STATISTIEK EN SAMENLEVING EN HET METEN VAN MENSEN

Ivo W. Molenaar[1]

SAMENVATTING

Het meten van mentale eigenschappen van mensen gebeurt op grote schaal, b.v. in het kader van schoolkeuze-adviezen, eindexamens, personeelselectie en psychiatrische diagnose. Het wekt dikwijls meer weerstand en twijfel op dan het meten van lichamelijke eigenschappen. Na een korte schets van de geschiedenis van het testen, probeert dit artikel na te gaan waar dit verschil in appreciatie vandaan komt, of het terecht is, en in welke mate het probleem door het toepassen van kwantitatieve methoden of door beleidsmaatregelen zou kunnen worden verholpen.

---

[1] Ivo W. Molenaar, Vakgroep Statistiek & Meettheorie; FPPSW, Rijksuniversiteit Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, tel. 050-636185, na 10 oktober 1995 050-3636185. E-mail: W.Molenaar@ppsw.rug.nl . Met dank aan de collega's Hofstee, Schoonman en Sijtsma voor enkele waardevolle suggesties.

1. INLEIDING

Soms worden mensen gemeten. Van zuigelingen wordt het geboortegewicht vastgesteld en van recruten de lichaamslengte. Bij koorts gebruikt men een thermometer en bij de donorkeuring van de bloedtransfusiedienst wordt het hemoglobinegehalte bepaald. Gemeenschappelijk aan deze voorbeelden is het bestaan van een zorgvuldig gemaakt meetinstrument, dat (mits zorgvuldig toegepast) een betrouwbare en geldige conclusie toelaat door het gevonden getal te vergelijken met eerder ontwikkelde normtabellen. Bij een te lichte baby, een te kleine recruut, te hoge koorts of een te laag HB-gehalte, kan de metende instantie tot een speciale actie besluiten. Meestal wordt daarbij ook nog andere informatie over de gemeten persoon in de overwegingen betrokken, omdat de keuze van de beste actie niet uitsluitend van de meetuitkomst afhangt.

Soms worden mensen gemeten. Bij leerlingen van 6-Atheneum wordt hun kennis van Biologie of Engels vastgesteld middels een door het CITO (Centraal Instituut voor Toets Ontwikkeling) ontworpen centraal schriftelijk eindexamen. Sollicitanten bij de Nederlandse Spoorwegen, of ze nu treinbestuurder of directeur willen worden, moeten zich onderwerpen aan selectietests waarbij zowel vaardigheden als karaktereigenschappen in testscores worden uitgedrukt. In een psychiatrische kliniek wordt een week na de opname van een patient door een behandelend staflid, op basis van een halfgestructureerd interview met de patient, de Present State Examination ingevuld, een gestandaardiseerde vragenlijst waaruit via een vast protocol scores op symptomen en syndromen worden berekend. Gemeenschappelijk aan deze voorbeelden is het bestaan van een zorgvuldig gemaakt meetinstrument, dat (mits zorgvuldig toegepast) een betrouwbare en geldige conclusie toelaat door het gevonden getal te vergelijken met eerder ontwikkelde normtabellen. Bij een te geringe kennis van Biologie of Engels, een te lage score op intelligentie of stressbestendigheid, resp. een sterke indicatie van bijvoorbeeld hallucinaties en wanen, kan de metende instantie tot een speciale actie besluiten. Meestal wordt daarbij ook nog andere informatie over de gemeten persoon in de overwegingen betrokken, omdat de keuze van de beste actie niet uitsluitend van de meetuitkomst afhangt.

Lijken de voorafgaande twee alinea's op elkaar? Jazeker, en wel met opzet. Maar er zijn ook flinke verschillen, en daarover gaat dit artikel. Verschillen in de statistiek,

d.w.z. in de kwantitatieve methoden die helpen om na te gaan of het meetinstrument wel zorgvuldig ontworpen is, of de normtabellen wel goed in elkaar zitten, of de risico's van een foutieve actiekeuze niet te groot zijn. Maar ook verschillen in de samenleving, die meting van lichaamsgewicht, lichaamslengte, lichaamstemperatuur en hemo-globinegehalte vrijwel steeds als vanzelfsprekend, betrouwbaar en nuttig percipieert, maar in doorsnee een veel lagere appreciatie heeft voor tests. Of het nu eindexamens, personeelselectie of psychiatrische diagnose betreft, er is alom een sterke neiging om de kwaliteit van de meting, het nut van de meting, ja zelfs de mogelijkheid van meting en het recht om over een ander te beslissen (al dan niet op basis van een meting) luidkeels ter discussie te stellen.

De dokter mag wel aan mijn lijf komen maar de psycholoog niet aan mijn ziel? De dokter kan meten want die heeft ervoor geleerd? Zielezaken vallen niet te meten, alleen te voelen? Ik wil tot elke prijs het diploma, de baan, ontslag uit de kliniek, dus deugen je metingen niet? In het volgende zal ik proberen dit kluwen van onvrede en kritiek een beetje te ontrafelen.

Is het echt zo slecht gesteld met de meetkwaliteit van tests? Wat valt er aan te verbeteren op het punt van theorievorming, instrumentontwerp, toepassingsvoorschrift, voorlichting aan beslissers en aan te meten personen? Door welke bronnen van variabiliteit zijn de metingen minder nauwkeurig dan bij de weegschaal en de thermometer? Valt die variabiliteit te kwantificeren, of nog beter, te reduceren, met het arsenaal van kwantitatieve methoden, of met heel andere maatregelen? Waarom wordt het meten van mensen als het over hun mentale eigenschappen gaat zo gauw als bedreigend, duur, onbetrouwbaar, ongeldig of zelfs onmogelijk beschouwd?

Voor iedereen afdoende antwoorden op deze serie vragen heb ik niet in huis. Sterker nog, ik weet nu al dat een ander die ook niet heeft, dus het helpt niet als de redactie dit stuk door een ander laat schrijven. Maar de VvS bestaat vijftig jaar, en er zijn nog minstens vijftig leden die geen enkel verband weten te noemen tussen de VvS en het meten van mentale vermogens van mensen. Die mogen vandaag op ontdekkingsreis naar Psychometrika.

2. GESCHIEDENIS EN AARDRIJKSKUNDE

Aan Drenth & Sijtsma (1990, hoofdstuk 1) ontleen ik de volgende samenvatting van de geschiedenis van het testen, door mij naar eigen inzicht bekort en toegelicht. Vierduizend jaar geleden liet een Chinese keizer zijn dienaren eens per drie jaar een vorderingentoets ondergaan die werd gebruikt voor beslissingen omtrent ontslag of promotie. Vierhonderd jaar geleden kende de orde der Jezuieten schriftelijke examens voor toelating en evaluatie. Tweehonderd jaar geleden betoogde de Franse arts Pinel dat krankzinnigen niet misdadig maar ziek waren. Daarmee begon een serie onderzoekingen die leidde tot het onderscheid tussen krankzinnig en zwakzinnig, en de ontdekking dat men zwakzinnige kinderen met veel training en veel geduld eenvoudige vaardigheden kon leren. Ruim honderd jaar geleden bestudeerden experimenteel psychologen, vooral in Duitsland, de zintuigen en de motoriek van gezonde mensen; op zoek naar algemene wetmatigheden merkten zij tot hun ongenoegen dat er systematische individuele verschillen bestonden. De eerste Nederlandse hoogleraar in de psychologie, Heymans, begon in dezelfde psychofysische traditie maar hield zich later meer met persoonlijkheidstypen bezig. In Engeland bestudeerde de bioloog Galton de erfelijkheid van zowel lichamelijke als psychische eigenschappen, en in Chicago konden de bezoekers van een tentoonstelling in 1893 zich aan enkele tests onderwerpen waarbij hun prestatie aan algemene normen werd gerelateerd.

Gedachten omtrent systematisch meten en normeren van prestaties zijn dus al heel oud. Maar een adhoc samengesteld examen is nog geen psychometrisch verantwoorde test (Dousma & Horsten, 1994). Daarom wordt vaak gezegd dat Binet in Parijs in 1904 de eerste echte test construeerde: hij ontwierp een serie opgaven die hij zag als een steekproef uit de taken die een schoolkind zoal tegenkwam, legde die aan een representatieve groep kinderen voor, verving opgaven die niet goed functioneerden, bepaalde de moeilijkheidsvolgorde van de overige opgaven, en gebruikte de gegevens om met het begrip "mentale leeftijd" de vertraagde geestelijke groei van sommige kinderen te meten. Dit werk vond navolging in Duitsland, Engeland, en vooral de Verenigde Staten, waar de Stanford-Binet test en het intelligentiequotient in veel onderzoek werden gebruikt.

In de Eerste Wereldoorlog werden voor het eerst allerlei tests schriftelijk en

groepsgewijs afgenomen voor de selectie van grote groepen mannen voor functies in het leger. In Amerika, met een combinatie van snelle industriele groei, grote groepen immigranten en een optimistisch positivistisch idee over de bijdrage van wetenschap en techniek aan de samenleving, bleef ook na de vrede van Versailles het massaal gebruik van tests voor selectie en plaatsing bestaan, zowel bij bedrijven als in het onderwijs. In 1928 publiceerde Thurstone onder de triomfantelijke titel "Attitudes can be measured" een methode waarmee ook houdingen en opvattingen van mensen konden worden gemeten, en vanaf dat moment wordt ook in sociologisch en politicologisch onderzoek met gestandaardiseerde schalen gemeten; het voornaamste verschil is dat de scoring eens/oneens de scoring goed/fout uit de vaardigheids- of capaciteitentest vervangt.

Ook in Europa werd steeds vaker de hulp van een psycholoog ingeroepen bij beslissingen omtrent leerproblemen, beroepskeuze, personeelselectie en deviant gedrag. Maar hier domineert tussen ongeveer 1920 en 1960 een fenomenologische of personalistische opvatting van de psychologie, waarbij een kwalitatief oordeel op grond van individuele observatie van de proefpersoon van veel meer waarde werd geacht dan een onder gestandaardiseerde omstandigheden verkregen numerieke score. Een psychologisch rapport heeft daarbij meer het karakter van een essay over een ontmoeting met een medemens, terwijl in de V.S. eerder een interpretatie van numerieke testgegevens zou worden gepresenteerd. Als in de personalistische visie tests worden gebruikt, zijn dat minder vaak opgaven met goed/fout of eens/oneens scoring, en vaker observaties hoe de proefpersoon een taak aanpakt, of vrije associaties geuit door de proefpersoon naar aanleiding van inktvlekken of foto's, die vervolgens door de psycholoog worden geduid. Aanpassing van de testsituatie aan het te testen individu krijgt daarbij nogal eens de voorkeur boven objectieve vergelijking van individuen onder strikt identieke testomstandigheden. Dit dilemma is nog steeds actueel (zie b.v. Silva, 1993).

De Tweede Wereldoorlog, waarin negen miljoen Amerikanen met de Army General Classification Test werden onderzocht, gaf een impuls tot verfijning, zowel van de tests zelf als van de modellen om tests te scoren, te verbeteren en betrouwbaarder te maken. Om in die behoefte tot verfijning te voorzien werd in 1947 in Princeton NJ de Educational Testing Service opgericht; de tegenhanger CITO in Arnhem volgde pas in 1969. Uit tijdschriften zoals Psychometrika en Educational and

Psychological Measurement wordt duidelijk dat de ontwikkeling van statistische modellen voor psychologische tests tot circa 1975 vrijwel geheel een Amerikaanse aangelegenheid was; het werk van Rasch (1960) in Kopenhagen werd pas lang na publicatie naar waarde geschat, toen het bekend raakte via Fischer (1974) in Oostenrijk, Andersen (1980) in Denemarken en Wright & Stone (1979) in Chicago.
Het standaardwerk Lord & Novick (1968) markeerde de overgang van de klassieke testtheorie (geobserveerde somscore gesplitst in ware score plus meetfout, analyse gebaseerd op correlaties en varianties) naar de item response theorie (de kans op een positief antwoord op een item verklaard door itemparameter(s) en persoons-parameter, analyse via een model met latente variabelen voor de discrete datamatrix). Maar het markeert ook het begin van een periode waarin het aandeel van Europese, en in het bijzonder Nederlandse, auteurs binnen de statistisch-psychometrische modellenbouw spectaculair zou groeien.

In de periode 1970-79 bevat het belangrijkste tijdschrift op dit vakgebied, Psychometrika, ruim 90% artikelen waarvan de auteur werkzaam is in Noordamerika; in de periode 1983-94 daalt dat tot 64%. Het aandeel van Nederlandse auteurs stijgt van 2% naar 20%. Voor de rest van Europa gaat het van 5% naar 12%, vrij homogeen verdeeld over elf landen, en dan zijn er nog enkele artikelen uit Australië, Japan en Israel. Bij de auteurskeuze voor twee binnenkort te verschijnen handboeken (Van der Linden & Hambleton, 1995; Fischer & Molenaar, 1995) is Nederland eveneens sterk vertegenwoordigd. Binnen Nederland komen de bijdragen van het CITO en van vrijwel alle universiteiten, die in de Onderzoekschool IOPS nauw samenwerken.

Ik beschik niet over gegevens over het aantal toepassingen van tests in b.v. onderwijs, personeelbeleid of psychiatrie. Mijn globale indruk is dat de V.S., Canada en Australië dit klassement aanvoeren, op afstand gevolgd door Engeland, Japan, de Benelux, Oostenrijk, Duitsland en de Scandinavische landen. Het lijkt erop dat er minder testgebruik plaats vindt in Frankrijk, de Zuid- en Oosteuropese landen en de rest van de wereld.

3. MEETPRETENTIE EN ONZEKERHEID BIJ TESTGEBRUIK

Na dit uitstapje in tijd en ruimte keer ik terug naar de in de inleiding opgeworpen vragen. In deze paragraaf staat centraal of mentale eigenschappen zich wel met tests

laten meten, en welke meetfouten en onzekerheden daarbij een rol spelen.

In mijn eerste voorbeeld werden menselijk gewicht, lengte, lichaamstemperatuur en hemoglobinegehalte vergeleken met kennis van een eindexamenvak, intelligentie, stressbestendigheid of vatbaarheid voor hallucinaties en wanen. De eerstgenoemde groep van eigenschappen valt als volgt te kenmerken:

- het is vrij duidelijk wat met het begrip wordt bedoeld;
- er is voldoende kennis van observeerbare manifestaties om een meetinstrument te bouwen, dit goedkoop te repliceren, en de nauwkeurigheid van het instrument te bewaken;
- de gemeten persoon kan de meetuitkomst nauwelijks beinvloeden;
- er treedt weinig variatie van uur tot uur of dag tot dag op;
- de meting vereist weinig tijd en weinig deskundigheid;
- meting onder identieke omstandigheden is vrij eenvoudig.

Als we erg precies gaan kijken valt er op het bovenstaande wel wat af te dingen. Gewicht en hemoglobinegehalte op zeespiegelnivo of op 2000 meter hoogte (de moderne fysica onderscheidt massa van zwaartekracht, de moderne topsport kent hoogtestages)? Gewicht, lengte en temperatuur schommelen enigszins met het uur van de dag. Naakt of gekleed op de weegschaal, thermometer in de anus of onder de tong? De ene weegschaal is de andere niet. De gemeten persoon kan zich uitrekken om langer te lijken. Maar goed, deze zaken veroorzaken maar een kleine variatie en/of kunnen vrij makkelijk onder controle worden gehouden. Zowel de afleesfout als de natuurlijke variatie over korte tijdsperioden zijn doorgaans klein genoeg om voor de meeste meetdoeleinden niet terzake te doen. Zoals Wright & Stone (1979) opmerken, wordt maar zelden gevraagd op welke meetlat of onder welke meetomstandigheden iemand 178 cm lang was.

Hoe anders is dit bij een psychologische test. We volgen de definitie van Drenth & Sijtsma (1990, p.31): 'Een test is een systematische classificatie- of meetprocedure, waarbij het mogelijk wordt een uitspraak te doen over een of meer empirisch-theoretisch gefundeerde eigenschappen van de onderzochte of over specifiek niet-testgedrag, door uit te gaan van een objectieve verwerking van reacties van hem/haar, in vergelijking tot die van anderen, op een aantal gestandaardiseerde, zorgvuldig gekozen stimuli'. Het valt te verwachten dat de hierboven met zes gedachtenstreepjes aangeduide kenmerken voor de meting van de eerder genoemde mentale eigenschap-

pen nogal wat problemen lijken op te roepen.

Voor de eindexamenkennis Biologie van het Atheneum bestaat er wel een door de Minister van Onderwijs vastgesteld examenprogramma plus enige consensus over de uitwerking daarvan in de meest gebruikelijke leerboeken, en een instructie voor de tijdsduur van het examen en de correctie van het werk, maar de opgavenkeuze, de omstandigheden in de examenzalen en de scoring van het werk geven toch regelmatig aanleiding tot boze brieven. Brieven over onjuiste meting van lichaamslengte of lichaamstemperatuur heb ik nog nooit gezien; als er veel van afhangt (gewicht bij boksers, bloedsamen- stelling bij athleten of wielrenners) ontstaat overigens wel kritiek op meetomstandigheden.

Over de empirisch-theoretische fundering van het intelligentiebegrip zijn hele boekenkasten volgeschreven. Een testpsycholoog weet dan ook meestal welke deelscores op welke test het best kunnen worden gebruikt bij welke specifieke keuring of advisering. Variatie per uur of per dag, bijvoorbeeld door vermoeidheid of motivatieverlies, kan wel een storende invloed hebben, de proefpersoon kan zich van den domme houden of zich juist tevoren laten trainen, een compleet intelligentie-onderzoek duurt meestal enkele uren, en het testen van alle kandidaten onder dezelfde omstandigheden vereist speciale zorg.

Voor stressbestendigheid bestaat veel minder theoretisch en empirisch onderzoek naar de aard van het begrip, de geschiktheid van stimuli en de objectieve verwerking van reacties. Men kan werken met zelfbeoordeling op schriftelijke vragen, uitvoering van lastige opdrachten in een kleine groep, reactie bij verbale aanvallen door de psycholoog of bij het tonen van realistische videobeelden. Maar het blijft moeilijk uit een onvermijdelijk kunstmatige situatie nauwkeurig te voorspellen hoe iemand op straat of in de beroepspractijk zal reageren, en bovendien zijn er soorten van stressbestendigheid (al naar de aard van de situatie) zoals er soorten van intelligentie zijn (al naar de aard van het op te lossen probleem). Iemands stressbestendigheid kan met de tijd of de motivatie sterk uiteenlopen, en de proefpersoon kan de testscore ook hier opzettelijk beinvloeden.

De psychiatrie maakt vorderingen bij het streven naar een consensus over begripsvorming en diagnosevorming, maar er is vaak sprake van grote variatie in gedrag en stemming van de patient, van al dan niet bewuste pogingen om zich anders voor te doen, en van het dilemma dat informatie dikwijls op een onverwacht moment van rust

en vertrouwen beschikbaar komt, waarbij standaardisering en objectivering van de meting nauwelijks mogelijk zijn. Naast informatie van de behandelende psychiater of psycholoog en van de verpleging wordt vaak ook informatie van de patient zelf of van het thuisfront (partner, ouders, huisgenoten) gebruikt, hoewel die sterk gekleurd kan zijn en bij vergelijking tussen patienten het probleem oproept dat de mate van kennis en betrokkenheid tussen informanten sterk uiteenloopt. Als statistisch consulent ben ik nogal eens betrokken geweest bij constructie, validering en afname van tests bij patienten, en eigenlijk ben ik telkens verbaasd dat er ondanks de formidabele obstakels bij de dataverzameling toch nog repliceerbare en interpreteerbare schalen worden gevonden, die met feitelijk gedrag en met de directe indruk van deskundigen vrij hoog samenhangen. Het voorspellen van toekomstig gedrag, reactie op behandeling of ziekteverloop bij afzien van behandeling is in de psychiatrie uitermate lastig, of men nu wel of niet van gestandaardiseerde tests gebruik maakt. Vooral bij diagnosevorming en bij communicatie binnen de staf over het ziektebeeld zijn zulke tests wel een nuttig hulpmiddel, dat dan ook op vrij grote schaal wordt gebruikt.

We moeten concluderen dat de taak die een psychologische test moet vervullen niet gering is: het te meten begrip is dikwijls nogal breed en vaag, de uitkomst zal vaak met het tijdstip en de omstandigheden fluctueren, zowel het ontwerpen als het gebruiken van het meetinstrument vereist veel tijd en deskundigheid, de consequenties van de meting zijn meestal drastisch en de gemeten persoon heeft dus aanleiding om gebreken in de meetprocedure met veel kritiek te begroeten. Het is geen wonder dat testuitkomsten vaak lang niet zo nauwkeurig zijn als de psycholoog, de gemeten proefpersoon en de samenleving zouden willen, en dat dit bijdraagt aan de slechte reputatie van tests.

Het is een bekend probleem dat de nauwkeurigheid van een psychometrische testscore niet kan worden opgevoerd door een duplometing in de strikte zin. Bij examens en intelligentietests zijn de opgaven immers de tweede keer tevoren bekend, en bij het meten van houdingen en persoonlijkheidstrekken kunnen herinnering en motivatieverlies er evenzo voor zorgen dat we geen onafhankelijke en identiek verdeelde tweede waarneming kunnen doen. Wordt de tweede meting uitgesteld totdat we vermoeden dat de herinnering aan de stimuli is verdwenen (niemand weet hoe lang dat is!) dan valt een verandering van de ware score (door groei van kennis of verandering van mening) niet meer te onderscheiden van een verandering door een

andere meetfout bij dezelfde ware score.

Wel geldt in alle statistische modellen dat het gebruik van meer stimuli tijdens de testsessie tot nauwkeuriger meting leidt. Dit veronderstelt wel dat er nog meer geschikte stimuli aanwezig zijn en dat de test-tijd ongestraft kan worden uitgebreid. Omdat de meeste tests toch al lang duren, is het van groot belang dat de meest geschikte items worden gekozen. Binnen de item response theorie heeft men met succes gebruik gemaakt van discrete optimaliseringstechnieken om uit een voorraad van eerder onderzochte items (stimuli) juist diegene te selecteren die binnen practische randvoorwaarden tot een optimale meting leiden, hetzij item voor item afhankelijk van de scores tot nu toe (adaptief, Weiss 1982; Schoonman, 1989) hetzij simultaan voor de gehele test (Theunissen, 1985; Timminga & Adema, 1995).

De psychometrie heeft op diverse manieren en via diverse modellen inhoud gegeven aan begrippen zoals betrouwbaarheid en meetfout. Vergeleken met de fysieke metingen uit de inleiding van dit artikel is de betrouwbaarheid relatief laag en de meetfout relatief groot; enkele oorzaken hiervoor zijn hierboven genoemd. Het is een verdienste van de psychometrie, en meer algemeen van de statistiek, dat variabiliteit duidelijk aan het licht komt en onzekerheid van schattingen en conclusies getalsmatig kan worden aangegeven. Dat maakt ons niet populair bij de klant en de samenleving die zekerheid wensen, maar het maant hen tot enige voorzichtigheid. Soms komen er ook suggesties uit voort om de meetnauwkeurigheid te verbeteren en de onzekerheid te verminderen. Maar uitingen van menselijke capaciteiten en menselijk gedrag worden door een veelheid van onvoorziene omstandigheden beinvloed. De gemeten persoon heeft - gelukkig - een vrije wil om bij allerlei prikkels uit zijn/haar omgeving de gedragswijze te kiezen die hem/haar invalt of het beste voorkomt. Mijn conclusie is dat er nog wel wat aan de meetnauwkeurigheid van tests kan worden verbeterd, maar dat het grootste deel van de variantie van structurele aard is.

Dit is niet de plaats om in detail de wiskundige vorm te beschrijven van de verschillende statistische modellen voor tests, of de manier waarop in deze modellen parameters worden geschat en de passing van het model onderzocht. Inleidingen hierover zijn b.v. Allen & Yen (1979), Wright & Stone (1979), Molenaar (1981, 1992). Drenth & Sijtsma (1990) behandelen allerlei aspecten van de psychologische test, waaronder ook de voornaamste modellen. Eggen & Sanders (1993) geven een breed

overzicht van het analyseren en construeren van studietoetsen. Hambleton & Swaminathan (1985) en Van der Linden & Hambleton (1995) bestrijken vrijwel de gehele item response theorie, en Fischer & Molenaar (1995) is een actueel handboek over het Rasch model, dat zowel conceptueel als technisch als de interessantste vorm van item response theorie kan worden beschouwd. Hofstee (1983) geeft een heldere niet-technische inleiding op de problemen bij selectie door werkgevers en scholen; Cohen e.a. (1988) is een lijvig en vlot geschreven compendium van tests, testbegrippen en testtoepassingen in Amerika, waarin de statistische modellering er helaas nogal bekaaid afkomt.

4. TESTS EN SAMENLEVING: VERZET EN ACCEPTATIE

Het meten van mentale eigenschappen is in een flink aantal landen in de laatste honderd jaar op grote schaal in zwang geraakt bij examens, school- en beroepskeuze, personeelselectie, diagnose van afwijkend gedrag en wetenschappelijk onderzoek in de gedrags- en maatschappijwetenschappen. Dit heeft allerlei vormen van verzet en kritiek opgeroepen. Anderzijds wordt door de voorstanders betoogd dat dankzij de tests in de genoemde sectoren doorgaans met minder kosten betere beslissingen zijn genomen, en dat zowel de geteste personen als de samenleving daarvoor erkentelijk zou moeten zijn. Wie heeft er nu volgens mij gelijk?

Dikwijls wordt gesteld dat ieder mens uniek is en dat het niet mogelijk is zijn/haar functioneren in getallen uit te drukken en mensen op basis van die getallen onderling te vergelijken. Het antwoord kan kort zijn: volgens mij is dat inderdaad niet mogelijk. Maar het is ook niet nodig voor de diverse doelen waarvoor tests worden gebruikt. Het gaat er daarbij immers om, voor een welomschreven en beperkt gebied van menselijke vermogens of menselijk gedrag vast te stellen, op basis van een zo eerlijk mogelijke meting, wat iemand kan en doet op dit beperkte gebied. Vervolgens wordt die informatie, vrijwel steeds aangevuld met andere informatie omtrent de persoon, gebruikt voor een beslissing of advies. Dat de getalsmatige informatie geen recht doet aan de rijkdom van iemands persoonlijkheid is onomstreden. Maar dat is ook niet nodig, en vaak zelfs ongewenst. Onze samenleving vindt overwegend, en naar mijn mening terecht, dat de beslissing om een diploma of een betrekking te weigeren niet mag berusten op informatie over iemands ras, sexe of godsdienst. In Amerika is rond

1980 in jurisprudentie vastgelegd dat een werkgever alleen tests mag gebruiken waarvan is aangetoond dat de testuitslag voorspellende waarde heeft voor het functioneren in de beoogde functie. Ook wordt veel aandacht besteed aan een cultuur-eerlijke selectie van de testitems: als een opgave voor een vrouw moeilijker is dan voor een op de te meten eigenschap even begaafde man, of een attitudevraag voor een Moslim een ander verband heeft met de te meten eigenschap dan voor een Christen, dan dient dit item verwijderd te worden alvorens de test in een gemengde groep wordt gebruikt. Terzijde wordt opgemerkt dat het vaststellen van zulke DIF (differential item functioning) een vrij uitvoerig onderzoek met geavanceerde statistische modellen vereist, en dat aan dit aspect in de afgelopen tien jaar in de V.S. maar ook in Nederland veel aandacht is besteed.

Een tweede groep van bezwaren richt zich op de gebreken van de meetprocedure. Die zijn in de vorige paragraaf uitvoerig aan de orde geweest. Er wordt nog voortdurend vooruitgang geboekt bij de keuze van de stimuli, de inrichting van de testomstandigheden, de correctie voor ongewenste invloeden en de statistische verwerking van de testresultaten. Maar de samenleving zal volgens mij twee onaangename waarheden moeten accepteren. De eerste is dat tests altijd een vrij grote meetfout zullen houden en maar in beperkte mate toekomstig gedrag kunnen voorspellen. De tweede is dat men wel kan proberen zonder tests, bijvoorbeeld via expert-oordelen, te examineren, sollicitanten te selecteren, beroepskeuze te adviseren of psychiatrische diagnoses te stellen, maar dat uit zorgvuldig empirisch onderzoek keer op keer is gebleken dat daarbij in doorsnee nog grotere voorspelfouten worden gemaakt (Meehl, 1954; Goldberg, 1968; Hofstee, 1974). In de dagelijkse beroepspractijk blijven die fouten bij de meer intuitieve methoden meestal verborgen, terwijl men via de psychometrie verantwoorde, zij het wijde, betrouwbaarheidsintervallen krijgt. Veel statistici zullen dit herkennen: als zij aangeven hoe weinig we uit beperkte data verantwoord kunnen concluderen, wordt dat meestal niet in dank afgenomen.

In tabel 1, ontleend aan Altink (1992), wordt aangegeven in welke mate 70 Nederlandse bedrijven gebruik maakten van diverse selectiestrategieen voor drie soorten functies, en welke voorspellende waarde (correlatie, validiteit) voor latere arbeidsprestaties men volgens een pooling van diverse Amerikaanse onderzoekingen per methode ongeveer mag verwachten. Merk op dat geen van de validiteiten hoog is, maar dat referenties wel erg weinig kunnen voorspellen. Gelukkig wordt dikwijls een

combinatie van methoden gebruikt. Als een organisatie betere sollicitanten weet te selecteren, die bijvoorbeeld 10 procent productiever zijn, is de meeropbrengst al gauw vele malen hoger dan de testkosten. Vandaar de raad: altijd testen bij selectie (Schoonman, 1989, 1990; Van der Maesen de Sombreff, 1992).

Tabel 1. Gebruiksfrequentie van enkele selectiestrategieën bij 70 Nederlandse bedrijven en hun uit diverse Amerikaanse onderzoeken geschatte validiteit

methode	funktieniveau			validiteit
	hoog	middel	laag	
Psych. Capaciteitentests	61%	39%	8%	0.53
Gestructureerd interview				0.45
Open interview				0.14
Assessment Center	10%	7%	0%	0.36
Arbeidsproef	7%	8%	28%	0.46
Referenties	74%	68%	41%	0.26

Met het voorbeeld van weging van boksers en bloedonderzoek bij athleten heb ik al aangeduid dat kritiek op de meting al gauw losbarst als het meetresultaat ingrijpende gevolgen heeft. Waar een psychologische test wordt gebruikt is dat meestal het geval. Examens, sollicitaties en psychische problemen zijn voor de betrokkene en voor zijn/haar naaste omgeving ingrijpende gebeurtenissen. Weliswaar geldt dit ook voor medische keuringen en chirurgische operaties, maar als daar metingen aan te pas komen zijn die vaker onomstreden (thermometer) of zo technisch dat kritiek ongepast lijkt (hemoglobinemeter, PET-scan).

Voor kritiek op de dokter, die overigens wel in omvang toeneemt, is meestal evident een contra-expertise nodig, terwijl de psycholoog meer spontane kritiek uitlokt. Hoewel iedereen zowel een lichaam als een geest heeft, geeft menigeen de beoordeling en de zorg omtrent het eigen lichaam gemakkelijker in handen van de arts dan de beoordeling en de zorg omtrent geestelijke vermogens in handen van de

psycholoog. Uitingen zoals "ik verdien te slagen voor Biologie", "ik heb genoeg in huis voor die baan", "ik ben toch niet gek" liggen voor in de mond, terwijl "mijn bloed deugt wel voor donatie" of "mijn lever functioneert prima" al gauw een beetje belachelijk lijken. Kunnen we echt beter meepraten over onze geest, of is de psychologie nog onbeholpen in vergelijking met de geneeskunde? Die bestond al toen haar patienten onmondig en onkundig waren en heeft een traditie van respect voor haar kundigheid waar de veel jongere psychologie jaloers op kan zijn. Het kan ook zijn dat de arts meer vertrouwen inboezemt omdat hij uiteindelijk over leven of dood gaat, en in sommige gevallen levensreddende ingrepen in huis heeft. Anderzijds zijn er denk ik meer onbetrouwbare medische metingen en medische diagnoses dan ons lief is, maar het lijkt soms wel alsof noch de arts noch de patient deze vuile was graag buiten ziet hangen.

In beide beroepen gaat het niet zelden over ingrijpende beslissingen over mensen. Beide beroepen kennen dan ook een ethische code, al is die bij artsen beter bekend en verder uitgewerkt. Zoals een nieuw geneesmiddel op werkzaamheid wordt beproefd alvorens te worden toegelaten, onderzoeken de American Psychological Association en het Nederlands Instituut van Psychologen de nieuwe tests op kwaliteit, waarbij ook richtlijnen over toepassing en rapportage onder de beroepsbeoefenaren worden verspreid.

Het gedachten-experiment "beslissen zonder tests" leed schipbreuk op de aantoonbare inferioriteit van het ongewapend expert-oordeel. Het nog radicalere experiment "niet beslissen" is nog niet genoemd. Dat is een variant op "geen gezeik, iedereen rijk": iedereen krijgt een diploma, alle sollicitanten worden aangenomen, de kliniek geeft iedereen de diagnose "normaal" (behalve degenen die niet naar huis willen; die noemen we "gestoord" en ze mogen blijven). De ongewenste gevolgen zijn zonneklaar. Toch loont het de moeite er eens over na te denken, omdat dit duidelijk maakt dat er in onze voorbeelden eigenlijk drie partijen zijn: de school/het bedrijf/de kliniek, de gemeten persoon en de samenleving. De psycholoog is gehouden de belangen van alle drie in het oog te houden bij zijn/haar professionele werk. Soms kennen de arts, de accountant, de advocaat en niet te vergeten de statisticus soortgelijke ethische problemen.

Wijlen Melvin Novick, een inspirerende collega, heeft me geleerd dat in zulke gevallen een open dialoog over de minst slechte oplossing moet worden nagestreefd.

De psycholoog of statisticus doet er goed aan de beschikbare kennis, maar ook de onvermijdelijke onzekerheid, boven tafel te krijgen en de voor- en nadelen van diverse alternatieven te schetsen. Het is eerder aan de samenleving, waar nodig de wetgever, om luisterend naar de vereniging van beroepsbeoefenaren de grenzen te stellen waarbinnen het specifieke conflict van waarden en normen dient te worden opgelost.

Literatuur

Allen, M.J. & Yen, W.M. (1979). Introduction to Measurement Theory. Monterey: Brooks/Cole.

Altink, W.M.M. (1990). De plaats van psychologische tests binnen het moderne personeelsbeleid. Congresbundel "Psychologische tests: trends en toepassingen. Nederlands Studiecentrum, mei 1990.

Andersen, E.B. (1980). Discrete Statistical Models With Social Science Applications. Amsterdam: North-Holland.

Cohen, J.R., Montague, P, Nathanson, L.S. & Swerdik, M.E. (1988). Psychological Testing: An introduction to Tests & Measurement. Mountain View: Mayfield.

Dousma, T. & Horsten, A. (1994). Tentamineren. Groningen: Wolters-Noordhoff.

Drenth, P.J.D. & Sijtsma, K. (1990). Testtheorie: Inleiding in de theorie van de psychologische test en zijn toepassingen. Houten: Bohn, Stafleu Van Loghum.

Eggen, T.J.H.M. & Sanders, P.F. (1993). Psychometrie in de praktijk. Arnhem: CITO.

Fischer, G. (1974). Einführung in die Theorie psychologischer Tests. Bern: Huber.

Fischer, G.H., & Molenaar, I.W. (eds.) (1995). Rasch Models: Foundations, Recent Developments and Applications. New York: Springer.

Goldberg, L.R. (1968). Simple Models or Simple Processes? Some research on clinical judgement. American Psychologist 23, 483-496.

Hambleton, R.K. & Swaminathan, H.S. (1985). Item Response Theory: Principles and Applications. Boston: Kluwer-Nijhoff.

Hofstee, W.K.B. (1974). Psychologische uitspraken over personen: beoordeling, voorspelling, advies, test. Deventer: Van Loghum Slaterus.

Hofstee, W.K.B. (1983). Selectie: begrip, theorie, procedures en ethiek. Aula Pocket 736. Utrecht: Het Spectrum.

Lord, F.M. & Novick, M.R. (1968). Statistical Theories of Mental Test Scores.

voorspelling, advies, test. Deventer: Van Loghum Slaterus.

Hofstee, W.K.B. (1983). Selectie: begrip, theorie, procedures en ethiek. Aula Pocket 736. Utrecht: Het Spectrum.

Lord, F.M. & Novick, M.R. (1968). Statistical Theories of Mental Test Scores. Reading, MA: Addison Wesley.

Meehl, P.E. (1954). Clinical versus Statistical Prediction. Minneapolis: Univ. Minn. Press.

Molenaar, I.W. (1982). Mensen die het beter meten. Kwantitatieve Methoden 5, 3-29.

Molenaar (1992) Statistical models for educational testing. In: Van der Heijden, P.G.M., Jansen, W., Francis, B. & Seeber, G.U.H. (eds), Statistical Modelling, 249-262. Amsterdam: North-Holland.

Rasch, G. (1960). Probabilistic Models for Some Intelligence and Attainment Tests. Chicago: The Univ. of Chicago Press.

Schoonman, W. (1989). An Applied Study on Cumputerized Adaptive Testing. Amsterdam, Lisse: Swets & Zeitlinger.

Schoonman, W. (1990). Altijd testen bij selectie. De Psycholoog, 25, 526-528.

Silva, F. (1993). Psychometric Foundations and Behavioral Assessment. Newbury Park CA: Sage.

Theunissen, T.J.J.M. (1985). Binary Programming and Test Design. Psychometrika 50, 411-420.

Thurstone, L.L. (1928). Attitudes can be measured. American Journal of Sociology 33, 529-544.

Timminga & Adema (1995). Test construction from item banks. In: Fischer, G.H., & Molenaar, I.W. (eds.), Rasch Models: Foundations, Recent Developments and Applications, 109-125. New York: Springer.

Van der Linden, W.J., & Hambleton, R.K. (eds) (1995). Handbook for modern item response theory. New York: Springer.

Van der Maesen de Sombreff, P.E.A.M. (1992). Het rendement van personeelsselectie. Proefschrift R.U. Groningen.

Weiss, D.J. (1982). Improving Measurement Quality and Efficiency with Adaptive Testing. Applied Psychological Measurement, 6, 473-492.

Wright, B.D. & Stone, M.H. (1979). Best Test Design. Rasch measurement. Chicago: Mesa Press.