Dawson, Th. (2004). Assessing Intellectual Development: Three approaches, one sequence. *Journal of Adult Development, 11,* 2, 71-85.
Diagnostic and Statistical Manual of Mental Disorders. DSM-5™ (5th Ed.). American Psychiatric Association (2013). Washington DC: APA.

Fava, G.A., Ruini, C. & Rafanelli, C. (2004). Psychometric theory is an obstacle to the progress of clinical research. *Psychotherapy and Psychosomatics, 73,* 2, 145-148.
Feinstein, A.R. (1987*). Clinimetrics.* New Haven: Yale University Press.
Fernbach, Ph.M., Macris, D.M. & Sobel, D.M. (2012). Which one made it go? The emergence of diagnostic reasoning in preschoolers. *Cognitive Development, 27,* 39-53.
Fisher, R.A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society, 17*, 1, 69-78.
Fisher, R.A. (1970, 16th edition). *Statistical methods for research workers*. New York: Hafner.
Fritz, A., Scherndl, Th. & Kühberger, A. (2012). A comprehensive review of reporting practices in psychology journals: Are effect sizes really enough? *Theory & Psychology, DOI: 10.1177/0959354312436870.* 0(0) 1-25.
Funder, D.C. (2001, 2nd Edition). *The personality puzzle*. New York: W.W. Norton &Company.

Garb, H.N. (2003). Incremental validity and the assessment of psychopathology in adults. *Psychological Assessment, 15,* 4, 508-520.
Gigerenzer, G. (2000). *Adaptive thinking: Rationality in a real world.* Oxford: Oxford University Press.
Gigerenzer, G. (2008). *Rationality of Mortals: How people cope with uncertainty*. Oxford: Oxford University Press.
Gillis, R.L. & Nilsen, E.S. (2012). Children's use of information quality to establish speaker preferences. *Developmental Psychology, DOI: 10.1037/a002947.9*
Herrnstein, H. & Wood, R. (1989). Five decades of item response modeling. *British Journal of Mathematical and Statistical Psychology, 42*, 1, 139-167.
Goodwin, L.D. (2001). Inter-rater Agreement and Reliability. *Measurement in Physical Education and Exercise Science, 5,* 1, 13-34.

Greve, W. (2001). Traps and gaps in action explanation: Theoretical problems of a psychology of human action. *Psychological Review, 108, 2*, 3, 435-451.
Grice, H.P. (1975). Logic and conversation. In: D. Davidson & G. Harman (Eds.) (1975). *The Logic of Grammar* (pp. 64-75). Encino CA: Dickenson.
Groenier, M., Vos, R.J., Pieters, J.M., Witteman, C. & Swinkels, J.A. (2011). Psychologist's diagnostic processes during a diagnostic interview. *Free Access Journal*: *Psych20110900010 - 54937165-1.*
Grossman, V. *Leven en Lot* (vertaald door Froukje Slofstra, 2014). Amsterdam: Uitgeverij Balans.
Guion, R.M. (1980). On Trinitarian conceptions of validity. *Professional Psychology, 11*, 3, 395-398.
Gulliksen, H. (1950). *Theory of Mental Tests*. New York: Wiley.
Guttman, L. (1965). A faceted definition of Intelligence. *Scripta Hierosolumitana, 14,* 66-181.

Harris, P.L. (2007). Trust. *Developmental Science, 10,* 1, 135-138.
Haynes, S.N. & Lench, H.C. (2003). Incremental validity of new clinical assessment measures. *Psychological Assessment, 15,* 4, 456-466.
Hartshorne, H. & May, M.A. (1928). *Studies in Deceit*. New York: McMillan.
Hendriks, J. (1997). *The five factor personality inventory* (FFPI). Academisch proefschrift. Rijks Universiteit Groningen (PhD thesis: University of Groningen, The Netherlands).
Hoffman, P.J. (1968). Cue-consistency and configurality in human judgment. In: B. Kleinmuntz (Ed.). *Formal representation of human judgment.* New York: J. Wiley.
Hofstee, W.K.B. (1994). Who should own the definition of personality? *European Journal of Personality, 8,* 149-162.
Hood, S.B. (2009). Validity in psychological testing and scientific realism. *Theory and Psychology, 19,* 4, 451-473.
Houdt, A. van (1994). *Waarheid rondom de leugen van het kind*. (Truth and lie in children) MA Thesis University Utrecht, The Netherlands, Department of Developmental Psychology.
Hox, J. (2002). *Multilevel analysis. Techniques and applications*. L. Erlbaum, Mahwah, New Jersey.
Hubbard, R. & Lindsay, R.M. (2008). Why p values are not a useful measure of evidence in statistical significance testing. *Theory and Psychology, 18,* 69-88.
Hunsley, J. (2003). Introduction to the special section of incremental validity and utility in clinical assessment. *Psychological Assessment, 15*, 4, 443-445.
Hunsley, J. & Meyer, G.J. (2003). The incremental validity of psychological assessment and testing: conceptual, methodological, and statistical issues. *Psychological Assessment, 15*, 4, 446-455.

Ioannidis, P. (2005). Why most published research findings are false. Open access journal: *PLOS (Public Library of Science) Medicine. 2: e124. Doi: 10.1371. journal.pmed.0020124.*

Jacobson, N.S. & Truax, P. (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 1, 12-19.
Jaffee, S.R., Strait, L.B., Odgers, C.L. (2012). From correlates to causes: Can quasi-experimental studies and statistical innovations bring us closer to identifying the causes of antisocial behavior? *Psychological Bulletin, 138,* 2, 272-295.
James, L.R., Demaree, R.G., Mulaik, S.A. & Ladd, R.T. (1992). Validity generalization in the context of situational models. *Journal of Applied Psychology, 77,* 1, 16-36.

Johnston, C. & Murray, C. (2003). Incremental validity in the assessment of children and adolescents. *Psychological Assessment, 15,* 4*,* 496-507.
Jonson, J.L & Plake, B.S. (1998). A historical comparison of validity standards and validity practices. *Educational and Psychological Measurement, 58*, 5, 736-753.

Kant, I. (1878/1961). *Kritik der reinen Vernunft*. (Critics on Pure Reason) Edited by Ingeborg Heidemann; P. Reklam Jun. Stuttgart. (Nederlandse vertaling beschikbaar).
Katzko, M. (2002). Unity versus multiplicity: A conceptual analysis of the term 'self' and its use in personality theories. *Journal of Personality, 71*, 1, 83-114.
Kelley, T.L. (1927*). Interpretation of educational measurements*. New York: MacMillan.
Kitchener, K.S. & King, P.M. (1981). Reflective judgment. Concepts of justification and their relationship to age and education. *Journal of Applied Developmental Psychology, 2,* 1, 89-116.
Kitchener, S., Lynch, C.L. Fischer, W. &Wood, P.K. (1993). Developmental range of reflective judgment: the effect of contextual support and practice on developmental stage. *Developmental Psychology, 29,* 893-906.
Kitchener, K.S., King, P.M., Wood, P.K. & Davison, M.L (1989). Sequentiality and consistency in the development of reflective judgment. *Journal of Applied Developmental Psychology, 10,* 73-95.
Kraemer, C.K. (2005). A simple effect size indicator for two-group comparisons? A comment on r-equivalent. *Psychological Methods, 10,* 4*,* 413-419.

Lakatos, I. (1968). Criticism and the methodology of scientific research programs. *Proceedings of the Aristotelian Society, 69*, 149-186.
Lambdin, Ch. (2012). Significance testing as sorcery: Science is empirical- significance tests are not. *Theory and Psychology, 22,* 1, 67-90.
Linschoten, J. (1959). *Op weg naar een fenomenologische psychologie: De psychologie van William James.*Utrecht: Bijleveld.
Lissitz, R.W. (Ed.). 2009). *The concept of validity: Revisions, New Directions, and Applications*. Charlotte NC: Information Age publishing, Inc.
Lissitz, R.W. (2009). Introduction. In: R.W. Lissitz (Ed.). *The Concept of Validity: Revisions, New Directions, and Applications* (pp. 1-15). Charlotte NC: Information Age publishing, Inc.
Loevinger, J. (1957). Objective tests as instruments of psychological theories. *Psychological Reports Monographs Supplement, 3,*635-694.
Lovasz, N. & Slaney, K.L. (2013). What makes a hypothetical construct 'hypothetical? Tracing the origins and uses of the 'hypothetical construct' concept in psychological science. *New Ideas in Psychology, 31,* 22-31.
Lucke, J.F. (2005). The alpha and omega of congeneric test theory: an extension of reliability and internal consistency to heterogeneous tests. *Applied Psychological Measurement, 29,* 1*,* 65-81.
Lumsden, J. (1978). Tests are perfectly reliable. *British Journal Mathematical and Statistical Psychology, 31,* 1*,* 19-26.

MacCorquadale, K. & Meehl, P.E. (1948). On a distinction between hypothetical constructs and intervening variables*. Psychological Review, 55,* 2, 95-107.
Maraun, M.D. & Gabriel, S.M. (2013). Illigitimate concept equating in the partial fusion of construct validation theory and latent variable modeling. *New Ideas in Psychology, 31,* 32-34.

Marra, R. & Palmer, B. (2004). Encouraging Intellectual growth: Senior college student profiles. *Journal of Adult Development, 11,* 2, 111-122.
Messick, S. (1988). The once and future issues of Validity. In: H. Wainer & H. Braun, (1988). *Test Validity*. Hillsdale NJ: Erlbaum (pp. 33-48).
Messick, S. (1989). Validity. In: R.L Linn (Ed.), (3nd ed.). *Educational Measurement.* National Council on Measurement in Education. London: Collier Macmillan Publishers.
Messick, S. (1994). Foundations of Validity: Meaning and Consequences in Psychological Assessment. European *Journal of Psychological Assessment, 1*, 1, 1-9.
Meyer, G.J., Finn, S.E., Eyde, L.D., Kay, G.G., Moreland, K.L, Dies, R.R., Eisman, E.J., Kubiszin.T.W. & Reed, G.M. (2001). Psychological testing and psychological assessment. A review of evidence and issues. *American Psychologist, 56,* 2,128-165.
Meyer, G.J., Finn, S.E., Eyde, L.D., Kay, G.G., Moreland, K.L, Dies, R.R., Eisman, E.J., Kubiszin.T.W. & Reed, G.M. (2002). Amplifying issues related to psychological testing and psychological assessment. A review of evidence and issues. *American Psychologist,* February*,* 2002, pp. 140-143.
Michigan Department of Education (1989). The Michigan Employability Survey. Ann Arbor: Author.
Miller, J. & Schwartz, W. (2011). Aggregate and individual replication probability within an explicit model of the research process. *Psychological Methods, 16,* 337-360.
Mitchell, G. (2012). Revisiting truth or triviality. The external validity of research in the psychological laboratory. *Perspectives on Psychological Science, 7,* 109-117.
Montaigne, Michel de. (16^{th}-17^{th} century). *Essays.* Dutch translation 2000: Amsterdam, The Netherlands: Boom.
Mumma, G.H. & Smith, J.L. (2001). Cognitive-behavioral-interpersonal scenarios: inter-formulator reliability. and convergent validity. *Journal of Psychopathology and Behavioral Assessment, 23,* 4, 203-221.
Mummendey, A. (2012). Scientific misconduct in Social Psychology: Towards a currency reform in Science. *European Bulletin of Social Psychology, 24,* 4-7.

Nevo, B. (1993). In search of a correctness typology for intelligence. *New Ideas in Psychology,* 391-397.
Nunnally, J.C. & Bernstein, I.H. (1994). *Psychometric Theory*. New York: McGraw Hill Inc.

Ones, D.S., Viswesvaran, C. & Schmidt, F.L. (1993). Comprehensive meta-analysis of Integrity Test Validities: Findings and Implications for Personnel Selection and Theories of Job Performance. *Journal of Applied Psychology, 78*, 679-703.

Perry, J. (1 970). *Forms of intellectual and ethical development in the College Years. A Scheme*. New York: Academic Press.

Raykov, T., Dimitrov, D.M., Von Eye, A. & Marcoulides, G.A. (2012). Inter-rater agreement evaluation: A latent variable modeling approach. *Educational and Psychological Measurement, DOI: 1177/0013164412449016* (pp. 1-20).
Roskam, E.E. Ch. I. (1991) Construct validity as explanatory theory. *Paper prepared for the First European Conference on Psychological Assessment.* Barcelona, Sept. 23-24, 1991.

Rubin, D.B. (2010). Reflections stimulated by the comments of Shadish (2010) and West and Thoemmes (2010). *Psychological Methods, 15*, 1, 38-46.
Rushton, J.P., Brainerd, C.J. & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin, 57*, 2, 210-228.

Sackett, P.R. (1994). Integrity testing for Personnel Selection. *Current Directions in Psychological Science, 3,*1, 69-73.
Sartori, R & Pasini, M. (2007). Quality and quantity in test validity: How can we be sure that psychological tests measure what they have to? *Quality & Quantity, 41,* 3, 359-374.
Smaling, Adri (1987). *Methodologische objectiviteit en kwalitatief onderzoek.* Lisse: Swets & Zeitlinger.
Schmidt, F.L (1992). What do data really mean? Research findings, Meta-analysis and cumulative & knowledge in psychology. *American Psychologist, 47*, 202-214.
Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1,* 115-129.
Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13,* 90-100. Doi: 10.1037/a0015108.*psychology.*
Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training researchers. *Psychological Methods, 1,* 2, 115-129.
Schmidt, F.L. & Hunter, J.E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1,* 2, 199-221.
Schmidt, F.L. & Hunter, J.E. (1998).The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 2, 262-274.
Schmidt, R.L, Law, K., Hunter, J.E., Rothstein. H.R., Pearlman, K. & McDaniel, M. (1993). Refinements in validity generalizations methods: Implications for the situational specificity hypothesis. *Journal of Applied Psychology, 78,* 1, 3-12.
Shadish, W.R. (2010). Campbell and Rubin: A primer and comparison of their approaches to causal inference in field settings. *Psychological Methods, 15,* 1*,* 3-17.
Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for generalized causal inference.* New York: Houghton Mifflin Cy.
Sheppard, L.A. (1993). Evaluating test validity. In: L. Darling-Hammond (Ed.). *Review of Research in Education (Volume 19)*. Washington DC: American Educational Research Association.
Simmons, J.P., Nelson, L.D. & Simonsohn, U. (2011). False positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Perspectives on Psychological Science, XX (X) 1-8. DOI: 10.1177/095679761147632.*
Sireci, S.G. (2007). On validity theory and test validation. *Educational Researcher, 36,* 477-481.
Slaney, K.L. (2012). Laying the cornerstone of construct validity theory: Herbert Feigl's influence on early specifications. *Theory and Psychology 0(0)* 1-20. *DOI: 10.1177/0959354311400659.*
Slaney, K.L. & Racine, T.P. (2013). What's in a name? Psychology's ever evasive construct. *New Ideas in Psychology, 31,* 4-12.
Smedslund, J. (1999). Psychology and the study of memory. *Scandinavian Journal of Psychology, 40,* (suppl. 3), pp. 3-17.
Standards for Educational and Psychological Testing (1999). Washington DC: American Educational Research Association American Psychological Association National Council on Measurement in Education.

Steinberg, L.,Thissen, D. & Wainer, H. (1990).Validity. In: Wainer, H. *Computer Adaptive testing.* A primer. Hillsdale NJ: L. Erlbaum (pp. 187-231).
Stouthard, M.A.E. & Peetsma, Th.T.D. (1999) Future-time perspective: Analysis of a facet-designed Questionnaire. *European Journal of Psychological Assessment, 15,* 99-105.
Strauss, M.E. & Smith, G.T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology, 5,* 1-25.

Tacq, J. (2011). Causality in quantitative and qualitative research. *Quality and Quantity, 45,* 263-291.
Tenney, E.R., Small, J.E., Konrad, P.L., Jaswal, V.K. & Spellman, B.A. (2011). Accuracy, confidence, and calibration: How young children and adults assess credibility. *Developmental Psychology, 47,* 4, 1065-1077.
Thorndike, E.L. (1918). *The seventh yearbook of the national society for the study of education. Pt. II.* Bloomington, IL: Public School Publishing Co.
Thorndike, R.L. (1949). *Personnel Selection*. New York: Wiley.
Trotsky, L. (27 November, 1932). Cited after T. Rüting: *Pavlov und der neue Mensch. Diskurse über Disziplinierung in Sowjetrussland.* München, 2002 (p. 179). (Pavlov and the new man: Arguing about discipline in the USSR).

Van Hezewijk, R. & Stam, H.J. (2014). Door het besluit tot eenzijdigheid... Biografische schets van Johannes Linschoten (1925-1964). In: *Van Fenomenologie naar Empirisch-Analytische Psychologie.* Busato, V., Van Essen, M. & Koops, W. Redactie. Amsterdam: Prometheus, Bert Bakker.
Verschueren.K., & Koomen, H. (2007) (redactie). *Handboek diagnostiek in de leerlingbegeleiding. Antwerpen-Apeldoorn: Garant.*

Waller, N. & Jones. J. (2011, 31 March). Investigating the performance of alternate regression weights by studying all possible criteria in regression models with a fixed set of predictors. *Psychometrika, DOI:1007/s11336-011-9209-5.*
West, E.J. (2004). Perry's legacy: Models of epistemological development. *Journal of Adult Development 11,* 2, 61-70.
Westmayer, H. (2004). Against confounding predictors and criteria in psychological assessment. *Key note lecture on the VIIth conference of the European Society of Psychological Assessment, Malaga: Spain*.
Wheelright, P.E. (1968). *The burning fountain. Bloomington*, IN: Indiana University Press.
Williams, John (1957*). Stoner*. New York: The New York Review of Books.
Williams, Bernard (1985). *Ethics and the limits of Philosophy*. London: Fontana.
Williams, Bernard (2002). *Truth and Truthfulness, an essay in genealogy*. Princeton and Oxford; Princeton University Press.
Wittman, W.W. (1988). Multivariate reliability theory. Principles of symmetry and successful Validation strategies. In: R.J. Nesselroade & R.B. Cattell. *Handbook of Multivariate Experimental Psychology (*2nd Edition). New York: Plenum Press.

Hoofdstuk IV

APA Standards for Educational and Psychological Tests. (1999). Washington DC. American Psychological Association. (er is een voorlopige nieuwe versie uitgekomen in 2014).

Baird, B.M., Le, K. & Lucas, R.E. (2006). On the nature of intra-individual personality variability, reliability, validity, and associations with well-being. *Journal of Personality and Social Psychology, 90,* 3, 512-527.

Barelds, D.H.P. & Kooij A. (2013). Het meten van intelligentie bij volwassenen met de WAIS-IV-NL. *De Psycholoog,* Oktober, 2013, 13-24.

Bechger, T.M., Maris, G., Verstraten, H.H.F.M. &. Beguin, A.A. (2003). Using classical test theory in combination with Item Response Theory. *Applied Psychological Measurement, 27,* 5, 319-334.

Bentler, P.A. & Woodward, J.A. (1980). Inequalities among lower bounds to reliability: With applications to test construction and factor analysis. *Psychometrika, 45,* 249-267.

Blinkhorn, S.F. (1997). Past imperfect, future conditional: Fifty years of test theory. *British Journal of Mathematical and Statistical Psychology, 50,* 2, 175-185.

Bollen, K.A. (2002) Latent variables in psychology and the social sciences. *Annual Review of Psychology, 53*, 4, 605-634.

Bond, T.G. & Fox. C. M. (2001). *Applying the Rasch model: Fundamental Measurement in the Human Sciences*. London: L. Erlbaum Publishers.

Borsboom, D., Mellenbergh, G.J. & Heerden, J. Van (2003). The theoretical status of latent variables. *Psychological Review, 110,* 2*,* 203-219.

Borsboom, D., Mellenbergh, G.J. & Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 4, 1061-1071.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart & Winston.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 3, 297-334.

Cronbach, L.J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12,* 671-684.

Cronbach, L.J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30,* 116-127.

Cronbach, L.J. (1990). *Essentials of Psychological Testing.* (6th edition). New York: Harper & Row.

Cronbach L.J., Gleser, G.C., Nanda, H & Rajaratnam, N. (1970). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles.* New York: J. Wiley & Sons.

Cronbach, L.J. & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 3, 281-302.

Cronbach, L.J. & Snow, R.E. (1977). *Aptitudes and Instructional Methods.* New York: Wiley.

Csikszentmihalyi, M. & Csikszentmihalyi, I.S. (Eds.) (1988). *Optimal experience: Psychological studies of flow in consciousness*. New York: Cambridge University Press.

Dam, K. van (1991). *Fixatie op fouten.* Lisse: Swets & Zeitlinger.

De Ayala, R.J. (2009). *The theory and practice of Item Response Theory.* NY: The Guilford Press.

Devlin, K. (2008). *The unfished game: Pascal, Fermat and the 17th century letter that made the world modern.* New York: Basic Books.

Drenth, P.J.D. & Sijtsma, K. (2006). *Testtheorie: Inleiding in de theorie van psychologische tests en zijn toepassingen* (4de Edition). Houten, The Netherlands: Bohn, Stafleu van Loghum. (Test theory: Introduction in the theory of psychological tests and their applications).

Diagnostic and Statistical Manual of Mental Disorders. DSM-5™ (5th Ed.). American Psychiatric Association (2013). Washington DC: APA.

Edwards. A.L. (1957). *Techniques of attitude scale construction*. New York: Appleton-Century-Crofts.
Embretson, S.E. (1999). Cognitive psychology applied to testing. In: FT. Durso (Ed.). *Handbook of Applied Cognition*. New York: John Wiley & Sons.
Embretson, S. & Prenovost, K. (1998). Item response research in assessment theory. In: P.C. Kendall, J.N.Butcher& G.N. Holmbeck (Eds.). *Handbook of Research Methods in Clinical Psychology* (pp. 276- 294).
Embretson, S.E. & Reise, S.P. (2000*). Item response theory for psychologists.* London: Lawrence Erlbaum Publishers.

Fisher, R.A. (1970, 14th edition; 1st edition, 1918). *Statistical Methods for Research Workers*. New York: Hafner.
Fischer, G.H. (1974). *Einführung in die Theorie psychologischer Tests*. Bern: Huber.

Gilovitch, T.D, Griffin, D.W. & Kahneman, D. (Eds.) (2001). *Heuristics and Biases: the Psychology of intuitive judgment.* Cambridge: Cambridge University Press.
Goldstein, H, & Wood, R. (1989). Five decades of item response modeling. *British Journal of Mathematical and Statistical Psychology, 42*,139-167.
Gregson, R.A.M. (1975). *Psychometrics of Similarity.* New York: Academic Press (Harcourt Brace Jovanovich Publishers).
Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.
Guttman. L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10,* 2, 255-282.
Guttman, L. (1971). Measurement as a structural theory. *Psychometrika, 36*, 3, 329-347.
Guttman, L. (1978, August). Recent structural laws of human behavior. *Paper read at the 9th International Sociological Congress, Uppsala Sweden.*

Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory: Principles and Applications*. Boston, Ma: Kluwer Academic Publishers.
Hambleton, R.K. (1986). (Review Editor). Standards for educational and psychological testing: Six Reviews. *Journal of Educational Measurement, 23,* 1, 83-98.
Hambleton, R.K. (1989). Principles and selected applications of Item Response Theory. In: R.L Linn (Ed.). *Educational Measurement* (3rd Edition). New York: American council on Education. Macmillan 'Publishing company, London: Collier Macmillan Publishers (pp. 147-201).
Heider, F. (1958). *The Psychology of Interpersonal Relations*. New York: Wiley.
Hogan, Th. P. (2003). *Psychological testing: A practical Introduction.* New Yotk: John Wiley & Sons, Inc.
Hoijtink, H., Klugkist, I. & Boelen, P.A. (2008). *Bayesian evaluation of informative hypotheses in psychology*. New York: Springer.

Johnson, J.H. & Sheeber, L.B. (1999). Developmental Assessment. In: WK. Silverman &T.H. Ollendieck (Eds.). *Developmental issues in the clinical assessment of children* (pp. 44-60). Needham Heights, Ma: Allyn & Bacon.

Kuhn, T.S. (1962). The structure of scientific revolutions. *International Encyclopedia of unified Science (Vol. 2, 2).* Chicago: The University of Chicago Press (2nd enlarged edition: 1970).

Latour, B. (1977). Pourquoi Péguy se répète-t-il? Pourquoi est-il-illisible? (Why does Péguy repeat himself? Why is he unreadable? In: *Péguy ecrivain: Colloque du centenaire*, Orleans, Septembre, 1973. Paris: Klinsieck (pp. 76-102).
Latour, B. (1994). *Wij zijn nooit modern geweest.* Amsterdam: Van Gennep. (vertaling van Nous n'avons jamais été modernes. Parijs: La Dévouverte, 1991).
Levy, Ph. (1973). On the relation between test theory and psychology. In: P. Kline (Ed.). *New approaches in psychological measurement.* London: J. Wiley.
Lord, F.M. & Novick, M.R. (1968). *Statistical Theories of Mental Tests Scores*. Reading: Ma.: Addison Wesley Publishing Company.

Mellenbergh, G.J. (1994). Generalized Item Response Theory. *Psychological Bulletin, 115*, 300-307.

Nunnally, J.C. & Bernstein, I.H. (1994). *Psychometric Theory*. New York: McGraw Hill Inc.

Peterson, R.A. & Kim, Y. (2012). On the relationship between Coefficient Alpha and Composite Reliability. *Journal of Applied Psychology, DOI: 10.1037/a0030767.*

Rasch, G. (1960). *Probabilistic Models for some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research.
Reise, S.P. & Waller, N.G. (2009). Item Response Theory and Clinical Measurement. *Annual Review of Clinical Psychology, 5,* 27-48.

Sijtsma, K. (2009). On the use, the misuse and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74,* 107-120.
Sijtsma, K. (2012). Future of Psychometrics: ask what psychometrists can do for psychology. *Psychometrika, 77,* 1, 4-20.
Spearman, C. (1904). 'General intelligence objectively determined and measured. *American Journal Of Psychology, 15*, 210-293.
Spearman, C. (1927). *Abilities of Men: Their Nature and Measurement*. New York: Macmillan.
Stouthard, M.A.E. & Peetsma, Th.T.D. (1999). Future-time perspective: Analysis of a facet-designed questionnaire. *European Journal of Psychological Assessment, 15,* 1, 99-105.

Ten Berge, J.M.F. & Soçan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika, 69,* 4, 613-625.
Torgerson, W.H. (1958). *Theory and Methods of Scaling*. New York: J. Wiley.
Thurstone, L.L. (1935). Vectors of Mind. Chicago: University of Chicago Press.
Thurstone, L.L. (1938). Primary Mental Abilities. Chicago: University of Chicago Press.
Trafimov, D. (2013). Are measurement theories falsifiable, and should we care? *Theory and Psychology 0 (0),* 1-5. DOI: 10.1177/095935431 348979.

Walsh, W.B. & Betz, N.E. (1990, 1st Edition; 2001, 3rd Edition). *Tests and Assessment*. New Jersey: Prentice Hall.
Wang, L. (P).), Hamaker, E. & Bergeman, C.S. (2012). Investigating Inter-Individual Differences in short-term Intra-Individual Variability. *Psychological Methods, DOI: 10. 1037/a0029317,* 1-15.
Westmayer, H. (2003). On the structure of case formulation. *European Journal of Psychological Assessment, 19,* 3, 210-217.
Wilson, M. (2013). Seeking balance between the statistical and scientific elements in psychometrics. *Psychometrika,* DOI: 10.1007/s11336-013-9327-3.

Hoofdstuk V

Andrews, G, & Hobbs, M.J. (2010). The effect of the draft DSM-5 criteria for GAD on prevalence and severity. *Australian and New Zealand Journal of Psychiatry, 44,* 784-790.
APA Standards for Educational and Psychological Tests. Washington DC: American Council for Educational Measurement; Washington DC: APA. American Psychological Association (APA, 1999). (Nieuwe voorlopige versie is van 2014).

Batstra, L. & Thoutenhoofd, E. (2013). Overleeft de psychiatrie de DSM-5? (Will psychiatry survive the DSM-5?). *De Psycholoog, 48, 3,* 10-17. (The Psychologist: Journal of the Dutch Psychological Association).
Berkhout, K. & Rosenberg, E. (NRC Magazine: 2012: 1-14-2012). Op zoek naar zonden (Searching for sins). Scientific Fraud.
Birkeland, S.A., Lismore, J.L., Manson, T.M., Brannick, M.T. & Smith, M.A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment 14,* 4, 317-335.
Brenner, E. (2003). Consumer-focused psychological assessment. *Professional Psychology: Research and Practice, 34,* 3, 240-247.

Cates, J.A. (1999). The art of assessment in psychology: ethics, expertise, and validity. *Journal of Clinical Psychology, 55,* 5, 631-641.
CAS. Utrecht Centre for Child and Adolescent Studies: *CAS PhD Program* (z.j.). Verzonden November, 2014 by J. Tenkink-de Jong, Secretary.

Dautzenberg, A.H.J. (2013). *Rafelranden van de moraal, Novelle.* Antwerpen Amsterdam: Atlas Contact.
De Groot, A.D. (1961, Methodologie: Grondslagen van onderzoek en denken in de gedragswetenschappen (12de druk 1994). Den Haag: Mouton.
De Groot, A.D. (1946). *Het denken van de schaker*. Den Haag: Mouton.
De Waal, F. (2015). *Bonobo en de tien geboden*. Amsterdam: Uitgeverij Atlas Contact.
Dijksterhuis, A., Van Knippenberg, A. & Veling, H. (2014). Newell and Shanks'approach to psychology is a dead end. *Behavioral and Brain Sciences, 37,* p. 25.

Fanelli, D. (2010). Do pressures to publish increase scientists' bias? An empirical support from US States data. *PLoS (Public Library of Science) One, 5,* 4, e 10271. DOI: 10.1371journal.pone.0010271.
Fanelli, D. (2010). 'Positive' results increase down the hierarchy of the sciences. *PLoS (Public Library of Science) One, 5,* 3, e10068. DOI: 10.1371/journal/pone.0010068.

Fanelli, D. (2011). Negative results are disappearing from most disciplines and countries. *Scientometrics.* DOI: 10.1007/s11192-011-0494-7. (published online: 11 September 2011).
Ferguson, C.J. (2015). 'Everybody knows Psychology is not a real science'. Public perceptions of psychology and how we can improve our relationship with policymakers, the scientific community, and the general public. *American Psychologist, 70,* 6, 527-542.

Gigerenzer, G. (2000). *Adaptive thinking: Rationality in the real world*. New York: Oxford University Press.
Gigerenzer, G. (2008). *Rationality for mortals: How people cope with uncertainty*. Oxford: Oxford University Press.
Griffin, B., Hesketh, B. & Grayson, D. (2004). Applicants faking good: evidence of item bias in the NEO PI-R. *Personality and Individual Differences, 36,* 1545-1558.

Holden, R.R., Wood, L.L.& Tomashewski, L. (2001). Do response time limitations counteract the effect of faking on personality inventory items? *Journal of Personality and Social Psychology, 81,* 1, 160-169.

James, W. (1902, 1982). *Varieties of religious experience*. Hammondsworth. In 1982 printed as Penguin Book.
James, W. (1909/ 1996). *A pluralistic universe.* Lincoln: University of Nebraska Press.
Jensen, A.R. (2011). The theory of intelligence and its measurement. *Intelligence, 39,* 171-177.
Johnson, E.L. (2012). Mapping the field of the whole human: Toward a form psychology. *New Ideas in Psychology,* http://dx.doi.org/10.1016/j.newideapsych.2012.09.002

Kessler, R.C., Chiu, W.T., Demler, O & Walters, E.E. (2005), Prevalence, severity, and comorbidity of twelve-month DSM-IV disorders in the national comorbidity survey replication (NCS-R). *Atchives of General Psychiatry, 62,* 617-627.
Kolfschoten, F. (2012). *Ontspoorde wetenschap: Over fraude, plagiaat en academische mores.* Amsterdam: Uitgeverij De Kring.
Kruglanski, A.W. & Gigerenzer, G. (2011). Theoretical Note: Intuitive and deliberate judgments are based on common principles. *Psychological Review, 118,* 1, 97-109.
Kruglanski, A.W. & Higgins, E.T. (2004). Theory construction in social personality psychology: Personal experiences and lessons learned. *Personality and Social Psychology Review, 8,* 2, 96-97.
Kupfer, D.J., First, M.B. & Regier, D.A. (2012). Introduction. In: D.J. Kupfer, M.B. First & D.A. Regier (Eds.). *A research agenda for DSM-5* (pp. xv-xxiii). Washington: American Psychological Association.

Laak, J. ter, Van Leuven, M. & Brugman, G. (2000). The effect of the accountability instruction and two job types on the Big Five scores. *European Journal of Psychological Assessment, 16,* 2, 209-214.
Lilienfeld, S.O. (2011). Public skepticism of psychology. Why many people perceive the study of Human behavior as unscientific. *American Psychologist, DOI: 10.1037/a0023963. Pp 1-19.*
Lilienfeld, S.O., Wood, J.M. & Garb, H.N. (2006). Why questionable psychological tests remain popular. *Scientific Review of Alternative Medicine, 10,* 6-15.

Mewton, L., Slade, T., McBride, O., Grove, R. & Teesson, M. (2011). An evaluation of the proposed

DSM-5 Alcohol use disorder criteria using Australian national data. *Addiction, 106,* 941-950.
Moreland, K.I., Eyde, L.D., Robinson, G.J., Primoff, E.S. & Most, R.B. (1995). Assessment of test user Qualifications. A research-based measurement procedure. *American Psychologist, 50,* 1*,* 14-23.

Newell, B.R. & Shanks, D.R. (2014). Unconscious influences on decision making: a critical review. *Behavioral and Brain Sciences, 37,* 1-19.

Ones, D.S., Viswesvaran, C. & Schmidt, F.L. (1993). Comprehensive meta-analysis of Integrity Test validities: Findings and Implications for Personnel Selection and Theories of Job Performance*. Journal of Applied Psychology, 78*, 5, 679-703.

Proudhon, P.J. (z.j.). *Wat is eigendom?* Vertaald door Z. Pennings. Utrecht: Uitgeverij IJzer.

Reichenbach, H. (1938). *Experience and Prediction: An analysis of the foundation and structure of knowledge.* Chicago Illinois.: The University of Chicago Press
Roskes, M., De Dreu, C.K.W. & Nijstad, B.A. (2012). Necessity is the mother of invention: Avoidance motivation stimulates creativity through cognitive effort. *Journal of Personality and Social Psychology, 103,* 2, 242-256.

Scherbaum, Ch.A., Sabet, J., Kern, M.J. & Agnello, P. (2013). Examining faking on personality Inventories Using unfolding Item Response Theory models. *Journal of Personality Assessment, 95,* 2, 207-216.
Schmand, B., Lindeboom, J., Schagen. S. Heijt, R, Koenen, T. et al. (1998). Cognitive complaints in Patients after whiplash injury. The impact of malingering. *Journal of Neurology, Neurosurgery and Psychiatry, 64,* 339-343.
Schmand, B. De Sterke, S. & Lindeboom, J. (1999). *Amsterdamse Korte Termijn Geheugen Test*. Amsterdam: Pearson.
Singh, J., Avasthi, A & Grover, S. (2007). Malignering of psychiatric disorders: A review. *German Journal of Psychiatry, 10,* 126-132.
Stark, S., Chernyshenko, A. S. & Yin Chan, K. (2001). Effect of the testing situation on item responding: Cause for concern. *Journal of Applied Psychology, 86,* 5*,* 943-953.

Tett, R.P., Freund, K.A., Christiansen, N.D., Fox, K.E. & Coaster, J. (2012). Faking on self-report Emotional intelligence and personality tests: Effects of faking opportunity, cognitive ability, and job type. *Personality and Individual Differences, 52,* 195-201*.*
Timpano, K.R., Exner, C., Glaesmer, H., Rief, W.Keshaviah, A., Brähler, E., & Wilhelm, S. (2011). The epidemiology of the proposed DSM-5 Hoarding disorder: Exploration of the acquisition Specifier, associated features, and distress. *Journal of Clinical Psychiatry, 72,* 780-786.

Van der Meulen, B.F. (2008). *Opwaartse druk.* Afscheidsrede Rijks Universiteit Groningen: Groningen: Stichting Kinderstudies.
Vecchione, M., Allesandri, G. & Barbaranelli, C. (2012). The five-factor-model in personnel selection: Measurement equivalence between applicant and non-applicant groups. *Personality and Individual Differences, 52,* 503-508.

Viswesvaran, C. & Ones, D.S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59,* 2, 197-210.
Voorbrood, C & Luijn, H. van (2010). Data: bread of psychologists? Archiving, making available and re-use of research data in psychology (Dutch: *DANS studies in digital archiving 4.* Amsterdam: Aksant Academic Publishers.

Walters, G.D., Wilson, N.J. & Glover, A.J.J. (2011). Predicting recidivism with the Psychopathy Checklist: Are factor score composites really necessary? *Psychological Assessment, 23,*552-557.
Wicherts, J.M., Borsboom, D., Kats, J. & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist, 61,* 726-728.
Wicherts, J.M., Bakker, M. & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting statistical results. *Public Library of Science*: *PLoS ONE*, 6, e 26828.

Auteur

Jan J.F ter Laak (1944) was werkzaam aan de universiteiten van Nijmegen, Tilburg, Leiden en Utrecht. Bij de afdeling Ontwikkelingspsychologie van de Universiteit Utrecht doceerde hij gedurende bijna 20 jaar psychologische diagnostiek. Hij was in de praktijk werkzaam in Enschede, Haarlem en Breda. Hij was lid van het hoofdbestuur van het Nederlands Instituut van Psychologen (NIP) voor psychologische diagnostiek, lid en voorzitter van de Commissie Testaangelegenheden van het NIP (Cotan) en voorzitter van de sectie Kinder & Jeugd van het NIP.