Het hertesten van intelligentie bij kinderen: een literatuurstudie

Mark Schittekatte [1]

Samenvatting

In deze literatuurstudie gaan wij op zoek naar antwoorden op een vraag uit de praktijk van de psychodiagnosticus: wanneer is het zinvol om kinderen te hertesten, in het bijzonder wat de intelligentietesting betreft? We tonen eerst aan dat het intelligentiequotiënt, gemeten via Wechsler-instrumenten, een stabiele maat is. Toch kunnen er concrete aanleidingen zijn om tot hertesten te beslissen: "foute" medische of psychopathologische diagnoses en uitgesproken motivatie- of emotionele problemen zijn voorbeelden van problemen tijdens een eerste testafname die de clinicus kunnen doen besluiten opnieuw te meten. De impact van een hertest op korte termijn is in de literatuur goed gedocumenteerd, met als meest opvallende kenmerk een sterkere stijging voor het performale gedeelte. Op basis van ons literatuuronderzoek adviseren wij minstens één jaar, bij voorkeur twee jaar te wachten voor het hertesten met eenzelfde test. Parallelversies van tests en paralleltests vormen ons laatste aandachtspunt. Ondanks het feit dat heel wat deelaspecten van onze centrale vraag via wetenschappelijk onderzoek konden worden onderbouwd, besluiten wij met een oproep voor meer onderzoek rond hertesteffecten en de ontwikkeling van meer diagnostisch parallelmateriaal voor intelligentietesting.

[1] Mark Schittekatte is dr. in de experimentele psychologie en werkzaam aan de faculteit Psychologie en Pedagogische Wetenschappen van de Universiteit Gent als coördinator van het Testpracticum PPW. Hij is tevens lid van de Commissie Psychodiagnostiek van de Belgische Federatie van Psychologen.

Het hertesten van intelligentie bij kinderen: een literatuurstudie

Mark Schittekatte

Inleiding

In dit artikel formuleren we een zo volledig mogelijk antwoord op de vraag: wanneer is het zinvol om kinderen te hertesten, in het bijzonder wat de intelligentietesting betreft? Bij onze zoektocht naar een wetenschappelijk gefundeerd antwoord op voorgaande vraag werd geopteerd voor een literatuuronderzoek van publicaties tussen 1971 en 1999. Via *Psychlit*, een database om vooral Engelstalige, maar ook Nederlandstalige publicaties over "psychologische onderwerpen" terug te vinden, leverde het onderwerp "hertesten én intelligentie" 86 referenties (tijdschriftartikels of hoofdstukken uit boeken) op, waarvan er een dertigtal in dit artikel werden verwerkt.

De rode draad van dit artikel, hierna gestoffeerd met wetenschappelijke bevindingen, kan als volgt worden geschetst:

1. Kan men aannemen dat het intelligentiequotiënt een stabiele maat is?
2. Wat zijn concrete aanleidingen die hertesten zinvol maken?
3. Bij het hertesten op korte termijn is de belangrijkste vraag: wat is de impact van een hertest op het IQ?
4. Wat is het minimumtijdsinterval bij hertesten met hetzelfde instrument?
5. Wat is van belang bij hertesten met een parallelversie of een paralleltest?

De antwoorden op deze deelvragen vormen de structuur van ons betoog.

De besproken instrumenten in dit artikel zijn bijna uitsluitend de Wechsler-intelligentietests[2]. Deze beperking is enerzijds te verklaren door het wetenschappelijk onderzoek dat overwegend deze instrumenten als voorwerp van studie neemt. Anderzijds, als men het belang van een test kan afmeten aan de frequentie waarmee deze in de praktijk wordt toegepast, dan zijn de Wechsler-tests sinds de jaren '60 zonder meer de meest invloedrijke op het vlak van meten van

[2] WPPSI, WISC (1949), WISC-R (1974), WISC-RN (1986) = de Nederlandstalige versie van de WISC-R, WISC-III (1991), WAIS en WAIS(-R).

intelligentie, zowel in de Verenigde Staten als in Europa[3]. Bevragingen van testgebruik bij klinische en schoolpsychologen tonen consistent aan dat de laatste decennia de Wechsler-schalen de meest gebruikte tests zijn voor het meten van cognitieve capaciteiten (Canivez & Watkins,1998). Van 1916 tot de late jaren '50 was de meest gebruikte individuele intelligentietest de Stanford-Binet (Bolen, 1998).

De recente internationale wetenschappelijke literatuur betreffende intelligentiemeting bij kinderen hanteert sinds 1993 meestal de WISC-III als onderzoeksinstrument. Van de WISC-III bestaat echter nog geen Nederlandstalige versie; toch menen wij dat wij de informatie in dat verband in onze studie moesten opnemen.[4]

Telkens synthetiseren wij de belangrijkste bevindingen, wat ons uiteindelijk tot een globaal antwoord zal leiden. Tussendoor nemen wij de vrijheid enkele zijwegen te bewandelen, over aanverwante onderwerpen die ons boeiden tijdens ons literatuuronderzoek.

1 Het IQ is een stabiele maat

De nauwkeurigheid van een testscore, afgezien van wat de test meet, staat bekend als de testbetrouwbaarheid (De Zeeuw, 1996). De betrouwbaarheid is een van de vijf criteria[5] bij de beoordeling van de kwaliteit van tests, gehanteerd door de Commissie Testaangelegenheden Nederland (COTAN) en gepubliceerd in *Documentatie van Tests en Testresearch in Nederland*. Goed nieuws in de nieuwe editie van dit standaardwerk voor Nederlandstalige psychodiagnostici (in druk; zie ook Schittekatte, 1999) is de gestage daling sinds 1982 van het aantal "onvoldoendes" voor het criterium betrouwbaarheid van Nederlandstalige tests.

Er zijn een aantal methoden om de mate van testbetrouwbaarheid te schatten, waarvan de bekendste zijn (De Zeeuw, 1978): de testherhalingsmethode, de paralleltestmethode, homogeniteits- of splitsingsmethode. Bij de testherhalingsmethode gaat men op zoek naar de stabiliteit of consistentie van een test over een tijdsinterval. Het is deze vorm van meten van "nauwkeurigheid" die hier centraal staat. Ook Murphy & Davidshofer (1998) wijzen op het onderscheid dat gewoonlijk wordt gemaakt tussen *betrouwbaarheid*, die de verhouding weer moet geven tussen de werkelijke en geobserveerde variantie, en *stabiliteit*, die refereert aan consistentie van testscores over een bepaalde tijd. Men spreekt in dit verband betreffende de correlaties dan ook eerder van stabiliteitscoëfficiënten dan van betrouwbaarheidscoëfficiënten.

[3] Het Vlaams Forum voor Diagnostiek organiseert in 2000 een "Rondvraag naar diagnostische noden in Vlaanderen" waarin ook de frequentie van het gebruik van psychodiagnostische instrumenten wordt nagegaan.
[4] Heel recentelijk kwam, eindelijk, een initiatief vanuit het Nederlands Instituut voor Psychologen, met budgetten, om een Nederlandstalige versie (voor Nederland én Vlaanderen) van de WISC-III tot stand te brengen.
[5] Andere criteria die de COTAN hanteert, zijn: uitgangspunten van testconstructie, kwaliteit van het testmateriaal en de handleiding, normen én begrips- en criteriumvaliditeit.

De stabiliteit van intelligentietests is een belangrijke karakteristiek aangezien intelligentie als construct een stabiele "trait" veronderstelt (Canivez & Watkins, 1998). Daarbij komt dat intelligentie een hoge graad van erfelijkheid vertoont, wat echter niet betekent dat de individuele intelligentie een vaste maat is. Aanvaarden wij dat intelligentie onderhevig is aan wijzigingen, dan rijzen een aantal vragen zoals: "zijn, onder gewone omstandigheden, intelligentietestscores stabiel of veranderen zij significant tijdens de levensloop?" en "onder welke omstandigheden mogen wij al dan niet systematische veranderingen in intelligentie verwachten?". De eerste vraag komt nu aan de orde, de tweede vraag staat centraal in het volgende deel.

We gaan eerst enkele auteurs bespreken die zelf al overzichten rond onze probleemstelling formuleerden:

1. Murphy & Davidshofer (1998) in hun algemeen handboek *Psychological Testing*.
2. Schuerger & Wit (1989) met hun artikel van een meta-analytische studie.
3. Canivez & Watkins (1998) over langetermijnstabiliteit van IQ als inleiding op hun studie met de WISC-III.
4. Neyens & Aldenkamp (1999) in hun pas verschenen artikel over stabiliteit van intelligentiescores en neuropsychologische maten bij kinderen met ten minste gemiddeld intelligentieniveau.

Vervolgens gaan we enkele afzonderlijke studies nader bekijken:

5. Arbuckle, Maag, Pushkar & Chaikelson (1998) over intellectuele ontwikkeling in levenslopen (met tussenperiodes van 45 jaar).
6. Lally, Lloyd & Kulberg (1987), Sarazin & Spreen (1986) en Vance, Brown & Hankins (1987) over de stabiliteit van de WISC, WISC-R en WAIS-R bij specifieke populaties.
7. Canivez & Watkins (1998), Bolen (1998) en Zimmerman (1996), recente studies over de stabiliteit van de WISC-III, eveneens bij specifieke populaties.

1.1 Longitudinale en cross-sectionele studies suggereren een vrij hoge stabiliteit in IQ's, soms zelfs van de vroege kinderleeftijd tot in de late volwassenheid. Murphy & Davidshofer (1998) citeren in dit verband een tiental studies, vooral uit de jaren '50 tot '70. De laagste stabiliteit blijkt bij vroegkinderlijke metingen (moeilijk meetbaar) en late volwassenheid (WAIS-IQ-veranderingen op latere leeftijd). Het IQ blijkt het meest stabiel tussen de leeftijd van 20 en 35 jaar. Rekening houdend met de methodologische problemen (met als bekendste het Flynn-

effect[6]) toont de literatuur over leeftijdsgebonden afname van intelligentie na 35 jaar aan dat de "echte" afname in algemene intelligentie gradueel is en later start dan voorheen aangenomen. Intelligentie blijkt relatief stabiel, ook voorbij de leeftijd van 60 jaar, en toont pas veel later duidelijke achteruitgang. Dus, ondanks het feit dat intelligentietests geen gefixeerd aspect van een individu beschrijven, blijkt er voldoende bewijs voor de consistentie van het IQ over lange periodes van individuen (Murphy & Davidshofer, 1998). Ook in andere recente handboeken zoals van Anastasi & Urbina (1997), Janda (1998) en Kline (2000) vinden wij wetenschappelijke funderingen ter ondersteuning van de stabiliteit van cognitieve capaciteiten.

1.2 In het artikel van Schuerger & Witt (1989) werden "test-hertestdata" van 79 bronnen gereviewd. Via multiple-regressietechnieken werden de effecten geanalyseerd van verschillende factoren op de stabiliteit van individueel geteste intelligentie. Vijf intelligentietests werden onderzocht: de Stanford-Binet Intelligentie Schaal, de WISC, WISC-R, WAIS en WAIS-R. Tijdsinterval en leeftijd waren significante predictoren van betrouwbaarheid, geslacht en instrument niet. De twee belangrijkste tendensen zijn als volgt samen te vatten:

a. Hoe groter het tijdsinterval tussen twee metingen, hoe lager de stabiliteitscoëfficiënten;
b. Hoe jonger het kind bij de eerste testing, hoe lager de stabiliteitscoëfficiënten.

1.3 Canivez & Watkins (1998) bespreken eveneens de langetermijnstabiliteit van de WISC (9 studies) en de WISC-R (15 studies). Significante en gemiddelde tot hoge test-hertestbetrouwbaarheidscoëfficiënten werden gerapporteerd (*r*'s van .50 tot .90). Belangrijker nog, de leereffecten bleken te verdwijnen als het hertestinterval groter dan een jaar was. Als leereffecten werden geobserveerd in langetermijnstabiliteitsstudies, was de effectgrootte gewoonlijk erg klein of van geen enkel praktisch belang. Juliano, Haddad & Carroll (1988) vonden ook langetermijnstabiliteit voor de WISC-R-factorstructuur bij jongeren met leermoeilijkheden.

1.4 Neyens & Aldenkamp (1999) analyseren de resultaten van enkele studies over test-hertestbetrouwbaarheid van de WISC-R bij specifieke populaties en concluderen: "*over het algemeen worden lagere stabiliteitscoëfficiënten gerapporteerd voor verstandelijk gehandicapten dan voor kinderen met leerstoornissen*". Zij stellen verder vast dat ook hoogbegaafde kinderen een relatief lagere stabiliteit vertonen, zowel met korte als met langere tijdsintervallen tussen twee testfasen. Ten slotte onthouden wij uit hun onderzoek een lage en in het beste geval een redelijke stabiliteit van de subtestscores van de WISC-RN. Dit bevestigt het idee dat veranderingen in subtestscores niet moeten worden meegenomen bij de evaluatie

[6] Flynn documenteerde het stijgende IQ in de USA tussen 1932 en 1978; ook meer recent werden stijgingen van 5 tot 25 IQ-punten in één generatie vastgesteld. Er is nog steeds geen sluitende verklaring voor de gestage stijging over de laatste 75 jaar: de stijging is te sterk om het aan genetische factoren te wijten; wel worden betere scholing, hogere levensstandaard, betere voeding en stijgende blootstelling aan geavanceerde technologie vermeld als mogelijke verklaringen. De stijgingen in het performale

van de cognitieve ontwikkeling van kinderen met een gemiddeld intelligentieniveau. Neyens & Aldenkamp pleiten ervoor dat indien onderzoekers de cognitieve ontwikkeling op specifieke domeinen willen onderzoeken, het beter is om hiervoor specifieke tests af te nemen of gebruik te maken van de Kaufman-factoren van de WISC-RN.

1.5 De hypothese dat individuele verschillen in volwassen intellectuele ontwikkeling variatie reflecteren in levenscontext en persoonlijkheid, werd onderzocht bij 132 wereldoorlog II-veteranen door Arbuckle e.a. (1998). Intelligentiedata van drie tijdstippen over 45 jaar werden geanalyseerd. De prestatie nam af na 45 jaar maar nam bv. toe voor de subtest Woordenschat. De correlaties tussen metingen tijdens Wereldoorlog II en de scores van de jaren '90 toonden vrij stabiele resultaten voor de intelligentie. Een meer geëngageerde levensstijl voorspelde minder achteruitgang voor de meeste subtests, betere gezondheid en sterkere introversie voorspelden minder achteruitgang voor sommige subtests. De individuele variatie in het traject gevonden in deze studie is relatief klein t.o.v. van de algehele stabiliteit van individuele verschillen in intelligentie in de volwassenheid. Zij concluderen dat intellectuele groei en achteruitgang geen gefixeerd ontwikkelingstraject volgen. Factoren gerelateerd aan individuele en omgevings-karakteristieken modereren groei en achteruitgang, en leiden tot individuele verschillen in het behoud van intellectuele mogelijkheden. Zij steunen recente modellen op zoek naar faciliterende processen voor de intellectuele ontwikkeling i.p.v. de "onvermijdelijke achteruitgang". Ook de studie van Gold, Andres, Etezadi, Arbuckle, Schwartzman & Chaikelson (1995) toonde voorgaande aan en zij concludeerden verder dat jongvolwassen gemeten intellectuele mogelijkheden de belangrijkste determinant is voor intellectuele ontwikkeling op latere leeftijd.

1.6 De studies van Lally e.a. (1987), Sarazin & Spreen (1986) en Vance e.a. (1987) hebben met elkaar gemeen dat hun onderzoek het hertesten op langere termijn met de Wechsler-instrumenten bij specifieke populaties betreft. I.v.m. de vastgestelde stabiliteit worden telkens "positieve" conclusies getrokken.

Lally en haar collega's onderzochten de stabiliteit van de WISC-R over drie jaar bij 60 leergestoorde leerlingen (gemiddeld 8,8 jaar oud), die "*special education service*" ontvingen. Hun besluit is dat de WISC-R ook bij leergestoorde jongeren als een stabiel instrument mag worden erkend. In hun artikel wijzen zij verder op het belang van de WISC-R als instrument om leerstoornissen te detecteren. De federale definitie in de USA van "leergestoord" is immers "…*when there is a significant discrepancy between ability and achievement…* ". De wet vraagt ook om kinderen die speciaal onderwijs krijgen elke drie jaar te herevalueren, zolang ze worden begeleid. Lally e.a. houden in hun artikel een pleidooi voor niet-systematisch herhalingsonderzoek daar de WISC-R heeft bewezen stabiel te zijn bij vele populaties, nu ook

---

gedeelte zijn het sterkst voor maten van visueel-spatieel redeneren, wat een reflectie kan zijn van de toenemende omgang met

bij een leergestoorde populatie. Zij stellen dat wanneer een leergestoorde leerling twee WISC-R-afnames kreeg die binnen de standaardmeetfout vallen, de waarschijnlijkheid dat een derde afname significant zal verschillen erg laag is. Alternatieven voor de driejaarlijkse evaluatie worden gesuggereerd, waaronder werken met een verkorte versie van de WISC-R.

Sarazin & Spreen (1986) deden onderzoek naar stabiliteit van de WISC en WAIS-R (en zeven andere neuropsychologische tests) over een interval van 15 jaar bij leergestoorde en neurologisch gestoorde personen. Hun algemeen eindresultaat: er bleken hoge en significante correlatiecoëfficiënten tussen tijdstip 1 (meting op 10 jaar) en tijdstip 2 (meting op 25 jaar), zelfs met gebruik, 15 jaar later, van een versie voor volwassenen voor heel wat tests.

Vance e.a. (1987) onderzochten adolescenten uit het buitengewoon onderwijs met een tijdsinterval van drie jaar. Eerst werd de WISC-R afgenomen en later de WAIS-R: hogere scores op VIQ en TIQ bleken op tijdstip 2. De auteurs verwachtten geen leereffecten, gelet op het ruime tijdsinterval en het werken met een ander instrument. Het VIQ steeg gemiddeld van 72.4 naar 77.9, het PIQ van 77.2 naar 79.5 en het TIQ van 72.4 naar 77.5. De absolute verandering bleek ongeveer 5 IQ-punten, wat aantoont dat clinici redelijk zeker mogen zijn dat de scores op de WISC-R van jongeren uit het buitengewoon onderwijs niet veel zullen veranderen als zij worden hertest met de WAIS-R. De WAIS-R-stabiliteit bleek ook bevredigend bij volwassen psychiatrische patiënten na een hertest na 15 maanden (Hawkins & Sayward, 1994) en bij "normale" 65-plussers bij een hertest na een jaar (Snow, Tierney, Zorzitto, Fisher & Reid, 1989).

1.7 Laat ons ook enkele recentere studies in dit verband apart bespreken die telkens de WISC-III als onderzoeksinstrument hanteerden.

Canivez & Watkins (1998) bestudeerden recentelijk de langetermijnstabiliteit van de WISC-III bij 667 leerlingen (gemiddelde leeftijd bij eerste testing 9,18 jaar) die getest werden om al dan niet in het buitengewoon onderwijs terecht te komen. Met een hertestinterval van gemiddeld 2,87 jaar werden stabiliteitscoëfficiënten van .87, .87 en .91 (telkens $p < .0001$) vastgesteld voor respectievelijk VIQ, PIQ en Totaal-IQ. Deze hoge waarden qua stabiliteit voldoen aan de strengste normen (dicht bij of meer dan het .90-criterium, aangeraden door psychometrici zoals bv. Salvia & Ysseldycke, 1991) en zijn duidelijk hoger dan voor vorige WISC-versies bij soortgelijke populaties. De stabiliteit bleek ook adequaat voor diagnostische doelen voor de factoren "Verbaal Begrip" en "Perceptuele Organisatie", maar niet voor "Vrijheid van Afleidbaarheid" en de "VIQ-PIQ-discrepantie".

Bolen (1998) benadrukt anderzijds bij zijn studie met 70 licht tot matig mentaal gehandicapte jongeren (gemiddeld bij eerste testing 10,7 jaar), eveneens met de WISC-III, een vrij grote

---

complexe visuele stimuli (bv. televisie, computers, enz.).

individuele variatie bij hertesting, in het bijzonder voor het VIQ. In deze populatie werden na een gemiddeld hertestinterval van ongeveer drie jaar slechts stabiliteitscoëfficiënten tussen .62 en .74 vastgesteld. Dat testbetrouwbaarheden altijd lager zijn aan de uiteinden van de distributie, vormt hiervoor een mogelijke verklaring.

Zimmerman & Woo-Sam (1996) rapporteren een meta-analyse van 16 studies waar jongeren eerst werden getest met de WISC-R en gemiddeld 29 maanden later met de WISC-III. Voor het Totaal-IQ bleek gemiddeld een achteruitgang van 6.4 IQ-punten, voornamelijk toegeschreven aan de nieuwe, "strengere" normen van de WISC-III. Stabiliteitscoëfficiënten werden in deze studie niet gerapporteerd.

Afsluitend kunnen wij concluderen dat veel wetenschappelijk onderzoek een sterke stabiliteit van het IQ op het spoor kwam. De consistentie van het cognitief functioneren van individuen werd meermaals onderschreven. Er werd in dit verband gewezen op het belang van individuele en omgevingskarakteristieken als moderators van groei en achteruitgang van intellectuele mogelijkheden. Als er een lagere stabiliteit wordt gerapporteerd, betrof het metingen in de vroege kindertijd of in de latere volwassenheid. Vrij hoge stabiliteitswaarden voor het IQ worden ook bij specifieke populaties teruggevonden. Als tendens komt naar voren: lagere stabiliteitscoëfficiënten bij verstandelijk gehandicapten dan bij kinderen met leermoeilijkheden. Tenslotte onthouden wij de slechts redelijke of zelfs lage langetermijnstabiliteit van de subtestscores van de WISC-RN.

2 Wat zijn concrete aanleidingen voor hertesten?

Over deze tweede gedachte werd nauwelijks wetenschappelijke literatuur gevonden; enkel in handboeken komt dit onderwerp ter sprake, waarbij de nadruk dan wordt gelegd op de clinicus die de specifieke situatie moet inschatten. Zo kunnen bv. emotionele factoren een "afgeremd IQ" tot gevolg hebben, en na één jaar therapie blijkt een cliënt vele IQ-punten hoger te scoren richting een meer "reëel IQ". Hierover zijn echter geen "groepsuitspraken" mogelijk en bijgevolg ook geen wetenschappelijk onderzoek; wel moet elk individueel geval apart worden beoordeeld. De scores op de subtests worden regelmatig vermeld als aanhechtingspunten.

Welke factoren leiden nu tot variabiliteit in testscores? Murphy & Davidshofer (1998) verwijzen naar een lijst van Thorndike (uit 1949) van mogelijke bronnen van variabiliteit of ongewilde inconsistentie in scores op een test. Er worden zes groepen onderscheiden, geïllustreerd in Tabel 1. Ook Drenth (1975; vermeld in De Zeeuw, 1978) geeft een gelijkaardig overzicht van de factoren die "bronnen der testvariantie" kunnen vormen. Hij maakt daarbij een onderscheid tussen blijvende factoren die niet aan het moment van testafneming zijn gebonden en tijdelijke factoren die dat wel zijn. Voorts maakt hij een onderscheid tussen specifieke factoren gebonden

aan de bijzonderheden van de desbetreffende test en algemene factoren, waarvan de erdoor veroorzaakte variantie ook bij andere (equivalente) tests aanwezig is. Tenslotte onderscheidt hij de variantie veroorzaakt door de interindividuele verschillen en de variantie veroorzaakt door factoren die buiten de onderzochte personen zijn gelegen.

De belangrijkste bron om te beslissen om te hertesten is o.i. groep 3: "tijdelijke, maar algemene karakteristieken van het individu" met o.a. gezondheidsproblemen, vermoeidheid (slaap), motivatie, emotionele druk, testcondities (warmte, licht, ventilatie,…). Ook bronnen uit groep 5, nl. klaarheid van instructies, respecteren van de tijdslimieten,…, m.a.w. de testleider/geteste-interactie (qua persoonlijkheid, ras, geslacht,…) en fouten in het beoordelen van een testprestatie, kunnen aanleiding geven tot de beslissing om te hertesten.

Tabel 1: Mogelijke bronnen van variabiliteit, gebaseerd op Thorndike (1949; beschreven in Murphy & Davidshofer, 1998)

1. Blijvende en algemene karakteristieken van het individu
   o.a. "test-wijsheid" of "test-naïviteit", niveau van kunde van een proefpersoon
2. Blijvende en specifieke karakteristieken van het individu
   a. specifiek t.o.v. de test als geheel
      o.a. stabiele response-sets, specifieke vaardigheid
   b. specifiek t.o.v. bepaalde test-items
      o.a. vertrouwdheid met specifieke opgaven
3. Tijdelijke en algemene karakteristieken van het individu
   o.a. gezondheid, vermoeidheid, motivatie, emotionele druk, externe condities
4. Tijdelijke en specifieke karakteristieken van het individu
   a. specifiek t.o.v. de test als geheel
      o.a. niveau van ervaring, specifieke trucs
   b. specifiek t.o.v. bepaalde test-items
      o.a. fluctuaties in het menselijk geheugen, aandacht
5. Systematische factoren die de afname van de test of de testappreciatie beïnvloeden
   a. testcondities
      o.a. mogelijke distractoren, klaarheid van instructies, strengheid tijdslimieten
   b. interactie testleider/pp.
      o.a. invloed persoonlijkheid, geslacht, ras als facilitator of inhibitor
   c. fouten in beoordelen of scoren van de test
6. Andere variantiebronnen
   a. geluk (giswerk)
   b. tijdelijke afleidingen

Murphy & Davidshofer (1998) bespreken ook drie hoofdredenen waarom bij een eventuele tweede afname verschillende scores kunnen voorkomen:

a. reële veranderingen in het gemetene (bv. vooruitgang leesvaardigheid);
b. reactiviteit, nl. na afname van een test antwoorden gaan opzoeken; op die manier wordt een test soms een katalysator;
c. leer- of geheugeneffecten bij korte tijdsintervallen.

In andere algemene handboeken i.v.m. psychodiagnostiek worden enkele van de bij Thorndike vermelde factoren benadrukt.

Kievit e.a. (1992) wijzen op het mogelijke belang van het effect van een behandeling en de afname door al dan niet dezelfde proefleider (Murphy & Davidshofer, groep 5).

Anastasi & Urbina (1997) vermelden belangrijke veranderingen in de familiestructuur, adoptie, ernstige ziekte en therapeutische begeleiding als belangrijke factoren die instabiliteit in intelligentie kunnen teweegbrengen (Murphy & Davidshofer, groep 3). Zij verwijzen ook naar studies rond de invloed van ongunstige leer- en leefomstandigheden van kinderen op intellectuele ontwikkeling.

In het handboek van Drenth & Sijstma (1990) worden motivatie, emotionaliteit en stemming vermeld als mogelijke beïnvloedende factoren bij metingen van intellectuele capaciteiten. Zij onthouden zich van een opsomming van condities, maar gaan wel in op een aantal toevallige invloeden op testgedrag, met name een black-out, een "helder moment", tijdelijk concentratieverlies, slaperigheid, fluctuaties in het "arousal"-niveau en zich plots opdringende gedachten die niets met de test hebben te maken (Murphy & Davidshofer, groep 3). Het probleem bij deze voorbeelden, stellen de auteurs terecht, is dat men er zich niet veel kan bij voorstellen (wat is precies een helder moment?) en dat het volstrekt onduidelijk is hoe en in welke mate een bepaald fenomeen de testprestatie zal beïnvloeden. Naast de factoren die gebonden zijn aan een specifieke testsessie, wijzen ze ook op mogelijke beïnvloeding door meer "chronische" factoren (Murphy & Davidshofer, groep 1).

Bij kinderen met een chronische neurologische aandoening (bv. epilepsie) gaat men ook de cognitieve capaciteiten hertesten (Neyens & Aldenkamp, 1999). De cognitieve deterioratie wordt nagegaan als aanwijzing voor een verslechtering van de neurologische toestand. Hertesten binnen een relatief korte periode is dan soms noodzakelijk om een betrouwbare evaluatie uit te voeren, vooral bij neurochirurgische of experimenteel medicamenteuze behandeling.

Luteijn, Deelman & Emmelkamp (1990) bespreken een gelijkaardige diagnostische en/of therapeutische reden om tot hertesten te besluiten bij volwassenen, met name bij duidelijke cognitieve stoornissen, zoals dementie, waarbij een achteruitgang van intelligentie mogelijk is. Aan de hand van herhaalde intelligentiebepalingen wordt bijvoorbeeld de mate van herstel of achteruitgang nagegaan. De auteurs realiseren zich dat men in de praktijk vaak niet over het premorbide (= voor een aandoening optrad) intelligentieniveau beschikt. Zij spreken in dit verband van de mogelijkheid om in de Wechsler-subtests toch aanwijzingen te vinden. Er blijken subtests die relatief weinig ('hold'-subtests) door leeftijd worden beïnvloed en subtests die relatief gevoelig zijn voor leeftijd ('do not hold'-subtests). Er wordt verondersteld, en ook wel door onderzoek ondersteund, dat de 'hold'-subtests (Informatie, Woordenschat, Plaatjes Ordenen en Figuurleggen) minder door pathologie zullen worden beïnvloed dan de 'do not hold'-subtests (Overeenkomsten, Cijferreeksen, Substitutie en Blokpatronen). Door de verhouding tussen beide soorten subtests uit te rekenen, zou er in het geval van een duidelijk hogere score

op de 'hold'-subtests sprake zijn van pathologische achteruitgang in intelligentie. Dergelijke deterioratie-index bleek voor het individuele geval soms te weinig valide.

Foute diagnoses van medische of psychopathologische aard, zoals bijvoorbeeld het over het hoofd zien van duidelijke gehoorproblemen of een ernstige depressie, kunnen uiteraard ook aanleiding geven tot een besluit van hertesting.

Ten slotte kunnen er in de praktijk administratieve redenen opduiken die een hertesting van het intelligentieniveau noodzakelijk maken.

Samengevat, in handboeken van psychodiagnostiek worden een reeks concrete aanleidingen voor hertesten beschreven. Algemene richtlijnen werden in dit verband niet teruggevonden, tenzij: de clinicus moet de situatie zelf inschatten.

Een overzicht van Thorndike (Murphy & Davidshofer, 1998) van mogelijke bronnen van variabiliteit in intellectuele capaciteiten leverde ons een eerste reeks van concrete aanleidingen op. Vooral de derde groep factoren - " tijdelijke, maar algemene karakteristieken van het individu" met o.a. gezondheid, vermoeidheid, motivatie, emotionele druk,… - lijken ons belangrijke redenen om te beslissen tot hertesten.

Foute diagnoses van medische of psychopathologische aard en administratieve redenen zijn twee andere mogelijke aanleidingen waarbij het zinvol kan blijken om de cognitieve capaciteiten opnieuw te gaan testen.

3 De impact op het IQ van een hertest op korte termijn

Meten in de psychologische praktijk heeft dikwijls tot gevolg dat de gemeten eigenschap of beter nog, de persoon als drager van de eigenschap, door een meting verandert. De belangrijkste oorzaken hiervan zijn gewenning, bekendheid en aanleren (De Zeeuw, 1978). In positieve zin kan die verandering er komen door bijvoorbeeld het vertrouwd raken met de testsituatie of de testleider, of het bekend raken met de soort testopgaven door vroegere testafnames; in negatieve zin door bijvoorbeeld verveling of tegenvallende resultaten voorheen. M.a.w. de ervaring opgedaan tijdens een vorige meting zal in de meeste gevallen niet geheel zonder uitwerking zijn op de testscores bij een volgende meting. Als het hertesten op korte termijn gebeurt (i.c. binnen het jaar na vorige afname) zullen die invloeden sterker zijn. Hier gaan we nu dieper op in. Gebaseerd op tientallen stabiliteitsonderzoeken met de WISC en de WISC-R concluderen Canivez & Watkins (1998) dat leereffecten verdwijnen als het hertestinterval groter is dan één jaar. Uit meerdere studies blijkt de voor de hand liggende, reeds vermelde, algemene regel: hoe groter het interval tussen twee testmomenten, hoe kleiner

het leereffect (bevestigd in handboeken zoals De Zeeuw, 1996 en Anastasi & Urbina, 1997). Door het interval groot genoeg te nemen, kan men het geheugeneffect pogen tegen te gaan. Rond de vraag welk interval optimaal is voor hertestbetrouwbaarheidsmetingen, proberen wij in het volgende hoofdstuk een advies te formuleren.

De belangrijkste bevinding van Canivez & Watkins (1998) over het hertesten op korte termijn is een significant leereffect, in het bijzonder gereflecteerd in hogere scores voor het PIQ, een fenomeen dat wij ook in andere specifieke studies, hierna beschreven, regelmatig terugvinden. Neyens & Aldenkamp (1999) pleiten voor normgegevens over het te verwachten test-hertesteffect, om bij een interval korter dan anderhalf jaar een onderscheid te kunnen maken tussen het effect als gevolg van de hertest en de werkelijke verandering (bv. door behandeling). De vraag naar de impact van een hertest wordt regelmatig gesteld door practici die, omwille van een van bovenvermelde redenen, worden verplicht om personen te hertesten in een interval kleiner dan één jaar.

Het onderzoek van Bolen (1988), gebaseerd op een zestal studies met de WISC en de WISC-R, toont stevige test-herteststabiliteitsscoëfficiënten op korte termijn aan voor zowel het TIQ, VIQ als PIQ, meer precies correlaties tussen .80 en .90. Significante leereffecten werden echter gereflecteerd in hogere scores na hertesten, opnieuw in het bijzonder voor het PIQ. De test-hertestbetrouwbaarheidscoëfficiënten van de subtests waren in bijna alle gevallen lager dan de TIQ- test-hertestbetrouwbaarheidsintervallen.

In de Franstalige handleiding van de Wechsler III (1996) lezen we in dit verband: 180 personen werden tweemaal getest binnen een interval van gemiddeld 30 dagen. Tussen de eerste en tweede meting steeg het totale IQ gemiddeld een tiental punten. Het verbale IQ steeg gevoelig minder (5 à 6 punten) dan het performale IQ (13 à 14 punten). Er waren aanwijzingen van een leereffect dat de WISC-III-resultaten beïnvloedde. Bij subtests die een "nieuwe taak" brachten, was de vooruitgang het sterkst bij hertesten. Zo bleek bijvoorbeeld bij de proefpersonen van 6 à 7 jaar, de gemiddelde score op de subtest Plaatjes Ordenen er 2.7 punten op vooruit te gaan, terwijl er bij de subtest Woordenschat slechts een gemiddelde stijging van 0.3 punten werd vastgesteld.

In de Amerikaanse handleiding van de WISC-III (1991; cfr. Bolen, 1998) wordt een onderzoek gerapporteerd bij 353 kinderen waar bij hertesten op korte termijn (Mdn = 23 dagen) eveneens een duidelijk leereffect optreedt (voor TIQ 7 à 8 punten en voor PIQ 12 à 13 punten vooruitgang), vooral voor de subtests Plaatjes Ordenen en Substitutie. Bij de oudere kinderen bleek een zelfde effect voor de subtest Onvolledige Tekeningen. De test-hertestbetrouwbaarheidscoëfficiënten voor de drie IQ's en de vier factorstructuren waren in het algemeen uitstekend, variërend van .71 (FDI = freedom from distractibility index, voor 6- à 7-jarigen) tot .95 (TIQ voor 14- à 15-jarigen). Het vertrouwd zijn met een subtest blijkt al bij een

tweede afname. We vermoedden dan ook dat het effect nadien zou verdwijnen, als het nieuwe, het verrassende als het ware uitdooft.

De recente studie van Neyens & Aldenkamp (1999) stelt deze redenering echter in vraag. Zij testten 59 kinderen tussen 4 en 12 jaar drie keer met de Nederlandstalige WISC-R met een tijdsinterval van gemiddeld zes maanden. De stabiliteit van de scores van deze kinderen met een gemiddeld intelligentieniveau lag iets lager dan bv. in de Amerikaanse handleiding van de WISC-R, met name gemiddeld .78, maar een verschillende "strengere" statistische methode wordt als verklaring aangebracht. Het PIQ vertoont een lagere stabiliteit (0.68) dan het VIQ en het TIQ. Dit stemt overeen met eerdere bevindingen. De gemiddelde toename bij de eerste hertest was "+3" voor het VIQ, "+9" voor het PIQ en "+7" voor het TIQ. Bij de tweede keer hertesten namen het PIQ en (bijgevolg) het TIQ gemiddeld genomen nog steeds toe, terwijl zou worden verwacht dat dit niet meer het geval zou zijn. Dit betekent dat een zeker hertesteffect nog altijd een rol speelde bij de derde hertest voor de performale score, waarschijnlijk door de tijdswinsten die hierbij kunnen worden behaald. Een methode om een dergelijke overschatting van de score te vermijden, is het vergelijken van scores gebaseerd op dichotome maten (falen dan wel behalen) in plaats van scores gebaseerd op tijdswinst. Snelheid reflecteert namelijk een soort automatisatie en accuratesse van de oplossingsmethode die wordt gehanteerd. Scoring aan de hand van een dichotome maat houdt geen rekening met deze processen. De data van Neyens & Aldenkamp kunnen volgens de auteurs worden gebruikt als normgegevens voor verwachte test-hertesteffecten bij onderzoek bij kinderen met een gemiddeld intelligentieniveau. Hun gepubliceerde data kunnen echter niet zonder meer worden toegepast bij het hertesten van kinderen met een lager of hoger intelligentieniveau. Op grond van deze correlatiecoëfficiënten kan niet zomaar een bij het individu toepasbare regel voor correctie of te verwachten test-hertesteffect worden getraceerd: de individuele variabiliteit in scores was hiervoor te groot. Een hoog percentage kinderen scoort bij het hertesten meer dan twee standaardfouten hoger dan bij de eerste testafname. Dit betekent dat een vermindering van de IQ-score bij hertesten met de alhier gerapporteerde gemiddelde verschilscore in veel gevallen tot een onderschatting van het test-hertesteffect kan leiden. Subtestscores vertoonden slechts een redelijke of zelfs lage test-herteststabiliteit. We herhalen in dit verband het idee dat veranderingen in subtestscores niet moeten worden meegenomen bij de evaluatie van de cognitieve ontwikkeling van kinderen met een gemiddeld intelligentieniveau. Net als bij het hertesten met de WISC-RN, wordt bij de WPPSI bij de eerste hertest in dezelfde studie van Neyens & Aldenkamp (1999) de grootste toename in het PIQ (+6) waargenomen, gevolgd door het TIQ (+5) en het VIQ (+1). Bij de tweede hertest is het PIQ vrij stabiel (+1), terwijl het VIQ (+6) en (bijgevolg) het TIQ (+4) toenemen. Dit onverwachte uitblijven van een toename in score van het PIQ bij de tweede hertest kan veroorzaakt zijn door een plafondeffect, d.w.z. dat het gemiddelde PIQ bij de tweede afname al 118 was en zodoende niet veel meer kon stijgen bij de derde testafname. Neyens & Aldenkamp pleiten voor een normering van de WPPSI bij een Nederlandse populatie (ondertussen ook gebeurd voor Vlaanderen!), evenals voor

onderzoek naar de overgang tussen het WPPSI en WISC-RN, beide noodzakelijk om een vroege bepaling van veranderingen in cognitieve aanleg mogelijk te maken. Zij bestudeerden ook de stabiliteit van enkele andere instrumenten in hun studie. Alleen de Stroop-kaarten bereikten een uitstekende test-hertestbetrouwbaarheid over de drie testsessies; de correlatiecoëfficiënten voor de Trail Making Test, de 15-Woordentest en de Rey Complexe Figuur Test zijn redelijk tot goed.

Legagnoux, Michael, Hocevar & Maxwell (1990) onderzochten 2000 lagereschoolkinderen op 26 SOI-LA-subtests, een instrument gebaseerd op Guilford's Structure-Of-Intellect (SOI)-model. De kinderen werden hertest binnen 2 à 4 weken. Als algemene bevindingen kwamen naar voor:

- jongere kinderen maken meer vooruitgang;
- meer vooruitgang voor minderbegaafde dan voor hoogbegaafde kinderen;
- meer vooruitgang als de posttest gelijk is aan de pretest dan bij het gebruik van een paralleltest;
- geen verschillende scores tussen de geslachten.

Wij onthouden hier het belang van leeftijd en cognitieve capaciteiten als belangrijke variabelen bij het hertesten van intellectuele draagkracht. De auteurs pleiten afsluitend voor meer rapportage van hertesteffecten in handleidingen en hopen dat psychometristen meer gaan stimuleren tot de ontwikkeling van parallelversies van instrumenten.

Uit het overzicht van Klauer (1993) over het effect van hertesten bij onderzoek naar "learning potential" viel ons als belangrijkste aanbeveling op: werk experimenteel (met controlegroep) en niet quasi-experimenteel om leereffecten te onderscheiden van interventie(/behandelings)-effecten bij hertesting. In de literatuur rond intelligentiemeting vonden wij slechts één studie die de experimentele methode hanteerde.

Van der Doef, Kwint en van de Koppel (1989) rapporteren over een onderzoek waarbij de WISC-RN tweemaal werd afgenomen bij 42 moeilijk lerende kinderen van zeven tot twaalf jaar met een interval van vier weken. In de experimentele conditie werd bij de eerste afname op ieder gemist item de juiste oplossing gegeven. In de controleconditie werd de test afgenomen volgens de standaardprocedure. Ongeacht de conditie toonden zowel het PIQ als het TIQ een significante stijging, het VIQ bleef nagenoeg gelijk. De controle- en de experimentele groepen verschilden niet in niveau bij de eerste afname. Bij de tweede afname scoorde de experimentele groep (niet significant) hoger dan de controlegroep. De resultaten manen aan tot voorzichtigheid bij het op korte termijn herhalen van een onderzoek met de WISC-RN. De belangrijkste conclusie was: een hertest op korte termijn met de WISC-RN flatteert het TIQ en PIQ, terwijl dit voor het VIQ niet het geval is. Bovendien blijken moeilijk lerende kinderen in staat te leren van een afname van de WISC-RN. Van hulp in de vorm van het geven van goede oplossingen, lijken de leerlingen van dit onderwijs nauwelijks te leren.

De algemene regel "hoe groter het interval tussen twee testmomenten, hoe kleiner het leereffect" en de bewering dat leereffecten verdwijnen als het hertestinterval groter is dan één jaar, vormden twee belangrijke elementen, als inleiding van de bespreking van onze literatuurstudie aangaande de impact op het IQ van een hertest op korte termijn.

Bij het hertesten binnen het jaar is de belangrijkste tendens de hogere scores voor het performaal IQ. Dit effect blijkt niet alleen na een tweede maar ook na een derde afname, wat "vooruitgang in score omdat het nieuwe taken betrof" als verklaring ontkracht. Leeftijd en cognitieve capaciteiten kwamen als belangrijke variabelen naar voor i.v.m. het kortetermijnhertesten van cognitieve mogelijkheden. Het experimentele onderzoek moet in dit werkveld veel meer dan nu worden toegepast. Voorzichtigheid voor leereffecten is geboden bij het herhalingsonderzoek op korte termijn met de Wechsler-instrumenten. Toch blijken er uit veel studies stevige test-herteststabiliteitscoëfficiënten op korte termijn.

4 Het minimumtijdsinterval bij hertesten met eenzelfde test

Onze zoektocht naar indicaties in de wetenschappelijke literatuur over het minimumtijdsinterval voor het hertesten met een Wechsler-instrument leverde een magere oogst op.

Kievit e.a. (1992) halen aan dat sommige tests in de handleiding vermelden dat ze binnen een bepaald tijdsbestek geen tweede keer mogen worden gebruikt. Zo wordt ook de Nederlandstalige WISC-R vermeld met een minimumtijdsinterval van twee jaar.

Zoals reeds hoger aangehaald, geven Canivez & Watkins (1998) n.a.v. hun studie over langetermijnstabiliteit aan, dat voor de WISC en de WISC-R de leereffecten bleken verdwenen als het hertestinterval groter was dan één jaar.

Een van de besluiten van de meta-analyse i.v.m. hertesten met de WISC-R en de WISC-III van Zimmerman & Woo-Sam (1996) is het advies van een minimumtijdsinterval van negen maanden tussen twee afnames. De alom weergevonden stijging bij hertesten van het IQ, vooral van het performale gedeelte, bleek minimaal als minstens negen maanden tussen de eerste en de tweede testing werd gewacht.

Wij beschikken eveneens over het deskundig oordeel van Jacques Gregoire[7], auteur van de Franstalige versies van de WISC-III. Hij concludeert, samen met Bolen (1998b), dat op basis

[7] Hoogleraar aan de UCL in het vakgebied psychodiagnostiek

van het hem bekende onderzoek naar het hertesteffect het interval tussen twee afnames met de WISC-III niet minder dan één jaar mag bedragen.

In het algemeen wordt, volgens Neyens & Aldenkamp (1999), bij een interval tussen herhaalde testafnames dat langer is dan twee jaar, het test-hertesteffect als minimaal beschouwd en dus als een controle achterwege gelaten. Zij pleiten voor normgegevens over het te verwachten test-hertesteffect in het belang van heronderzoekingen uit te voeren na een kortere periode dan twee jaar, bv. bij hun onderzochte populatie, met name kinderen met chronische en maligne neurologische aandoeningen. Het is hierbij van belang, besluiten ze, om bij zo'n kort interval een onderscheid te kunnen maken tussen het effect als gevolg van de hertest en de werkelijke verandering (bv. door behandeling). Zij voorzien dan ook als een van de eersten in zulke normgegevens voor kinderen met een gemiddelde intelligentie.

Concluderend, het minimuminterval bij het hertesten met de WISC, WISC-R of WISC-III ligt ergens tussen negen maanden en twee jaar. Indien mogelijk moet men, om alle test-hertesteffecten te vermijden, tussen twee afnames twee jaar laten verlopen. Wacht men minimaal één jaar, dan zullen de invloeden van herhaalde meting beperkt zijn.

We vonden, afsluitend, wel nog drie studies terug, twee Nederlandse en één Amerikaanse, waarin men onderzoek uitvoerde naar hertesteffecten in het kader van het regulier herhalingsonderzoek binnen eigen land.

Dekker, Mackaay-Cramer & De Bruyn (1990) vermelden dat er in het speciaal (of buitengewoon) onderwijs in Nederland regulier met vaste intervallen een intelligentietest wordt afgenomen; degemiddelde tijdsduur tussen de afnames blijkt bijna twee jaar te zijn. Hij besteedt in zijn artikel ook aandacht aan het verouderen van normen met een overschatting van het IQ als gevolg, en aan een vergelijking van de WISC- R (vertaalde versie) en de WISC- RN (met Nederlandse en Vlaamse normen) bij hertesten van kinderen in speciaal onderwijs.

Ook Heessels, Meijs, Feltzer & Eilander (1992) voerden onderzoek uit dat kaderde binnen het reguliere herhalingsonderzoek. Zij bestudeerden de resultaten van dertig lichamelijk gehandicapte kinderen van een revalidatiecentrum, getest met dezelfde vertaalde WISC-R en gemiddeld drie jaar later met de WISC-RN. Zowel performaal als verbaal bleken lagere IQ-scores met de "nieuwe" test, vooral performaal verschenen grote verschillen.

Lally e.a. (1987) stellen dat de wet in de USA zegt dat herhalingsonderzoek bij kinderen in het buitengewoon onderwijs om de drie jaar noodzakelijk is, zolang zij "zorgen" ontvangen. De auteurs houden een pleidooi, op basis van hun onderzoek, voor het niet-systematisch uitvoeren van herhalingsonderzoek. Als bij herhaalde afname de scores van de WISC-R twee keer binnen de grenzen van de standaardmeetfout vallen, is de waarschijnlijkheid dat een derde afname

significant zal verschillen erg laag. Zij concluderen dat kinderen met leerproblemen vrij stabiele IQ's behalen zodat "automatisch" herhaalde cognitieve testing niet nodig blijkt.

5 Parallelversies van een test en paralleltests

Kievit e.a. (1992) vermelden dat men in sommige gevallen gebruik kan maken van een parallelversie van een test. Zo'n test heeft precies dezelfde statistische eigenschappen maar bevat andere items, zodat het leereffect voor een belangrijk deel wordt omzeild.

Voor de Wechsler-instrumenten bestaan er echter geen parallelversies. Voor de WISC-RN wordt in dit handboek (Kievit e.a., 1992) wel verwezen naar een paralleltest, met name de RAKIT (Revisie Amsterdamse Kinder Intelligentie Test). Om de IQ-scores van de twee tests met elkaar te vergelijken, heeft men equivalentietabellen ontwikkeld waarmee het WISC-RN-IQ in een RAKIT-IQ kan worden omgezet (en omgekeerd). In het algemeen kan men stellen dat bij hertesten met een paralleltest men best rekening houdt met de eventuele verschillen qua meetschaal (zie ook Scheiris, 1999) maar ook met de correlatie tussen beide tests en het betrouwbaarheidsinterval rond de bekomen scores.

Luteyn e.a. (1990) vermelden in hun handboek dat IQ's bepaald met verschillende tests zeker niet elkaars equivalent zijn. Uit onderzoek naar de relatie tussen de GIT (Groninger Intelligentie Test), de WAIS en de SPM (Progressieve Matrices van Raven) bij neurologische patiënten bleken voorspellingen van het ene IQ bij ongeveer 38 % van de patiënten verschillen groter dan 10 IQ- punten op te leveren.

Bolen (1998a) stelt dat de clinicus kan observeren, wanneer hij hertest met een paralleltest, dat de resultaten niet overeenstemmen met vorige resultaten. De cruciale vraag is hoe verschillend de testresultaten moeten zijn om significant te kunnen worden genoemd. Sattler (1982) suggereert als algemene regel dat een verschil gelijk of groter dan één standaarddeviatie tussen de twee tests als significant wordt aanzien. Dit is in het bijzonder van toepassing als beide tests een hoge betrouwbaarheid hebben en daarbij horend kleine standaardmeetfouten. Er zijn echter meerdere redenen waarom hoog betrouwbare tests die algemene capaciteiten meten toch discrepante resultaten kunnen teweegbrengen. Bracken (1988) schreef hieromtrent het artikel *Ten Psychometric Reasons Why Similar Tests Produce Dissimilar Results*, met verwijzingen naar floor- and ceiling-effecten, regressie naar het gemiddelde, verschillen in normtabellen, enz.

Ten slotte stelde men in de reeds hoger vermelde studie van Legagnoux e.a. (1990) vast dat de vooruitgang in IQ bij hertesten veel hoger bleek in het geval de posttest gelijk was aan de pretest dan in het geval van een tweede meting met een parallelversie van een test. Samen met hen pleiten wij ervoor dat psychometristen meer aandacht zouden hebben voor het ontwikkelen

van parallelversies en de eventuele hertesteffecten klaar en duidelijk in de handleiding zouden vermelden.

Conclusies

1. Ondanks het feit dat intellectuele capaciteiten geen gefixeerd aspect betreffen, is er toch voldoende bewijs van een consistent IQ over langere periodes, ook bij speciale populaties.
2. Bij "foute" medische of psychopathologische diagnoses, uitgesproken motivatie-problemen, emotionele druk, vermoeidheid,… tijdens een eerste testafname, kan hertesten zinvol zijn, maar het klinisch oordeel blijft de basis voor het beslissen tot hertesten.
3. Bij het op korte termijn hertesten is minstens een leereffect te verwachten. Het is dan vooral het PIQ dat wordt geflatteerd.
4. Wacht indien mogelijk minstens één jaar, bij voorkeur twee jaar voor het hertesten met eenzelfde Wechsler-instrument.
5. Als het interval korter moet: kijk na wat men in de handleiding zegt, en ga na of er een parallelversie of paralleltest voorhanden is.

Op de hoofdvraag verwachtten wij een meer wetenschappelijk onderbouwd antwoord dan bij aanvang van deze literatuurstudie. Er blijkt echter nog een duidelijke nood aan extra onderzoek op het gebied van hertesten van intelligentie. We denken hierbij aan specifieke studies over de hertesteffecten, eventueel experimenteeel geïnspireerd, en over parallelversies en –instrumenten. De ontwikkeling van deze laatste zijn ondermaats. In onze literatuurstudie bleek een ander opvallend onderbelicht aspect, met name de overgangen tussen de kleuter-, kind- en volwassenen-meting van intelligentie.

Via initiatieven zoals het pas opgerichte Vlaams Forum voor Diagnostiek, waarvan de auteur en de inspirator van dit artikel zelf deel uitmaken, hopen wij in de nabije toekomst enige van de gemelde tekorten te kunnen ondervangen.

Literatuur

Anastasi, A., Urbina, S. (1997). *Psychological Testing*. New Jersey: Printice Hall.

Arbuckle, T.Y., Maag, U., Pushkar, D., Chaikelson, J.S. (1998). Individual differences in trajectory of intellectual development over 45 years of adulthood. *Psychology and Aging,* 4, 663-675.

Bolen, L.M. (1998a). Assessing intelligence using the WISCIII. In: H.B. Vance (ed.), *Psychological assessment of children.* New York: John Wiley.

Bolen, L.M. (1998b). WISC-III score changes for EMH students. *Psychology in the schools,* 4, 327-332.

Canivez, G.L., Watkins, M.W. (1998). Long-term stability of the Wechsler Intelligence Scale for Children – Third edition. *Psychological Assessment,* 10, 285-291.

Dekker, R., Mackaay-Craamer, E.M., De Bruyn, E.E.J. (1990). De Nederlandse WISCR bij kinderen in het speciaal onderwijs. *Kind en Adolescent,* 4, 170-179.

De Zeeuw, J. (1978). *Algemene psychodiagnostiek II. Testtheorie*. Lisse: Swets & Zeitlinger.

De Zeeuw, J. (1996). *Algemene psychodiagnostiek I. Testmethoden* Lisse: Swets & Zeitlinger.

De Zeeuw, J. (1996). *Inleiding in de psychodiagnostiek*. Lisse: Swets & Zeitlinger.

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

Gold, D.P., Andres, D., Etezadi, J., Arbuckle, T.Y., Schwartzman, A.E., Chaikelson, J. (1995). Structural equation model of intellectual change and continuity and predictors of intelligence in older men. *Psychology and Aging,* 10, 294-303.

Hawkins, A., Sayward, H.K. (1994). Examiner judgment and actual stability of psychiatric inpatient intelligence quotients. *The Clinical Neuropsychologist,* 4, 394-404.

Heesels, N., Meijs, M.J.M., Feltzer, M.J.A., Eilander, H.J. (1992). Herhalingsonderzoek met de WISC-R bij kinderen met een lichamelijke handicap. *Kind en Adolescent,*3, 144-147.

Janda, L.H. (1998). *Psychological testing. Theory and applications.* Needham Heights, Massachusetts: Allyn and Bacon.

Juliano, J.M., Haddad, F.A., Carroll, J.L. (1988). Three year stability of WISC-R factor scores for Black an White, female and male children classified as learning disabled. *Journal of School Psychology*, 26, 317-325.

Kievit, T., de Wit, J., Groenendaal, J.H.A., Tak, J.A. (1992). *Handboek psychodiagnostiek voor de hulpverlening aan kinderen.* Leuven: Acco.

Klauer, K.J. (1993). Learning potential testing: the effect of retesting. In: J.H.M. Hamers, K. Sijtsma, A.

Kline, P. (2000*). Handbook of psychological testing.* New York: Routledge.

Ruijssenaars (eds.), *Learning potential assessment: theoretical, methodological and practical issues.* Amsterdam: Swets and Zeitlinger.

Lally, M.J., Lloyd, R.D., Kulberg, J.M. (1987). Is intelligence stable in learning disabled children? *Journal of psychoeducational Assessment,* 4, 411-416.

Legagnoux, G., Michael, W.B., Hovecar, D., Maxwell, V. (1990). Retest effect on standardized structure of intellect ability measures for a sample of elementary school children. *Educational and Psychological Measurement,* 50, 475-492.

Luteijn, F., Deelman, B.G., Emmelkamp, P.M.G. (1990). *Diagnostiek in de klinische psychologie.* Houten: Bohn Stafleu Van Loghum.

Murphy, K. R., Davidshofer, C.O. (1998). *Psychological Testing*. New Jersey: Printice Hall.

Neyens, L., Aldenkamp, A. (1999). Stabiliteit van intelligentiescores en neuropsychologische maten bij kinderen met een ten minste gemiddeld intelligentieniveau. *Nederlands Tijdschrift voor de Psychologie,* 54, 38-46.

Salvia, J., Ysseldyke, J.E. (1991). *Assessment* (5th ed.). Boston: Houghton Mifflin.

Sarazin, F.F., Spreen, O. (1986). Fifteen-Year stability of some neuropsychological tests in learning disabled subjects with and without neurological impairment. *Journal of Clinical and Experimental Neuropsychology,* 3, 190-200.

Snow, W., Tierney, M., Zorzitto, M., Fisher, R., Reid, D. (1989). Wais-R test-retest reliability in a normal elderly sample. *Journal of Clinical and Experimental Neuropsychology,* 4, 423-428.

Scheiris, J. (1999). Kwalitatieve en kwantitatieve diagnostiek. *SIGnaal,* 8 (26), 3-14.

Schittekatte, M. (2000). De Commissie TestAaangelegenheden Nederland. *BFP info,* 15-1, 15.

Schuerger, J.M., Witt, A.C. (1989). The temporal stability of individually tested intelligence. *Journal of Clinical Psychology,* 45, 294-302.

Vance, H.R., Brown, W., Hankins, N. (1987). A comparison of the WISC-R and the WAIS-R with special education students. *Journal of Clinical Psychology,*3, 377-380.

Van der Doef, M.P., Kwint, J.M., van de Koppel, J.M.H. (1989). *Kind en Adolescent;*3, 136-141.

Wechsler, D. (1991). *WISC-III. Manual.* New York: The Psychological Corporation.

Wechsler, D. (1996*). Manuel de l'échelle d'intelligence de Wechsler pour enfants, troisième édition*. Paris: Editions du Centre de Psychologie Appliquée.

Zimmerman, I.L., Woo-Sam, J.M. (1996). Is retesting with the WISC-III a defensible procedure? *Perceptual and Motor Skills,* 82, 349-350.