De WISC-III anno 2006:

een voorstel tot eenduidige en hiërarchische analyse, interpretatie en rapportage

De WISC-III is de meest gebruikte intelligentietest bij kinderen en jongeren. Na een kritische bespreking van de test en de belangrijkste ontwikkelingen van de laatste jaren volgt een stapsgewijze beschrijving van een analyse- en interpretatiemethode die verantwoord, eenduidig en statistisch onderbouwd is. Vooralsnog heeft deze methode bij het grote vakpubliek weinig voet aan de grond gekregen in rapportages en publicaties over de WISC-III. Naast een methode die houvast biedt, wordt de lezer ook voorzien van bruikbare digitale bestanden waarmee volgens deze methode gescoord en gerapporteerd kan worden.

Yaron Kaldenbach

INLEIDING

Binnen het kinder- en jeugdveld geniet de WISC *(Wechsler Intelligence Scale for Children)* al jarenlang grote populariteit onder psychologen, orthopedagogen en andere professionals die zich bezighouden met het meten van het intelligentieniveau bij kinderen en jongeren. Eens in de zoveel jaar verschijnt er een nieuwe en geactualiseerde versie van de test, voorzien van recente normgegevens. In 2002 werd in Nederland de WISC-III -officieel WISC-III^{NL}- (Kort e.a., 2002; 2005) geïntroduceerd, de opvolger van de in 1986 uitgebrachte WISC-R (Vander Steene, 1986). De WISC-III is een intelligentietest voor 6- t/m 16-jarigen en wordt in tientallen landen gebruikt. De test bestaat uit 13 subtests, aan de hand waarvan een verbale en performale schaalscore berekend kunnen worden. Analyse volgens een 'driefactorenstructuur' is eveneens mogelijk (Verbaal Begrip, Perceptuele Organisatie en Verwerkingssnelheid).

De WISC-III raakte echter in opspraak, onder meer vanwege de betwijfelde representativiteit van de normgegevens. Door aanhoudende kritiek vanuit het veld volgden meerdere aanpassingen en herzieningen, waarvan de laatste dateert uit april 2005. Van de WISC-III zijn sinds de introductie drie verschillende versies van de normgegevens in omloop gebracht (handleiding 2002, erratum oktober 2003 en de herziene handleiding 2005). Alleen gebruik van de meest recente versie is nu verantwoord.

Najaar 2005 werd de WISC-III door de COTAN (*Commissie Testaangelegenheden Nederland* van het *Nederlands Instituut van Psychologen*, het NIP) positief herbeoordeeld. Nu we 'met een gerust hart' de WISC-III kunnen gebruiken, is het tijd om iets anders aan de orde te stellen wat eigenlijk al veel langer speelt. Orthopedagogen en psychologen (gemakshalve zal aan hen gerefereerd worden als 'psychodiagnosten', waartoe we ook psychodiagnostisch werkenden kunnen rekenen) rapporteren uiterst verschillend over intelligentietests. Logischerwijs wat betreft vorm en stijl, maar ook op inhoudelijk vlak, zoals onder meer op het gebied van de analyse, de naamgeving en interpretatie van scores, en de conclusies en adviezen die men op deze tests baseert. Dit leidt ertoe dat de uitkomsten van een onderzoek in de huidige praktijk niet alleen afhangen van het kind, maar ook van degene die het onderzoek uitvoert. Het gebrek aan eenduidigheid verdient geen schoonheidsprijs en is bovendien kwalijk, omdat het soms leidt tot onverantwoorde individuele uitspraken.

In dit artikel zal, nadat eerst een aantal algemene zaken rondom de WISC-III de revue hebben gepasseerd, een concreet voorstel worden gedaan om op een systematische en verantwoorde manier stap voor stap de testresultaten te analyseren

Over de auteur

Drs. Y. Kaldenbach, gz-psycholoog en kinder- en jeugdpsycholoog NIP, Coördinator Psychodiagnostiek bij afdeling Jeugd van Altrecht GGZ te Utrecht. Naast diagnostieksupervisor is hij lid van de commissie Basisaantekening Psychodiagnostiek NIP en verzorgt hij WISC-III cursussen. E-mail: ykaldenbach@hetnet.nl.

Trefwoorden

WISC-III, intelligentietest
IQ
analysemethode
profielinterpretatie
rapportage

en te interpreteren. De lezer wordt van harte uitgenodigd om kritisch te reflecteren op zijn huidige manier van werken en over de beschreven getrapte methode binnen zijn instelling met collega's een discussie te initiëren.

COTAN

In 2003 gaf de COTAN (NIP, 2004) haar eerste oordeel over de WISC-III en dat was niet best (figuur 1): onvoldoendes voor de normen, betrouwbaarheid (de mate waarin een meting afhankelijk is van toevalsvariabelen), begripsvaliditeit (de mate waarin ook daadwerkelijk wordt gemeten wat men beoogt te meten, namelijk het theoretische begrip 'intelligentie') en criteriumvaliditeit (in hoeverre de testscore een goede voorspeller is van niet-testgedrag – retrospectief, gelijktijdig of predictief – , zoals bijvoorbeeld een CITO-score) (Swanborn, 1993; Drenth & Sijtsma, 1990). Hoewel met voetnoten door de COTAN werd toegelicht wat de motivatie voor een onvoldoende was, hetgeen soms de beoordeling enigszins nuanceerde (normen waren niet representatief en/of de representativiteit was niet te beoordelen, test-hertest gegevens ontbraken en er was te weinig onderzoek naar validiteit gedaan), bleef de test in opspraak. Bij herhaling werd ook de media gehaald (voor een uitgebreide kritische beschrijving, zie www.testresearch.nl, de website van Peter Tellegen, auteur van onder meer de SON-R, een non-verbale intelligentietest). Een van de voornaamste bezwaren had betrekking op de oververtegenwoordiging van de groep hoogopgeleiden in de steekproef waardoor de test te streng was. Deze kritiek werd ter harte genomen en in oktober 2003 verschenen de eerste herziene normen (NDC, 2003). Omdat de handleiding verder op een aantal punten ongemoeid was gelaten, werd met het verschijnen van deze nieuwe normen in één klap een aantal tabellen uit de handleiding onbruikbaar, bijvoorbeeld tabellen over het berekenen van significante verschillen tussen IQ-scores (intelligentie quotiënt). Markant was bovendien dat er een aantal verschillende versies circuleerde van de handleiding en testboekjes. De versies verschilden wat betreft de mate waarin errata/correcties waren doorgevoerd. In de update van 2003 bleef bovendien een aantal eerder geuite bezwaren bestaan, reden om in april 2005 na de aanhoudende kritiek een nieuwe handleiding met uitgebreidere en wederom herziene (norm)gegevens uit te brengen, en deze nogmaals aan de COTAN ter beoordeling voor te leggen. Ditmaal was de COTAN een stuk positiever. Van de vier onvoldoendes bleef er slechts één overeind (NIP, 2005), en over deze valt te steggelen. Het zou correcter zijn geweest wanneer de COTAN de criteriumvaliditeit niet als 'onvoldoende' had beoordeeld, maar vermeld zou hebben dat er te weinig onderzoek naar is gedaan om de criteriumvaliditeit goed te kunnen beoordelen. Overigens wordt op dit moment onderzoek gedaan naar de criteriumvaliditeit van de WISC-III, waarschijnlijk met het doel deze naderhand aan de COTAN voor te leggen om ook de laatste onvoldoende op te heffen.

Criterium	2003	2005
Uitgangspunten bij de testconstructie	voldoende	goed
Kwaliteit van het testmateriaal	goed	goed
Kwaliteit van de handleiding	voldoende	goed
Normen	onvoldoende	voldoende
Betrouwbaarheid	onvoldoende	voldoende
Begripsvaliditeit	onvoldoende	voldoende
Criteriumvaliditeit	onvoldoende	onvoldoende

Figuur 1. COTAN-beoordelingen van de WISC-III^{NL}

Belangrijk is verder dat de COTAN een halt toeroept aan de subtestanalyse zoals deze veelal in de diagnostische praktijk plaatsvindt. Ook de makers van de WISC-III onderschrijven dit (Kort e.a., 2005). Op basis van subtests worden regelmatig verregaande uitspraken gedaan over zwakke en sterke kanten van kinderen, soms zelfs vergezeld van uitspraken over cerebraal functioneren met een dringende suggestie voor neurologisch onderzoek. De COTAN is hierover duidelijk: analyseren op subtestniveau wordt afgeraden en kan hooguit met veel voorzichtigheid enkele hypotheses genereren. Dit onder meer vanwege de betrouwbaarheden van de afzonderlijke subtests. In dit kader wordt de subtest Doolhoven er speciaal uitgelicht.

Ook de factor Verwerkingssnelheid wordt genoemd als factor om extra prudent mee om te gaan wat betreft interpretatie. Deze factor heeft een relatief lage test-hertest betrouwbaarheid (de mate waarin resultaten over twee verschillende afnamen in de tijd gelijk zijn binnen en tussen personen) en is bij jongere leeftijdsgroepen niet goed op te sporen (NIP, 2005). In aanvulling hierop kan worden gesteld dat de onderzoeker zich dient te realiseren dat deze factor slechts uit twee subtests bestaat. Hierdoor wordt de factor pas bij grotere verschillen tussen de twee subtests intern inconsistent (6 punten of meer; er is een aantal redenen om binnen de methode van dit artikel niet te kiezen voor de SU-SV significantiegrenzen uit de handleiding, maar het voert te ver deze

hier nu uitvoerig toe te lichten). Bovendien is de naam van de factor misleidend. Het betreft namelijk *visuele* en *kortdurende* snelheid van informatieverwerking *onder tijdsdruk*. In de praktijk laten kinderen met een gemiddelde Verwerkingssnelheid op een Bourdon-Vos (volgehouden visuele aandachtstest) regelmatig een traag en instabiel werktempo zien.

Verder kwam ook uit onderzoek naar voren dat het zogenaamde 'Flynn-effect', de natuurlijke stijging van prestaties op intelligentietests met gemiddeld 3-5 punten per decennium (Flynn, 1994), bij de WISC-III niet werd teruggevonden zoals op basis van de literatuur was voorspeld (Kort e.a., 2005). Bij sommige groepen trad het effect niet op, bij kinderen met een lichte verstandelijke beperking was er wel sprake van het Flynn-effect maar weer minder ten aanzien van het performaal IQ.

Hoewel de Wechslertests worden beschouwd als intelligentietests met een sterk actief verbaal accent, blijken zowel bij de autochtonen als bij de allochtonen geen grote verschillen tussen het verbaal (VIQ) en performaal IQ (PIQ). Op alle subtests (behalve Substitutie) worden door de allochtoon-etnische groep gemiddeld lagere scores behaald, maar hierin zijn geen grote schommelingen waarneembaar (NIP, 2005). Er wordt bij deze doelgroep geen specifieke verbale uitval gevonden, wat vaak wel wordt verwacht en reden kan zijn te kiezen voor een niet-verbale test.

Ten slotte wordt geadviseerd om bij intelligentietests altijd gebruik te maken van de betrouwbaarheidsintervallen. Een enkel getal suggereert een mate van exactheid die niet waar te maken is en gaat voorbij aan omstandigheden die maken dat iemand wat hoger of lager kan scoren (de meetfout). Daarom is het ook zo onterecht dat instellingen, indicatieorganen en scholen absolute *cut-off* IQ's hanteren om iemand al dan niet te accepteren: een totaal IQ van 86 betekent dat een kind welkom is, als het totaal IQ 84 is, wordt het doorverwezen omdat het wegens het cognitieve niveau 'ineens' onvoldoende in staat zou zijn te profiteren van het aanbod.

Binnen de statistiek is het 95%-betrouwbaarheidsinterval het meest gangbaar (Swanborn, 1993). Daarom wordt geadviseerd bij de WISC-III het 90%-betrouwbaarheidsinterval uit de handleiding bij voorkeur niet te gebruiken. In vijf van de honderd gevallen zal er door toevalsfactoren sprake zijn van een interval dat niet overlapt met iemands 'echte' capaciteiten, een onder- of overschatting door de test dus. Wanneer men het 90%-interval uit de handleiding hanteert, wordt het interval smaller en dus overzichtelijker, en voorziet weer meer in de behoefte van een onderzoeker om een exacte uitspraak te kunnen doen. Het oogt nauwkeuriger, maar let wel: er is een kans van 10% dat de onderzoeker ernaast zit met dat smalle interval.

Ten slotte is het in algemene zin belangrijk te noemen dat de betrouwbaarheid (en validiteit) van de meting optimaal is wanneer de psychodiagnost zich houdt aan de richtlijnen van de handleiding. Het lijkt overbodig dit expliciet te benoemen, maar de praktijk leert dat veel onderzoekers naarmate ze 'feeling' met tests ontwikkelen, zichzelf steeds meer vrijheid permitteren met betrekking tot de testinstructies, subtestvolgorde, afbreekregels en itemscoring.

DE WISC-R: ENKELE REIS NAAR HET MUSEUM

De 'oude WISC' wordt nog steeds gebruikt in Nederland, meestal door individuele psychodiagnosten, incidenteel zelfs nog instellingsbreed. De COTAN noemt normen na 15 jaar verouderd en na 20 jaar onbruikbaar. De normgegevens van de WISC-R werden begin jaren tachtig verzameld en zijn nu dus onbruikbaar. Op alle overige categorieën is het oordeel van de COTAN over de WISC-III gelijk aan dat over de WISC-R, waarbij moet worden opgemerkt dat de COTAN-beoordeling van de WISC-R uit 1992 stamt en inmiddels ook aan geldigheid heeft verloren. De oude WISC overschat het niveau van cognitief functioneren. In de praktijk worden allerlei creatieve maar onverantwoorde oplossingen in verslagen teruggevonden om hiervoor te compenseren, zoals een automatische 'Flynn-effect correctie', toegepast door een aantal punten van de IQ's af te trekken. Eerder werd geconcludeerd dat het Flynn-effect rondom de WISC-III een stuk genuanceerder ligt en geen algemene correctie rechtvaardigt. De Sector Jeugd van het NIP gaf in januari 2005 al een flyer uit met veelgestelde vragen over intelligentietest (NIP, 2005), waarin gebruik van de WISC-R werd afgeraden. Met de herziene normen en de COTAN-herbeoordeling van 2005 is gebruik van de WISC-R niet langer verantwoord en bij een klacht bovendien moeilijk verdedigbaar. Het is goed elkaar hier als collega's op te wijzen en te discussiëren over het spanningsveld tussen de individuele beroepsverantwoordelijkheid enerzijds en de voorgeschreven tests door indicatieorganen anderzijds. Het gebruiken van recente normen verdient in principe altijd de voorkeur.

Bij de WISC-R werd regelmatig gebruikgemaakt van 'Kaufman en Bannatyne'-analyses (Kaufman, 1994, 1975; Bannatyne, 1974). Deze methoden zijn exclusief van toepassing op de WISC-R en de Kaufman-analyse (hier bedoeld als een analyse met vier factoren, waaronder de factor 'Vrijheid van Afleidbaarheid') is daarnaast toepasbaar bij sommige buitenlandse versies van de WISC-III (Kort e.a., 2005; Georgas e.a., 2003). Beide methoden mogen niet in relatie tot de *Nederlandse* versie van de WISC-III gehanteerd worden, ondanks incidentele ondersteuning voor de 'vierde factor' binnen kleinschalig Nederlands onderzoek (Oosterbaan e.a., 2006). Grotere inter-


Foto: Aleid Denier van der Gon

nationale onderzoeken (In: Georgas e.a., 2003) laten zien dat de factor 'Freedom from Distractibility' vaak niet uit factoranalyses naar voren komt en binnen het geheel van factoren vaak een relatief zwak onderdeel blijft. Ook wordt in het veld nog een aantal andere analysemethoden aangetroffen waarvan de herkomst, status en theoretische fundering soms onduidelijk is. In Nederlandstalige literatuur zijn voor zover de auteur van dit artikel bekend tot op heden geen beschrijvingen verschenen van analysemethoden voor de Nederlandse WISC-III.

DE METHODE: HIËRARCHISCHE ANALYSE EN INTERPRETATIE

In het volgende wordt een analysemethode beschreven voor de WISC-III. Elementen uit deze methode werden eerder onder meer beschreven in internationale literatuur (Kaufman & Lichtenberger, 2000) en binnen Nederland voor de WISC-R (Geelhoed, 1996). Hoewel er instellingen in Nederland zijn die (delen van) deze methode gebruiken, geniet de methode onvoldoende bekendheid. Omdat de statistische eigenschappen (gemiddelde, standaardafwijking en range) van de normscores bij de WISC-III ongewijzigd zijn ten aanzien van de WISC-R, is hiermee een verantwoorde basis gelegd om de methode, met lichte aanpassingen, ook op de WISC-III (en tevens de WAIS-III, de equivalent voor cliënten van 16-85 jaar; Uterwijk, 2000) toe te passen. Op www.kindenadolescent.nl zijn bestanden te downloaden, waaronder een 'Scorehulp', die snel en foutloos volgens de beschreven analysemethode berekeningen uitvoert. In figuur 2 is de methode verkort en stapsgewijs weergegeven, op de website kunt u een uitgebreide beslisboom downloaden.

Statistiek als basis: m = 10 en sd = 3

In de WISC-III worden ruwe scores omgerekend naar (gestandaardiseerde) normscores. Deze normscores hebben een range van 1 t/m 19, met een gemiddelde (m) van 10 en een standaardafwijking (de gemiddelde afwijking tot het gemiddelde) van 3. De hoogste en laagste score van de range worden bepaald door 10 ± driemaal de standaardafwijking (sd). In de statistiek wordt bij een normaal verdeelde variabele meestal uitgegaan van het principe dat afwijkingen van minder dan één standaarddeviatie van het gemiddelde als gemiddeld geclassificeerd worden (Nijdam & Van Buuren, 1994). Scores die meer dan twee standaardafwijkingen van het gemiddelde verschillen, worden uitzonderlijk genoemd (Groen e.a., 1991). Toegepast op de normscores betekent dit dat een score van

boven de 7 en *onder* de 13 binnen de range van het gemiddelde valt, in de praktijk geldt dan dat normscores tussen de 8 en 12 gemiddeld zijn (bij omrekening overeenkomstig met het interval voor gemiddelde IQ-scores 90-110). Wanneer deze systematiek verder wordt gevolgd, kan een classificatie van subtestscores worden gehanteerd zoals weergegeven in figuur 3. In de praktijk wordt hiervan vaak afgeweken, men vindt een normscore van 8 dan bijvoorbeeld al laaggemiddeld of benedengemiddeld. Er wordt dan voorbijgegaan aan een stukje statistiek dat op deze normscores van toepassing is.

Stap 1: een harmonisch profiel op schaalniveau?

Nadat de WISC-III gescoord is en alle normscores en IQ-waarden berekend zijn, wordt allereerst gekeken of de verbale en performale schaal significant van elkaar verschillen. De tabel in de handleiding geeft weer hoe groot het verschil bij elke leeftijd minimaal moet zijn om te mogen spreken van een significant verschil (de 'Scorehulp' berekent dit automatisch).

Een niet-significant verschil tussen het VIQ en PIQ is toevalstreffer

Ook hier wordt bij voorkeur $p < 0.05$ als richtlijn gehanteerd. Algemene vuistregels voor discrepanties van bijvoorbeeld 12 of 15 punten dienen niet gebruikt te worden, omdat deze aanzienlijk kunnen afwijken van de exacte grenswaarden per leeftijd zoals in de handleiding vermeld (deze variëren bij de schalen en factoren van 12 tot 18 punten).

Statistisch significante verschillen komen regelmatig voor en zijn niet per definitie zorgelijk (klinisch relevant) of reden voor vervolgonderzoek. Ieder mens heeft sterke en minder sterke kanten in zijn profiel en disharmonische profielen komen

Bereken alle normscores en IQ's

1. Bepaal of er een harmonisch profiel op schaalniveau is (verschillen de schalen onderling significant?)
2. Bepaal of de schalen intern consistent zijn
3. Bepaal of er op factorniveau sprake is van een harmonisch profiel (vergelijking *tussen* de factoren) en of de factoren intern consistent zijn (analyse *binnen* de factoren)
4. Analyse op subtestniveau (hypothesevormend)
5. Analyse op itemniveau (facultatief en hypothesevormend)

Figuur 2. Hiërarchische analyse en interpretatie in vogelvlucht

in normale populaties veel voor. Bij niet-significante verschillen wordt afgeraden uitspraken te doen over verschillen in verbale en performale capaciteiten. Het verschil is niet significant en moet veiligheidshalve als toevalstreffer worden bezien.

Als er *wel* een significant verschil tussen het VIQ en PIQ is,

> 15	zeer goed
13-15	goed
8-12	gemiddeld
5-7	zwak
< 5	zeer zwak

Figuur 3. Classificatie van WISC-III normscores

betekent dit dat de gemeenschappelijke noemer van het totale IQ (TIQ) minder betekenis heeft en niet geïnterpreteerd kan worden omdat deze de lading niet goed dekt. Het TIQ is dan een weinig zeggend getal tussen het VIQ en PIQ in geworden (overigens geen rekenkundig gemiddelde van het VIQ en PIQ) dat inhoudelijk te weinig informatief is om bijvoorbeeld schoolkeuzes op te baseren. Dit betekent echter niet dat het TIQ niet hoeft te worden vermeld. Het is beter een waarde binnen de context te plaatsen dan deze weg te laten (voor mogelijke toekomstige onderzoekers is het prettig om over alle gegevens te kunnen beschikken, wat vergelijking mogelijk maakt). Een afwijkende subtest wordt om deze reden ook niet weggelaten.

Stap 2: zijn de schalen intern consistent?

De term 'harmonisch' wordt dus gebruikt voor de beschrijving van verschillen *tussen* de schalen (en verderop ook tussen de factoren onderling). Voor de beschrijving van verschillen *binnen* een schaal (of factor) wordt de term 'interne consistentie' gehanteerd.

Na stap 1 wordt vervolgens voor beide schalen het schaalgemiddelde en per subtest de afwijking hiervan berekend. Uitgaande van m = 10 en sd = 3 kan gesteld worden dat een schaal intern consistent is wanneer geen van de subtests binnen die schaal 3 punten of meer afwijkt van het schaalgemiddelde. Zodra er wel één of meer subtests zijn die 3 punten of meer afwijken van het schaalgemiddelde, spreken we van een intern inconsistente schaal. Intern consistent houdt in dat de schaal de inhoud goed dekt. Er is sprake van een gemeenschappelijke noemer die uit betrekkelijk gelijkwaardige/samenhangende onderdelen is opgebouwd. Wanneer er op schaalniveau sprake is van interne *in*consistentie, is dit een reden om te analyseren op factorniveau en mag dus ook de algemene schaalomschrijving niet meer bij deze cliënt gebruikt worden. Zijn beide schalen intern consistent, dan kunnen deze geïnterpreteerd worden. Het is goed om evenwel stap 3 uit te voeren (analyse

op factorniveau). Indien namelijk ook op factorniveau een intern consistent beeld blijft, dan kan het informatief zijn aan de hand van de factorbeschrijvingen een gedifferentieerder beeld van het kind te geven dan bij de schaalbeschrijvingen (zie box 1 voor een beschrijving van de meetpretenties van de schalen, factoren en subtests). Internationaal is de trend om op factorniveau te analyseren en in de enkele jaren geleden in Amerika verschenen WISC-IV worden de verbale en performale schaal niet eens meer teruggevonden en is er naast het TIQ sprake van vier zogenaamde 'Index Scores' (Verbaal Begrip, Perceptueel Redeneren, Werkgeheugen en Verwerkingssnelheid).

Een voordeel van het gebruik van statistische onderbouwing is dat het een einde maakt aan het 'gevoelsmatig' bepalen of een profiel wel of niet harmonisch is, waarbij iedereen zijn eigen opvattingen heeft over wat hij wel of niet een groot verschil vindt.

Stap 3: analyse op factorniveau

Op factorniveau worden de stappen 1 en 2 herhaald, die eerder bij de schalen werden toegepast. Er wordt bepaald of er op factorniveau sprake is van een harmonisch profiel (zijn er significante verschillen tussen factoren?) en berekend of de factoren intern consistent zijn (zijn er subtests die minimaal 3 punten afwijken van het factorgemiddelde?). Vaak, maar zeker niet altijd, wordt gevonden dat op factorniveau vervolgens drie intern consistente factoren verschijnen nadat er op schaalniveau eerder sprake was interne *in*consistentie (zie box 2). De performale interne inconsistentie wordt bijvoorbeeld regelmatig verklaard door een 'afwijkende' subtestscore op Substitutie, onderdeel van de factor Verwerkingssnelheid. Bij intern consistente schalen en factoren wordt het afgeraden om te analyseren op subtestniveau. Aangezien de subtest niet noemenswaardig afwijkt van de overige subtests binnen die schaal of factor, is er geen reden om de schaal- of factorbeschrijving te verlaten, te meer omdat op dit niveau ook wat stelliger dan op subtestniveau geïnterpreteerd mag worden. De factor Verwerkingssnelheid is hierop een uitzondering (zie eerder). Wanneer analyse op factorniveau plaatsvindt, verdient het de voorkeur om in het rapport ook de subtests onder drie kopjes van de factoren weer te geven en niet onder de twee schalen. Eventuele inconsistenties op factorniveau kunnen met een '+' of '–' achter de normscore worden weergeven, respectievelijk voor een significante afwijking naar boven en naar beneden ten opzichte van het factorgemiddelde. Hetzelfde geldt voor schaalniveau, maar in de praktijk wordt dan dus vervolgd met analyse en beschrijving op factorniveau.

Stap 4: analyse op subtestniveau (hypothesevormend)

Eerder werd gezegd dat subtestanalyse overbodig is wanneer blijkt dat de schalen of factoren de lading goed dekken (lees: intern consistent zijn). In de 'klinische praktijk' zien we

Box 1. Meetpretenties van de schalen, factoren en subtests

Schalen

Verbale schaal:

Taken uit deze schaal doen een beroep op de auditief-verbale informatieverwerking en vragen een verbale respons. Taalvaardigheid, verbaal begrip en verbale (opgedane) kennis spelen een belangrijke rol, evenals het auditief geheugen.

Performale schaal:

Taken uit deze schaal doen een beroep op visueel-ruimtelijke informatieverwerking en vragen meestal een motorische respons. Visuele analyse en synthese, visuo-motoriek, snelheid van werken, visueel (associatief) en non-verbaal redeneren met betrekking tot sociale situaties zijn hier van belang.

Factoren

Factor Verbaal Begrip:

Deze subtests doen een beroep op inzicht in door middel van taal gepresenteerde problemen (definiëren van betekenis, verwoorden van kennis en verbaal abstract redeneren).

Factor Perceptuele Organisatie:

Deze subtests doen een beroep op onmiddellijke probleemoplossingsvaardigheden bij visueel-ruimtelijke problemen, visuo-motoriek en non-verbaal redeneren met betrekking tot sociale situaties.

Factor Verwerkingssnelheid:

Deze subtests doen een beroep op snelheid van visuele informatieverwerking, visueel associatief geheugen en visuele matching.

Subtests

Informatie (IN): algemene kennis

Overeenkomsten (OV): verbaal abstract redeneren

Rekenen (RE): rekenvaardigheid (auditief kortetermijngeheugen)

Woordkennis (WO): woordkennis

Begrijpen (BG): inzicht in dagelijkse (sociale) situaties

Onvolledige Tekeningen (OT): visuele detailwaarneming

Substitutie (SU): visueel associatief geheugen

Plaatjes Ordenen (PO): non-verbaal redeneren met betrekking tot sociale situaties ('sociaal sequentiëren')

Blokpatronen (BP): visuo-motoriek, visuele analyse en synthese, patroonwaarneming

Figuur Leggen (FL): visuo-motoriek, visualisatie, patroonherkenning

Symbool Vergelijken (SV): snelheid van visuele informatieverwerking, visuele matching

Cijferreeksen (CR): auditief sequentieel geheugen

Doolhoven (DH): visuele oriëntatie, planning

Box 2. Fragment uit een psychologisch rapport volgens de hiërarchische analysemethode: Casus Milan.

WISC-III^{NL} (herziene normen 2005)
Meetpretentie: intelligentie

Schalen:	
Verbaal IQ (VIQ):	121 (112-127)
Performaal IQ (PIQ):	99 (89-109)
Totaal IQ (TIQ):	112 (104-118)

Factoren:	
Verbaal Begrip (VB):	128 (117-134)
Perceptuele Organisatie (PO):	106 (95-116)
Verwerkingssnelheid (VS):	83 (75-96)

VB-factor	
Informatie (IN):	14
Overeenkomsten (OV):	17
Woordkennis (WO):	14
Begrijpen (BG):	13

PO-factor	
Onvolledige Tekeningen (OT):	11
Plaatjes Ordenen (PO):	12
Blokpatronen (BP):	10
Figuur Leggen (FL):	11

VS -factor	
Substitutie (SU):	6
Symbolen Vergelijken (SV):	7

Aanvullende subtests	
Rekenen (RE):	8
Cijferreeksen (CR):	11
Doolhoven (DH):	13

Milan heeft een totaal IQ van ongeveer 112 (95% betrouwbaarheidsinterval ligt tussen de 104-118) en presteert hiermee op bovengemiddeld intelligentieniveau. Er is sprake van een disharmonisch profiel op schaal- en factorniveau. Het verbaal IQ (121, begaafd niveau) is significant hoger dan het performaal IQ (99, gemiddeld niveau). Ook de factoren verschillen onderling allemaal significant. Zowel de performale als de verbale schaal zijn intern inconsistent, reden waarom analyse op factorniveau plaatsvindt.

De factor *Verbaal Begrip* (128, begaafd niveau) is intern consistent. Milans inzicht in door middel van taal gepresenteerde problemen (definiëren van betekenis, verwoorden van kennis en verbaal abstract redeneren) is beter ontwikkeld dan bij leeftijdgenoten.

De factor *Perceptuele Organisatie* (106, gemiddeld niveau) is eveneens intern consistent. Milans onmiddellijke probleemoplossingsvaardigheden bij visueel-ruimtelijke problemen, de visuo-motoriek en het non-verbaal redeneren met betrekking tot sociale situaties zijn leeftijdsadequaat ontwikkeld.

De factor *Verwerkingssnelheid* (83, benedengemiddeld niveau) is intern consistent. Milans snelheid van visuele informatieverwerking, visueel associatief geheugen en visuele matching lijken minder ontwikkeld dan bij leeftijdsgenoten.

De *aanvullende subtests*: Milans auditief kortetermijngeheugen en rekenvaardigheid lijken van gemiddeld niveau (RE), evenals zijn auditief sequentieel geheugen (CR). Zijn visuele oriëntatie en planning lijken goed ontwikkeld (DH).

echter dat de verleiding om aan allerlei subtiele verschillen betekenis te verlenen toch hardnekkig is.

Maar wat nu als het profiel inderdaad heel grillig is met uitschieters naar boven en beneden, waardoor een schaal- of factor-IQ weinig betekenis heeft vanwege interne inconsistentie? In dat geval mag de algemene schaal- of factorbeschrijving niet gebruikt worden om het niveau van de vaardigheden van het kind te beschrijven. Hoewel de COTAN analyse op subtestniveau afraadt (NIP, 2005), is het in zulke gevallen zinvol nader naar de subtestverschillen te kijken. Analyse op subtestniveau is echter 'tricky' en moet op een hypothesevormende manier gebeuren, met meer dan een enkele slag om de arm. Ook in de formulering dient dit tentatieve karakter duidelijk naar voren te komen. Bij een normscore van 6 op Onvolledige Tekeningen kan bijvoorbeeld vermeld worden dat iemands visuele detailwaarneming zwak ontwikkeld *lijkt*. Significante verschillen (zie handleiding voor grenswaarden) tussen bijvoorbeeld Blokpatronen en Figuur Leggen kunnen hypotheses genereren over verschillen in visuoconstructieve vaardigheden met betrekking tot abstracte en betekenisloze visueel-ruimtelijke informatie (BP) in vergelijking met concrete en betekenisvolle visueel-ruimtelijke informatie (FL). In het advies kan later in het rapport zo nodig verdiepingsonderzoek geadviseerd worden, mede bepaald door de klinische relevantie van het toetsen van deze richtinggevende hypothese (bijvoorbeeld voor de schoolsituatie of om een beter behandelaanbod te kunnen doen).

Stap 5: analyse op itemniveau (facultatief en hypothesevormend)

Het kan zinvol zijn ook op itemniveau te kijken hoe subtestscores zijn opgebouwd. Twee kinderen kunnen exact dezelfde subtestscore halen terwijl het ene kind veel antwoorden geeft die een 0 of een 2 als score krijgen (het kind is goed in staat wat het weet onder woorden te brengen), en het andere kind meer 1-punts antwoorden geeft (het kind weet bij meer vragen wel enige informatie te geven maar doet dit op een wat vage of te concrete manier). De betekenis van dezelfde subtestscore kan dan verschillen. Veel meer dan hierover iets in de observaties vermelden en eventueel een hypothese genereren, is echter niet geoorloofd.

Naamgeving van scores

Door de recente ontwikkelingen van de laatste jaren in 'IQ-test land' moeten we ons 'gevoel voor IQ's' drastisch herzien. Niet langer kan gezegd worden dat een IQ van 110-120 een

havoniveau weergeeft en de gemiddelde succesvolle vwo'er een 120+-IQ heeft. Er moet per test gekeken worden wat een IQ betekent; een WISC-III IQ is geen SON-R IQ. In de handleiding moet worden opgezocht welke IQ's gemiddeld zijn bij de opleidingsniveaus en welke standaarddeviaties daarbij horen. Op de WISC-III haalt de gemiddelde havoleerling bijvoorbeeld 'slechts' een TIQ van 106,9 (in de eerste normen uit 2002 was dit 103,8), een IQ waarmee hij voorheen nog naar de mavo zou zijn gestuurd (NIP, 2005/2002).

Daarnaast is het erg verwarrend voor collega's, opdrachtgevers en cliënten om in rapporten te zien hoe gegoocheld wordt met getallen en hun betekenis. Wat de een 'gemiddeld' noemt, vindt een ander alweer 'laaggemiddeld' of 'benedengemiddeld'. Volgens de indeling van Geelhoed (1996) zou een IQ van 89 bijvoorbeeld 'zwakbegaafd' genoemd moeten worden, terwijl anderen hieraan meestal refereren als 'benedengemiddeld'. Het hangt dus blijkbaar van de onderzoeker af hoe een IQ genoemd wordt. Het is belangrijk dat er één taal komt waarin over de IQ-scores gesproken wordt. Ook daarom is het zo belangrijk alle getallen in het rapport te vermelden, zodat het rapport transparant en toetsbaar is. Een indeling van IQ-scores wordt in figuur 4 weergegeven. Deze indeling is door Resing en Blok (2002), beiden lid van de COTAN, beschreven in *De Psycholoog* en wordt onder meer door de Sector Jeugd van het NIP gesteund. Daaronder staat de indeling volgens de DSM-IV-TR. Nadeel van deze laatste indeling is dat de range alleen de lagere IQ's van betekenis voorziet en hierbij niet altijd even flatterende termen gebruikt.

Een kritische lezer zal opmerken dat de range van een gemiddeld IQ (figuur 4) van 90 tot 110 loopt, terwijl op basis van m = 100 en sd = 15 een range van 85-115 statistisch correct zou zijn. De reden voor deze smallere en veelgebruikte range, heeft te maken met het risico van leerproblemen ('externe validiteit'). Een IQ van 86 zou statistisch wel 'gemiddeld' zijn, maar in de praktijk kunnen samengaan met leerproblemen (Geelhoed & Güldner, 2002). In de naamgeving zijn landelijk de marges 'verengd' om beter te kunnen differentiëren voor de praktijk. Statistisch zijn de IQ's 86 en 114 beide gemiddeld, maar de eerste leidt tot een vmbo-advies terwijl de tweede mogelijkheden binnen het vwo laat zien.

Resing & Blok (2002)

> 130	zeer begaafd
121-130	begaafd
111-120	bovengemiddeld
90-110	gemiddeld
80-89	benedengemiddeld
70-79	laag begaafd / moeilijk lerend
50-69	lichte verstandelijke beperking / licht zwakzinnig

DSM-IV-TR (2001)

71-84	zwakbegaafdheid
50-55 tot ±70	lichte zwakzinnigheid
35-40 tot 50-55	matige zwakzinnigheid
20-25 tot 35-40	ernstige zwakzinnigheid
< 20-25	diepe zwakzinnigheid

Figuur 4. Naamgeving van IQ-scores

Adviezen voor de rapportage

Om te zorgen dat op een verantwoorde manier wordt gerapporteerd, is het belangrijk kennis te hebben van de inhoud van de Beroepscode (NIP, 1998) en de Algemene Standaard testgebruik (AST) (NIP, 2004). In box 2 is een deel van een rapport volgens de hier beschreven methode weergegeven. Daarnaast zijn enkele specifieke zaken van belang om de kwaliteit van een WISC-III-rapport te verbeteren:

Vermeld altijd *alle* normscores en *alle* IQ's (het totaal IQ, de schaal IQ's en factor IQ's) met de gehanteerde betrouwbaarheidsintervallen (90% of 95%). Vermeld ook de gebruikte normgroep. Van veel tests zijn namelijk updates verschenen na de oorspronkelijke normen en een rapport kan beter op waarde geschat worden wanneer is aangegeven van welke normgegevens gebruik is gemaakt. Anders is bij hertesten niet na te gaan of een afwijkende uitkomst een gevolg is van prestatieveranderingen bij het kind, discontinuïteit in normgegevens (de metingen zijn dan niet meer vergelijkbaar), of verschillen in gehanteerde betrouwbaarheidsintervallen. Alleen als intervallen van gelijke significantieniveaus (bijvoorbeeld 95%) elkaar niet overlappen, kan op dat significantieniveau gesteld worden dat er sprake is van significante verschuivingen tussen de metingen. Het is goed om niet alleen in intervallen te rapporteren (wel altijd de punt-IQ's vermelden), maar ook meer in intervallen te *denken*. Ook wij psychodiagnosten zijn dol op een getalletje dat een exactheid suggereert die niet strookt met de werkelijkheid. Het getal is slechts een schatting; het interval geeft de marge aan waarbinnen de kans groot is dat die schatting accuraat is.

Een vraag naar niveaubepaling komt vaak voort uit iets dat in het dagelijks leven van het kind niet goed gaat en vragen oproept. Praktische implicaties van de bevindingen voor thuis en op school zijn daarom een belangrijk onderdeel van het advies, uiteraard afhankelijk van de specifieke vraagstelling van het onderzoek. Wat betekenen de resultaten voor de manier waarop het kind bejegend wordt? Voor ouders en school kan het heel belangrijk zijn om te weten dat iemand door taalvaardigheid heel slim imponeert, maar tegelijkertijd een laag werktempo kan hebben en veel moeite met visueel-ruimtelijke taken. Een disharmonisch profiel ten gunste van de verbale vaardigheden brengt het risico van overschatting

en overvraging met zich mee. Onkunde op performaal vlak kan door grote verbale capaciteiten gemaskeerd worden en ten onrechte gelabeld worden als onwil. Andersom kan bij een zwak verbaal IQ met een veel sterker ontwikkelde performale kant gedacht worden aan het bieden van visuele ondersteuning, bij lagere niveaus zelfs met pictogrammen ('eerst het plaatje, dan het praatje'). Ook kan een kind geleerd worden om visuele en verbale informatie aan elkaar te koppelen, waarmee iemands visuele capaciteiten als kapstok kunnen dienen voor verdere ontwikkeling op verbaal terrein. Praktische adviezen (op welke manier kan men het kind stimuleren in zijn tekorten en zijn sterke kanten hierbij inzetten?) zorgen ervoor dat het onderzoek een gunstige invloed op de ontwikkeling van het kind heeft. Echter, een testsituatie kent ook beperkingen zoals het gegeven dat een kind slechts enkele uren gezien is. Informatie van ouders en school is daarom minstens zo belangrijk.

Bij de kinderen bij wie een gericht onderwijsadvies gevraagd wordt, is het goed de implicaties voor het onderwijsniveau en -type te geven van de bevindingen. Bekijk in dit geval niet alleen het losse getal of interval, maar neem hierin de factoren mee die invloed hebben gehad op de prestatie van de dag en andere factoren bekend uit de schoolinformatie en anamnese die invloed hebben op iemands kans van slagen op een bepaalde school. Uiteindelijk is het de deskundigheid van de psychodiagnost om op basis van alle gegevens tot een weloverwogen en expliciet gemotiveerd advies te komen passend binnen het totale diagnostisch beeld, ook al ligt het TIQ dan misschien wel enkele punten aan de 'verkeerde' kant van de grens die door organisaties voor indicatiestelling wordt gehanteerd.

Slotbeschouwing

Al met al kunnen we de WISC-III beschouwen als een instrument dat zich aan de hand van de beschreven analysemethode goed leent voor een niveaubepaling bij kinderen, met daarin een globale sterkte-zwakteanalyse, hypotheses voor mogelijk vervolgonderzoek en praktische individuele adviezen voor kind, ouders en school.

De WISC-III kan via de Swets Test Manager (Harcourt) digitaal gescoord worden. De STM is duur en meer gericht op de volwassen doelgroep. Alternatief hiervoor is het handmatig scoren, waarbij de eerder genoemde Scorehulp veel werk uit handen kan nemen. De onderzoeker moet nog wel zelf de IQ-getallen opzoeken behorend bij de berekende scores. Voordeel van de Scorehulp is dat deze past binnen de beschreven analysemethode en automatisch informatie geeft over harmonie en interne consistentie. •

LITERATUUR

- Bannatyne, A. (1974). A note on the reorganization on the WISC scale scores. *Journal of Learning Disabilities, 1*, 272-274.
- Drenth, P.J.D. & Sijtsma, K. (1990). *Testtheorie. Inleiding in de theorie van psychologische tests en zijn toepassingen.* Houten/Diegem: Bohn Stafleu Van Loghum.
- Flynn, J.R. (1994). IQ gains over time. In: R.J. Sternberg (Ed.), *Encyclopedia of Human Intelligence* (pp. 616-623). New York: Macmillan.
- Geelhoed, J.W. (1996). Intelligentie-onderzoek binnen de klinische cyclus. In: Pijnenburg, Van Rijswijk, Ruijsenaars & Veerman (red.), *Pedologisch Jaarboek 1996*. Delft: Eburon.
- Geelhoed, J.W. & Güldner, M. (2002). De classificatie van intelligentiescores: een reactie. *De Psycholoog, 37* (10), 522-524.
- Georgas, J., Weiss, L.G., Vijver, F.J.R. van de & Saklofske, D.H. (2003). *Culture and children's intelligence: Cross-cultural analysis of the WISC-III*. New York: Academic Press.
- Groen, W.E., Hakkert, A.J., Muthert, J.P., Stobbelaar, A.K., Vaessens, J.H.G. & Zwaneveld, G. (1991). *Moderne wiskunde bovenbouw 6V-A*. Groningen: Wolters Noordhof.
- Kaufman, A.S. (1975). Factor analysis on the WISC-R at 11 age levels between 6 and 16 years. *Journal of Consulting and Clinical Psychology, 43*, 145-157.
- Kaufman, A.S. (1994). *Intelligent testing with the WISC-III*. New York: Wiley.
- Kaufman, A.S. & Lichtenberger, E.O. (2000). *Essentials of WISC-III and WPPSI-R assessment*. New York: Wiley.
- Kort, W., Compaan, L., Bleichrodt, N., Resing, W.C.M., Schittekatte, M., Bosmans, M., Vermeir, G. & Verhaeghe, P. (2002). *WISC-III^{NL} Wechsler Intelligence Scale for Children. Derde Editie NL. Handleiding. David Wechsler*. Amsterdam: NIP Dienstencentrum.
- Kort, W., Schittekatte, M., Dekker, P.H., Verhaeghe, P., Compaan, E.L., Bosmans, M. & Vermeir, G. (2005). *WISC-III^{NL} Wechsler Intelligence Scale for Children. David Wechsler. Derde Editie NL. Handleiding en Verantwoording*. Amsterdam: Harcourt Test Publishers. Amsterdam: NIP Dienstencentrum.
- NDC (2003). *Errata en Normtabellen WISC-III^{NL} oktober 2003*. Amsterdam: NIP Dienstencentrum.
- Nederlands Instituut van Psychologen, Sector Jeugd (2005). *Veelgestelde vragen over intelligentietests.* Amsterdam: NIP. Te downloaden via www.jeugdpsycholoog.nl.
- Nederlands Instituut van Psychologen (1998). *Beroepsethiek voor Psychologen: Nieuwe Beroepscode 1998*. Amsterdam: NIP. Te downloaden via www.psynip.nl.
- Nederlands Instituut van Psychologen (NIP) (2004). *Algemene Standaard Testgebruik*. Amsterdam: NIP. Te downloaden via www.psynip.nl.
- Nederlands Instituut van Psychologen (NIP) (2005). *Documentatie van Tests en Testresearch in Nederland, aanvulling 2005/03*, pp.23-38. Amsterdam: Boom test uitgevers.
- Nederlands Instituut van Psychologen (NIP) (2004). *Documentatie van Tests en Testresearch in Nederland, aanvulling 2004/01*, pp.17-18. Amsterdam: Boom test uitgevers.
- Nijdam, A.D. & Buuren, J.A. van (1994). *Statistiek voor de sociale wetenschappen (zesde druk)*. Alphen a/d Rijn/Zaventem: Samsom Bedrijfsinformatie.
- Oosterbaan, H., Kroes, G., Gent, B. van & Bruyn, E.E.J. de (2006). De WISC-III bij kinderen met ernstige gedragsproblemen, ontwikkelingsproblemen en/of psychiatrische problemen. *Kind en Adolescent, 27*, 57-68.
- Pesch, W. & Ponsioen, A. (2004). Flinterdunne en flagrante Flynn-effecten bij licht verstandelijk gehandicapte kinderen. Aanbevelingen voor het gebruik van de WISC-III. *De Psycholoog, 39* (2), 64-68.
- Resing, W. & Blok, J. (2002). De classificatie van intelligentiescores: voorstel voor een eenduidig systeem. *De Psycholoog, 37* (5), 244-248. Te downloaden via www.jeugdpsycholoog.nl.
- Swanborn, P.G. (1993). *Methoden van sociaal-wetenschappelijk onderzoek (nieuwe editie)*. Meppel/Amsterdam: Boom.
- Uterwijk, J. (red.) (2000). *WAIS-III Nederlandstalige bewerking. Technische Handleiding. David Wechsler*. Lisse: Swets Test Publishers.
- Vander Steene, G., Haassen, P.P. van, Bruyn, E.E.J. de, Coetsier, P., Pijl, Y.J., Poortinga, Y.M., Spelberg, H.C. & Stinissen, J. (1986). *WISC-R, Wechsler Intelligence Scale for Children-Revised. Nederlandstalige uitgave*. Lisse: Swets & Zeitlinger.