De waan van "het" IQ

Peter Tellegen
Persoonlijkheids- en Differentiële Psychologie, RuG, 2 juni 2004

"Onze zoon heeft de SON-R 2,5-7 gedaan en kwam uit op een IQ van 69, hij heeft hierbij ook PDD-NOS. Nu zijn we op zoek naar een goede school voor onze zoon, die hebben we ook gevonden en nu blijkt dat hij dan 1 puntje te weinig heeft om toegelaten te worden. Een IQ van 70 mag wel, een IQ van 69 mag niet. Is het wel de bedoeling om zo zwaar met een intelligentietest (wat een momentopname is en bij kinderen met PDD-NOS nogal moeilijk ligt) beoordeeld te kunnen worden."

Inleiding

In selectie-situaties binnen het onderwijs waarbij met vaste grenzen voor IQ-scores wordt gewerkt, wordt aan de IQ-score veelal een te absolute betekenis toegekend. Er wordt geen rekening gehouden met verschillen tussen tests, of met de omstandigheden waaronder de test is afgenomen. Ten onrechte worden de IQ-scores geacht onderling inwisselbaar te zijn. Ofschoon in theorie erkend wordt dat de uitkomst kan zijn beïnvloed door meetfouten, wordt het effect daarvan zo laag ingeschat dat daarmee in de selectieprocedure geen rekening wordt gehouden. In dit opzicht wordt het IQ-*getal* veel te absoluut geïnterpreteerd. Hetzelfde geldt voor classificatie van IQ-scores waarbij aan vastgelegde intervallen van IQ-scores een specifiek voorgeschreven betekenis wordt toegekend. Zo dienen personen met een IQ-score van 50 t/m 69 te worden omschreven als 'licht zwakzinnig' (Resing en Blok, 2002). Dus ook hier wordt een 'absolute' waarde aan de score toegekend die dan bepalend is voor de betekenis en wordt er geen rekening gehouden met de test die is afgenomen, met andere omstandigheden, of met de onbetrouwbaarheid.

De illusie van absolute betekenis, en daarmee verbonden de onderlinge inwisselbaarheid van IQ-scores, wordt in de hand gewerkt door de benaming "IQ", die wordt gebruikt ongeacht met welke test deze score is verkregen. De overeenkomst in uitgangspunten bij de normering, normaal verdeelde scores in de referentiepopulatie met een gemiddelde van 100 en een standaarddeviatie van 15, wekt ook ten onrechte de suggestie dat de betekenis van de score identiek is. Hierna zal een overzicht gegeven worden van belangrijke factoren die IQ-scores zodanig kunnen beïnvloeden, dat deze scores veel minder goed vergelijkbaar zijn dan veelal wordt verondersteld.

Voordat wordt ingegaan op beperkingen die inherent zijn aan het beoordelen van IQ-scores, is het goed om op het volgende te wijzen. Selectieprocedures zoals die nu bijvoorbeeld gehanteerd worden ten behoeve van toelating tot het speciaal onderwijs, en ook de classificatie van IQ-scores, wekken door het hanteren van vaste procedures en beslissingsregels, gerelateerd aan een getal, het IQ, de indruk van objectiviteit en wetenschappelijkheid. In feite echter zijn de beslissingsregels arbitrair tot stand gekomen. Een wetenschappelijke benadering van selectie vereist dat op grond van empirische criteria wordt onderzocht wat de optimale beslisregels zijn. Dat deze empirische onderbouwing in de huidige praktijk ontbreekt blijkt al uit de ronde getallen - 10-tallen of 5-tallen - die bij deze procedures worden gehanteerd. Zo dient het IQ lager dan 70 te zijn, of hoger dan 85, maar een grens van 72 of 93 wordt niet gehanteerd. Het gaat hierbij dan ook om criteria die vanachter het bureau worden bedacht en die niet eerst op zinvolheid worden getoetst.

Zo er, bijvoorbeeld, al een onderzoek zou zijn waaruit 15 jaar geleden bleek dat het bij test X zinvol is om bij IQ-scores onder de 70 te spreken van "licht zwakzinnig", dan is het zeer de vraag of deze grens ook zou moeten gelden voor test Y, en ook is het de vraag of nu, 15 jaar later, bij test X nog dezelfde grens van 70

gehanteerd zou moeten worden. Daarom is ook de objectiviteit van de procedure schijn; men hanteert wel vaste getallen maar onderkent niet dat de betekenis van deze getallen niet constant is en tussen tests kan verschillen.

Beperkende factoren

De volgende factoren stellen beperkingen aan de mogelijkheid om IQ-scores onderling te vergelijken en om deze absoluut te interpreteren.

Betrouwbaarheid

Betrouwbaarheid is de meest bekende beperkende factor bij de interpretatie van testscores. De betrouwbaarheid wordt standaard bij intelligentietests vermeld en vaak worden intervallen aangegeven die gebaseerd zijn op de betrouwbaarheid. Wel is er veel onduidelijkheid met betrekking tot het onderscheid tussen betrouwbaarheids- en waarschijnlijkheidsintervallen en is het veelal onduidelijk of deze intervallen al dan niet symmetrisch rond de geobserveerde score moeten worden geconstrueerd (zie Snijders, Tellegen & Laros, 1988; Evers, 2001).

De betrouwbaarheid geeft de overeenstemming in score aan met een strikt parallelle test. De 'ware' score is de hypothetische gemiddelde score bij afname van een onbeperkt aantal parallelle tests. De meetfout is de afwijking van de geobserveerde score ten opzichte van deze hypothetische ware score. Indien bij selectie een grenswaarde voor verwijzingen/beslissingen wordt gehanteerd dan zou deze grenswaarde idealiter betrekking moeten hebben op de ware score. Daarom zou men, als een grenswaarde wordt gebruikt, als eis kunnen stellen dat de geobserveerde IQ-score significant hiervan moet afwijken. Indien de score niet significant afwijkt dan is er onvoldoende informatie om tot een eenduidige beslissing te komen.

Op grond van de standaardmeetfout kan rond de grenswaarde een betrouwbaarheidsinterval worden geconstrueerd. Voor een grenswaarde van bijvoorbeeld 90, een relatief matige betrouwbaarheid van .90 en een tweezijdig significantieniveau van 5%, loopt dit interval van 81 tot en met 99. Voor zeer veel scores kan in deze situatie dus niet met voldoende zekerheid worden uitgemaakt of de ware score nu lager dan wel hoger is dan 90.

Indien de betrouwbaarheid vrij hoog is, bijvoorbeeld .95 en men bovendien gemakkelijker zou accepteren dat een onjuiste uitspraak wordt gedaan door een significantieniveau van 10% te hanteren, dan zou het interval lopen van 85 tot en met 95. Dit is de helft smaller dan het vorige interval maar nog steeds aanzienlijk en van praktische betekenis. Zo betekent het voor het classificatiesysteem dat door de COTAN naar voren wordt gebracht (Resing & Blok, 2002) en waarbij veel intervallen 10 IQ-punten breed zijn, dat voor vrijwel elke score geldt dat hij niet significant afwijkt van de voorgaande dan wel de opvolgende categorie. Ook in selectiesituaties zou men in veel gevallen moeten concluderen dat er onvoldoende informatie is om op een zorgvuldige wijze een beslissing te nemen.

Het is opvallend dat met betrekking tot de indicatiestelling en classificatie zo weinig aandacht wordt besteed aan (on)betrouwbaarheid en hoe men hier rekening mee zou moeten houden. Illustratief is hoe lichtvoetig in de recente uitgave van Algemene Psychodiagnostiek I (De Zeeuw, Dekker & Resing, 2004) met betrouwbaarheidsintervallen wordt omgegaan. Hierover zeggen de auteurs (p. 47): "De betrouwbaarheidsintervallen behorend bij bijvoorbeeld het 95% betrouwbaarheidsinterval zijn vaak nogal groot en daardoor in de praktijk van beperkte waarde" Nadat een interval is berekend dat loopt van 95–115 vervolgen de auteurs: "In de praktijk kan men hier niet veel mee: de ondergrens is bij dit kind gemiddeld en de bovengrens is een IQ op het niveau 'boven gemiddeld'. Het is natuurlijk zeer onwaarschijnlijk dat men er in de praktijk zover 'naast' zou zitten, maar dit is het statistische gegeven." Psychometrische argumenten om de testuitkomst te relativeren worden dus terzijde geschoven omdat men liever niet aanvaardt dat de nauwkeurigheid van de testuitkomst zo beperkt is.

Generaliseerbaarheid

In het kader van betrouwbaarheid wordt de vraag gesteld hoe de score er uit zou zien indien iedere subtest niet uit een *beperkt aantal items* zou bestaan, maar uit een zeer groot aantal items. Het corresponderende betrouwbaarheidsinterval geeft aan hoeveel onzekerheid dit geeft met betrekking tot uitspraken over het 'ware' IQ. Bij de generalisatie kan nog een niveau verder worden gegaan: nu wordt de vraag wat er zou gebeuren als de intelligentietest niet uit een *beperkt aantal subtests* zou bestaan maar uit een zeer groot aantal subtests. De subtests die zijn afgenomen worden hierbij beschouwd als een noodzakelijkerwijs beperkte steekproef uit het domein van relevante subtests. Deze maat van generaliseerbaarheid kan berekend worden met coëfficiënt alpha waarbij de genormeerde subtestscores de elementen van de berekening vormen.

De generaliseerbaarheid is in het algemeen aanmerkelijk lager dan de betrouwbaarheid. Zo is bij de SON-R 5,5-17 de betrouwbaarheid gemiddeld .93 en de generaliseerbaarheid gemiddeld .85 (Snijders, Tellegen & Laros, 1988). Bij de SON-R 2,5-7 zijn betrouwbaarheid en generaliseerbaarheid gemiddeld respectievelijk .90 en .78. (Tellegen, Winkel, Wijnberg-Williams & Laros, 1998). Indien men op grond van de uitkomst bij een specifieke test, uitspraken wil doen over 'het' intelligentieniveau, dan generaliseert men naar een meer algemeen niveau en het bijbehorende generaliseerbaarheidsinterval is dan ook veel breder dan het betrouwbaarheidsinterval. Voor de individuele beoordeling, maar zeker voor selectiesituaties waarbij bij de interpretatie van het IQ geen onderscheid tussen de verschillende intelligentietests wordt gemaakt, is het brede generaliseerbaarheidsinterval het meest relevant.

De betrouwbaarheid geeft de verwachte correlatie weer met een strikt parallelle test. De generaliseerbaarheid geeft de verwachte correlatie met een vergelijkbare testbatterij waarbij de subtests afkomstig zijn uit hetzelfde domein. Dat dit in de praktijk ook opgaat blijkt uit de correlatie van .81 tussen de WISC-R en de RAKIT bij een representatieve steekproef van ruim 400 leerlingen (Bleichrodt, Resing, Drenth & Zaal, 1987). Deze correlatie is aanzienlijk lager dan de betrouwbaarheid van deze tests (.94/.95) maar komt wel goed overeen met de generaliseerbaarheidscoëfficiënt (WISC-R:.81; RAKIT: .78; zie Snijders, Tellegen & Laros, 1988, p. 85). Indien beide tests een gemiddelde zouden hebben van 100 en een standaarddeviatie van 15 dan betekent dit dat de standaarddeviatie van de verschilscores gelijk is aan 9.25. Dit heeft als gevolg dat wanneer een kind getest is met de WISC-R of met de RAKIT, er een kans is van 5% dat de IQ-score op de andere test 19 punten of meer zal afwijken. De kans dat de score op de andere test 10 punten of meer afwijkt is groter dan 30%. Let wel, deze verschillen hebben betrekking op betrouwbare en goed geconstrueerde intelligentietests waarbij de afname onder gecontroleerde condities plaatsvond en het moment van afname niet sterk uit elkaar lag.

Zeer grote verschillen in IQ-scores zijn ook gevonden bij een onderzoek onder vierjarige kinderen die zowel met de SON-R 2,5-7 als met het performale deel van de WPPSI-R zijn getest. Bij twee procent van de kinderen was het verschil tussen SON-IQ en WPPSI-PIQ ongeveer 40 punten (Tellegen et al., 1998, p. 112).

Definitie normpopulatie

De IQ-score geeft weer hoe hoog/laag de score is ten opzichte van de normpopulatie. Een andere definitie van de normpopulatie geeft dan ook andere scores. Dit bleek bijvoorbeeld bij de vergelijking van de genormeerde scores van de WISC-R en de RAKIT. Aangezien bij de RAKIT leerlingen van het speciaal onderwijs buiten de normpopulatie waren gehouden wijkt de verdeling af van de WISC-R waar leerlingen van het speciaal onderwijs wel tot de normgroep behoren. Als gevolg van deze omissie vallen de laagste scores bij de RAKIT ongeveer 7 IQ-punten te laag uit (Tellegen, 2002a).

Bij de handleiding van de RAKIT wordt ook een aanbeveling gedaan voor de classificatie van testscores. Volgens deze classificatie (Bleichrodt, Drenth, Zaal & Resing, 1987, p. 21) zijn leerlingen met een IQ van 70-79 'debiel' en zijn leerlingen met een IQ lager dan 70 'zwakzinnig'. Aangezien de normering van de RAKIT beperkt is tot leerlingen van het reguliere basisonderwijs zou dit betekenen dat ongeveer 10% van de leerlingen van het reguliere basisonderwijs debiel of zwakzinnig is.

Eerder is aangegeven dat in ongeveer 30% van de gevallen de scores op de WISC-R en de RAKIT meer dan 10 IQ-punten van elkaar zullen verschillen. Deze berekening ging er echter van uit dat de

scoreverdelingen gelijk zijn. Aangezien dat door verschillende definities van de normpopulatie niet het geval is, zullen de onderlinge verschillen in scores nog groter zijn en nog frequenter voorkomen.

Representativiteit normgroep
De verdeling van de IQ-scores is bij de RAKIT afwijkend omdat een ongebruikelijke definitie van de normpopulatie is gehanteerd. Binnen het kader van deze definitie is de steekproef wel zorgvuldig getrokken en representatief. Het kan echter ook zo zijn dat scores op intelligentietests op grond van de overeenkomstige definitie van de normpopulatie vergelijkbaar zouden moeten zijn, maar dat in de praktijk niet zijn omdat de steekproef niet representatief is. Indien deze afwijkingen goed gedocumenteerd zijn, kan men daar nog enigszins rekening mee houden. Als in de handleiding geen, onjuiste, of onvolledige informatie wordt gegeven, kan eigenlijk niet beoordeeld worden of, en op welke wijze, de normen zullen afwijken. Dit bleek het geval bij de recente herziening van twee van de meest gebruikte intelligentietests zoals de WAIS-III (zie Span, 2002; Tellegen 2002bc, 2003a)
en de WISC-III (zie Köhler 2002; Tellegen, 2002d, 2003b, 2004a).

Nauwkeurigheid normen
Ook bij een goed getrokken steekproef en een adequate definitie van de normpopulatie moet rekening worden gehouden met systematische afwijkingen die te maken hebben met de omvang van de steekproef en de methode die is gevolgd om de scoreverdelingen te normaliseren.

Bij intelligentietests worden de normeringen meestal per leeftijdsgroep uitgevoerd. Bij individueel af te nemen tests zijn aantallen van 100-200 per normgroep gebruikelijk. Bij normale verdelingen met gemiddelde 100 en standaarddeviatie 15 zullen de systematische afwijkingen van het gemiddelde bij deze steekproefomvang in 5% van de gevallen tenminste 2 tot 3 punten bedragen. Als dan de standaarddeviatie op grond van de beperkte steekproefomvang ook nog 2 punten afwijkt, dan zijn de resulterende extreme IQ-scores aan een zijde van de verdeling 9 punten te hoog of te laag.

Als de scoreverdelingen niet normaal zijn verdeeld, en de normalisatie op grond van cumulatieve percentages wordt uitgevoerd, is het nog moeilijker om de meer extreme scores nauwkeurig te bepalen. Fit-methodes kunnen de nauwkeurigheid verbeteren maar worden nog weinig toegepast (Laros & Tellegen, 1991, p. 156). In het geval van leeftijdsgroepen is het continue normeringsmodel dat is ontwikkeld in het kader van de normering van de SON-tests, de ideale methode aangezien dan gelijktijdig en optimaal van de informatie van alle leeftijdsgroepen gebruik wordt gemaakt en de normen op de exacte leeftijd kunnen worden gebaseerd (Snijders, Tellegen & Laros, 1988; Tellegen et al., 1998).

Leeftijdsgroepen
Bij de meeste intelligentietests worden normen per leeftijdsgroep berekend. Deze norm is dan correct voor het midden van het leeftijdsinterval, maar aan het begin van het interval wordt het niveau van het kind onderschat en aan het eind van het interval wordt het niveau overschat. Vooral bij jongere kinderen, en wanneer relatief brede intervallen worden gebruikt, kan de systematische afwijking gemakkelijk 3 tot 5 IQ-punten bedragen. Dit betekent dat op de grens van het leeftijdsinterval, beoordeling volgens de ene of volgens de andere normgroep een verschil kan betekenen van 6 tot 10 punten. Zie Tellegen (2002b, 2004a) voor uitgewerkte voorbeelden met de WAIS-III en WISC-III.

Ofschoon dergelijke sterke afwijkingen zich alleen voordoen bij een deel van degenen die onderzocht worden, is het merkwaardig dat dit probleem niet beter wordt aangepakt. Deze fouten zijn namelijk te verhelpen met een gecomputeriseerde normering gebaseerd op de exacte leeftijd, maar kunnen ook eenvoudig worden opgelost door extra normtabellen in de testhandleiding op te nemen. Opmerkelijk is dat het NIP-Dienstencentrum heeft verklaard dat het dergelijke fouten niet wil aanpakken, ondanks eerdere toezeggingen. De reden hiervoor is door de directeur van het NIP in een brief aan de kopers van de WISC-III als volgt verwoord: "Een belangrijke overweging hierbij is dat bij een verdere verfijning ten onrechte zou worden gesuggereerd dat het mogelijk is tot een zeer nauwkeurige aanduiding van het IQ te komen." Er

zullen weinig wetenschappen zijn waar het handhaven van systematische meetfouten tot doel wordt verheven (zie Tellegen, 2004).

Veroudering normen

Testnormen zijn aan verandering onderhevig hetgeen ertoe kan leiden dat na een aantal jaren de scores te hoog uitvallen in vergelijking met meer recente normen. Dit wordt wel het Flynn-effect genoemd naar degeen die dit uitgebreid heeft onderzocht. Een globale schatting van dit effect is 3 IQ-punten per 10 jaar (Flynn, 1984, 1987). Een probleem bij de beoordeling van dit effect is dat het meer betrekking lijkt te hebben op inzicht en minder op kennis, en dat weinig bekend is hoe, naast het gemiddelde, mogelijk ook de vorm van de scoreverdeling verandert. Verder zijn er aanwijzingen dat het effect in Nederland aan het afzwakken is (zie Tellegen et al., 1998, p. 115).

Diverse nog in gebruik zijnde intelligentiestests hebben sterk verouderde normen: voor de GIT zijn de normen meer dan 40 jaar oud; de WAIS meer dan 35 jaar; en de RAKIT en de WISC-R ruim 20 jaar. Ook de SON-R 5,5-17 die enkele jaren na de WISC-R en RAKIT is genormeerd, begint verouderd te raken. Zelfs indien men er in zou slagen van de belangrijke tests binnen 20 jaar nieuwe normeringen uit te brengen, dan kunnen, op grond van de veroudering van de normen, systematische verschillen in scores optreden van 0-5 IQ-punten. Een eerste stap om daarmee systematisch rekening te houden is gezet bij de SON-tests waarbij met IQ* een correctie wordt toegepast op het IQ, gebaseerd op een schatting van het Flynn-effect (Tellegen, 2004c). Het is de bedoeling deze schatting te verbeteren door vergelijking met andere goede recent in Nederland genormeerde intelligentietests.

Inhoudelijke en methodische verschillen tussen tests

Bij het punt van generaliseerbaarheid is aan de orde gekomen hoe uitkomsten op intelligentietests kunnen verschillen wanneer de subtests als onafhankelijke steekproeven uit hetzelfde domein van vaardigheden kunnen worden beschouwd. De onderlinge verschillen tussen de IQ-scores kunnen nog groter worden als bij een test op specifieke vaardigheden sterk de nadruk wordt gelegd, zoals verbale vaardigheid of tests met name gericht op ruimtelijk inzicht. Tests kunnen ook in andere opzichten verschillen. Zo kunnen deze bijvoorbeeld geordend worden op een dimensie die aangeeft in hoeverre het leren van nieuwe vaardigheden wordt beoordeeld dan wel dat beroep wordt gedaan op schoolse kennis. De volgende hypothetische ordening (van inzicht naar kennis) van diverse tests is op dat onderscheid gebaseerd: LEM – SON – RAKIT – WISC – NIO/Givo – CITO-eindtoets. Te verwachten valt dat tussen de scores op de LEM en de CITO-eindtoets relatief grote discrepanties zullen bestaan.

Testbias

Van testbias is sprake wanneer personen of groepen op grond van een oneigenlijk kenmerk benadeeld worden bij de afname van een test. Dit kan zich voordoen indien personen waarvan Nederlands niet de moedertaal is, toch worden getest met een Nederlandstalige test waarbij begrip en kennis van de Nederlandse taal een belangrijk onderdeel vormt bij de beoordeling van de intelligentie. Uit allerlei onderzoeken, onder meer met de RAKIT, de GATB en de NIO, blijkt dat allochtonen juist op de verbale onderdelen sterk achterblijven. Een methodologische eis bij intelligentietests is dat verbale testonderdelen betrekking dienen te hebben op de moedertaal (Carroll, 1993, p. 145). Dit uitgangspunt wordt in Nederland echter niet algemeen aanvaard (zie Evers & Te Nijenhuis, 1999; Tellegen, 2000, 2001; Te Nijenhuis & Evers, 2000). Dit leidt tot 'wetenschappelijke' stereotypen over het lage intelligentieniveau van allochtonen, zoals een gemiddeld IQ van de Turkse en Marokkaanse bevolking in Nederland van 78 (Te Nijenhuis & Van der Flier, 2001).

Uit het onderzoek met de NIO, een klassikaal af te nemen intelligentietest met deels verbale onderdelen (Van Dijk & Tellegen, 2004) blijkt dat juist de verbale onderdelen het onderwijsniveau onderschatten dat de allochtone leerlingen volgen. Daarom is het zo verwonderlijk dat de NDT (Van Hoorn, Van der Kamp & Den Brinker, 2003), een test die juist bedoeld is voor de lagere onderwijsniveaus waar relatief veel allochtonen onderwijs volgen, de intelligentiescore vrijwel geheel baseert op verbale

onderdelen. Gevreesd moet worden dat gebruik van deze test bij allochtone leerlingen zal leiden tot een sterke onderschatting van hun onderwijsmogelijkheden waardoor deze leerlingen op grond van niet valide criteria op een voor hen te laag onderwijsniveau terecht komen.

Overige factoren

Naast de hierboven genoemde factoren zijn er nog een groot aantal invloeden die kunnen maken dat testscores minder goed vergelijkbaar zijn, of die kunnen leiden tot een onjuiste interpretatie. Te denken valt aan testleidereffecten bij de afname, gezondheidssituatie bij de afname, moment van afname (tijdstip en dag van de week), zoek- en rekenfouten bij de verwerking (optellen scores, berekenen leeftijd, hanteren tabellen) en storende factoren bij de afname (bijvoorbeeld ongeschikte stoel, lawaai, slecht licht etc.).

Verder is in dit artikel ook niet ingegaan op het probleem van de stabiliteit van de intelligentie. Op grond van een eenmalige testafname wordt vaak een beslissing genomen die niet gemakkelijk kan worden aangepast terwijl het intelligentieniveau aan veranderingen onderhevig kan zijn.

Discussie

De veronderstelling dat mensen een IQ 'hebben' en dat op grond van dit IQ een belangrijke beslissing kan worden genomen, of een typering van de persoon kan worden gegeven, berust op een misvatting. De uitkomst van een intelligentietest kan alleen beschouwd worden als een *indicatie* van het intelligentieniveau. Deze indicatie is verre van nauwkeurig en de betekenis van de IQ-score is evenzeer afhankelijk van de test, ouderdom van normen en tal van andere, deels irrelevante factoren. Het eerste wat een psycholoog zou moeten vertellen als hij of zij de uitkomst van een intelligentietest meedeelt, is dat de IQ-score er ver naast kan zitten. Afwijkingen tot 10 punten naar boven en naar beneden zijn normaal en afwijkingen van 10 tot 20 punten komen zo frequent voor dat de kans daarop bij de interpretatie van de uitkomst een belangrijke rol dient te spelen. Het is daarom volstrekt onzinnig als men zou willen voorschrijven dat bij een IQ-score van 69 vermeld dient te worden dat de persoon 'licht zwakzinnig' is en bij een score van 71 dat de persoon 'laag begaafd' is. Dit alles ook ongeacht de test die is afgenomen.

De nauwkeurigheid van IQ-scores zou kunnen toenemen door betere normeringen hetgeen onder meer mogelijk is door toepassing van het continu normeringsmodel. Daarnaast zouden constructeurs de onderlinge vergelijkbaarheid van de testscores als een belangrijke doelstelling moeten zien waarvoor kruisvalididatie van normen noodzakelijk is. Dit biedt dan ook de mogelijkheid om correcties voor het Flynn-effect standaard te integreren bij de presentatie van testscores.

Men zou kunnen denken dat deze verbeteringen, die de mogelijkheid tot onderlinge vergelijking van IQ-scores vergroten, tot bij benadering absoluut nauwkeurige IQ-scores zouden kunnen leiden waardoor classificatie en indicatiestelling wel verantwoord kunnen worden uitgevoerd. Dit is tot op zekere hoogte het geval, maar juist wanneer men zich serieus bezig houdt met de verbetering van de nauwkeurigheid van IQ-scores, wordt ook duidelijk hoezeer het noodzakelijk is de betekenis van de testuitkomst te blijven relativeren.

De meeste intelligentietests, en dit geldt zeker voor de individueel af te nemen tests, zijn ontwikkeld als hulpmiddel voor deskundigen die verantwoordelijk zijn voor hulpverlening en advisering aan kinderen en volwassenen. Hierbij staat de behartiging van de belangen van de cliënt voorop. In deze context kan een intelligentietest goede diensten bewijzen mits de deskundige door opleiding en ervaring in staat is de waarde en de beperkingen van de test af te wegen en hij deze in de context kan plaatsen van de persoonlijke ervaring met de cliënt, andere aanvullende informatie, en de doelstellingen en mogelijkheden die in het geding zijn. In het praktijkvoorbeeld waarmee dit artikel begon waren de ouders, de psycholoog die de test had afgenomen, en de school, het er allemaal over eens dat dit type school voor dit kind geschikt was. Vanwege een verschil van 1 IQ-punt heeft het echter een half jaar geduurd voordat de ouders te horen kregen dat toelating tot deze school alsnog zou worden toegestaan.

In de huidige situatie wordt de psycholoog/diagnosticus iemand die gegevens aanlevert aan een commissie die moet beoordelen, op grond van arbitraire criteria die weer door anderen zijn vastgesteld, welke beslissing genomen dient te worden. Van deze mechanische benadering is niet alleen de cliënt de dupe, maar ook de maatschappij en ook de psycholoog/diagnosticus. Deze laatste kan zijn deskundigheid niet meer inbrengen en moet toezien dat op grond van de door hem aangeleverde getallen beslissingen worden genomen die berusten op beperkt valide testuitkomsten die in kunnen gaan tegen de belangen van de cliënt. In de praktijk zal de psycholoog soms naar wegen zoeken om dan maar gegevens te leveren die wel tot een zijns inziens juiste beslissing leiden (zie Brouwers, 2003; Woudenberg, 2004).

De situatie die is ontstaan betekent eigenlijk dat met de indicatiestelling en classificatierichtlijnen een einde dreigt te worden gemaakt aan verantwoorde diagnostiek en aan verantwoord testgebruik. Het zal geen toeval zijn dat deze ontwikkeling gepaard gaat met een verschuiving in de wijze waarop tests geconstrueerd worden. De psychologische test is niet meer primair een wetenschappelijk instrument maar een commercieel product. Hierbij wordt het geaccepteerd dat de test na de normering wordt veranderd, dat gebreken in de normering worden verzwegen en dat fouten bij de analyse worden genegeerd.

Van deze ontwikkelingen is in de eerste plaats degene die wordt getest en beoordeeld de dupe. Een goede voorlichting aan het 'publiek' over de beperkte waarde van intelligentietests en de schade die kan worden aangericht door de huidige testpraktijk, is noodzakelijk om te voorkomen dat de beroepsgroep straks in de hoek van kwakzalvers en beunhazen terecht komt. De psychologen die zich met diagnostiek bezig houden zijn evenzeer de dupe, van 'huisarts' worden zij 'keuringsarts', een verandering waar niemand bij is gebaat en die maar door weinigen wordt geambiëerd. Ten slotte is ook de overheid de dupe, die efficiëntie probeert te bereiken zonder enig zicht op de deugdelijkheid van de methode en het effect daarvan.

De wal keert het schip, en dat zal hier ook wel gelden, maar de schade die ondertussen wordt aangericht is niet te herstellen.

Literatuur

Bleichrodt, N., Drenth, P.J.D., Zaal, J.N. & Resing, W.C.M. (1987). *RAKIT Handleiding*. Lisse: Swets & Zeitlinger.

Bleichrodt, N., Resing, W.C.M., Drenth, P.J.D. & Zaal, J.N. (1987). *Intelligentie-meting bij kinderen*. Lisse: Swets & Zeitlinger.

Brouwers, G.W. (2003). Zogenaamde Classificerende Diagnostiek als opmaat naar bureaucratie en gesjoemel. *Tijdschrift voor orthopedagogiek, 42,* 396-398.

Carroll, J.B. (1993). *Human cognitive abilities. A survey of factor-analytic studies.* Cambridge: Cambridge University Press.

Dijk, H. van & Tellegen, P.J. (2004). NIO Nederlandse Intelligentietest voor Onderwijsniveau. Handleiding en Verantwoording. Amsterdam: Boom testuitgevers.

Evers, A. (2001). *Beoordelingssysteem voor de Kwaliteit van Tests*. Amsterdam: Cotan/NIP.

Evers, A. & Te Nijenhuis, J. (1999). Liever speciale dan traditionele cognitieve capaciteitentests voor allochtonen? Een vergelijking. *De Psycholoog, 34*, 250-255.

Flynn, J.R. (1984). The Mean IQ of Americans: Massive Gains 1932-1978. *Psychological Bulletin, 95*, 29-51.

Flynn, J.R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*, 171-191.
Hoorn, W. van, Kamp, L. van der & Brinker, W. den (2003). *Nederlandse Differentiatie Testserie NDT-2003. Handleiding*. Lisse: Swets & Zeitlinger.

Kort, W., Compaan, E.L., Bleichrodt, N., Resing, W.C.M., Schittekatte, M., Bosmans, M., Vermeir, G. & Verhaeghe, P. (2002). *WISC-III NL. Handleiding*. London: The Psychological Corporation.

Laros, J.A. & Tellegen, P.J. (1991). *Construction and validation of the SON-R 5,5-17, the Snijders-Oomen non-verbal intelligence test*. Groningen: Wolters-Noordhoff.

Resing, W.C.M. & Blok, J.B. (2002). De classificatie van intelligentiescores: voorstel voor een eenduidig systeem. *De Psycholoog, 37*, 244-249.

Resing, W.C.M., Evers, A., Koomen, H.M.Y., Pameijer, N.K., Bleichrodt, N. & Boxtel, H. van (2002). *Indicatiestelling. Condities en instrumentarium.* Amsterdam: NDC/Boom.

Snijders, J.Th., Tellegen, P.J. & Laros, J.A. (1988). *Snijders-Oomen niet-verbale intelligentietest. SON-R 5,5-17. Verantwoording en handleiding*. Groningen: Wolters-Noordhoff.

Span, M.M. (2002). WAIS-III. De Stand van zaken. *De Psycholoog, 37*, 602-606.

Te Nijenhuis, J. & Evers, A. (2000). Is een wetenschappelijke benadering van testgebruik bij allochtonen onverantwoord? een reactie op Tellegen (2000). *De Psycholoog, 35*, 327-332.

Te Nijenhuis, J. & Flier, H. van der (2001). Group differences in mean intelligence for the Dutch and third world immigrants. *Journal of Biosocial Science, 33*, 469-475.

Tellegen, P.J. (2000). Verantwoord testgebruik bij allochtonen. een reactie. *De Psycholoog, 35,* 231-235.

Tellegen, P.J. (2002b). De kwaliteit van de normen van de WAIS-III. *De Psycholoog, 37*, 463-465.

Tellegen, P.J. (2002c). Afname van de WAIS-III of WISC-III. Verantwoord en verstandig? *De Psycholoog, 37*, 677-679.

Tellegen, P.J. (2002d). De WISC-III NL. Een illusie armer. *De Psycholoog, 37*, 607-610.

Tellegen, P.J., Winkel, M., Wijnberg-Williams, B. & Laros, J.A. (1998). *Snijders-Oomen Niet-verbale Intelligentietest. SON-R 2,5-7. Handleiding en Verantwoording*. Lisse: Swets & Zeitlinger.

Uterwijk, J. (2000). *WAIS-III Nederlandstalige bewerking. Technische handleidin*g. Lisse: Swets & Zeitlinger.

Zeeuw, J. de, Dekker, R. & Resing, W.C.M. (2004). *Algemene Psychodiagnostiek I. Testmethoden. Geheel herziene druk*. Leiden: PITS.