

## 10 VEROUDERING VAN DE TESTNORMEN

Een belangrijk, en voor de diagnostiek uitermate lastig probleem, is de veroudering van testnormen. De prestatie op intelligentietests van personen van dezelfde leeftijd neemt in de Westerse landen per tien jaar met ongeveer drie IQ-punten toe (Lynn & Hampson, 1986; Flynn, 1987, 2009). Per land en per type intelligentietest kunnen deze uitkomsten echter verschillen. Door het beter worden van de prestaties zullen bij een nieuwe test de normen 'strenger' worden en zullen de scores lager uitval- len dan op tests die langer geleden zijn genormeerd. Een dergelijk effect werd in Nederland onder meer geconstateerd bij de herziening van de WISC-R (Harinck & Schoorl, 1987), de SON-R 5½-17 (Snijders, Tellegen & Laros, 1988), de SON-R 2½-7 (Tellegen, e.a. 1998), bij de LDT (Resing & Nijland, 2002) en bij de herziening van de WISC-III-NL (Wechsler, 2005a). Bij de Amerikaanse vergelijking van de WISC-III met de WISC-R (Wechsler, 1991) en van de WPPSI-R met de WPPSI (Wechsler, 1989), bleek de toename per tien jaar (gemiddeld over beide tests) voor het *Full Scale IQ* (FSIQ) 3.4 punten, voor het *Performance IQ* (PIQ) 4.3 punten en voor het *Verbal IQ* (VIQ) 1.9 punt.

In de handleiding van de SON-R 2½-7 is een tabel opgenomen waarin per jaar het effect van de veroudering van de normen is aangegeven. Sinds 2005 is deze correctie voor de SON-tests in het computerprogramma geïmplementeerd en wordt het gecorrigeerde IQ als IQ\* weergegeven. Hierbij wordt uitgegaan van een verouderingseffect van de normen van één IQ-punt per drie jaar, ofwel tien punten per dertig jaar. Hoewel deze schatting aansluit bij hetgeen in de literatuur wordt ver- meld, bestaat er echter ook gerede twijfel of het Flynn-effect wel zo constant is en of er in de loop der jaren niet van een vermindering sprake is (Sundet, Barlaug & Torjussen, 2004). Anderzijds zijn er aanwijzingen dat het Flynn-effect juist sterker is bij tests die gerelateerd zijn aan *fluid intelligence*, zoals de SON-tests. Een juiste inschatting van de grootte van het effect is uitermate belangrijk om de correctie goed toe te kunnen passen. In de volgende paragraaf zullen eerst een aantal aanwijzin- gen, afkomstig uit Nederlands/Vlaams onderzoek, worden besproken dat het Flynn-effect in Neder- land minder sterk is dan door ons was aangenomen. Daarna worden de resultaten gepresenteerd van een onderzoek dat speciaal is opgezet om de sterkte van het Flynn-effect bij de SON-tests te bepalen. Bij dit onderzoek zijn bij honderd leerlingen van een basisschool zowel de SON-R 6-40 als de corresponderende subtests van de SON-R 5½-17 afgenomen.

### 10.1 AANWIJZINGEN VOOR EEN MINDER STERK FLYNN-EFFECT

Drie verschillende onderzoeken wijzen in de richting dat een schatting van het Flynn-effect voor de SON-tests van drie IQ-punten per tien jaar te hoog is. De eerste indicatie is gebaseerd op de verge- lijking tussen de IQ-scores van de SON-R 2½-7 en de IQ-scores op andere tests die in Nederland bij dezelfde kinderen zijn afgenomen. De tweede aanwijzing volgt uit het normeringsonderzoek met de WISC-III-NL waarbij de scores met de WISC-R zijn vergeleken en de derde aanwijzing volgt uit het vergelijkend onderzoek van de WNV met de SON-R 2½-7 en de SON-R 5½-17.

### Vergelijking IQ van de SON-R 2½-7 en scores op andere tests

In de handleiding van de SON-R 2½-7 is een vergelijking gemaakt van het SON-IQ met de gemiddelde scores op andere tests met Nederlandse normen (Tellegen, e.a. 1998, paragraaf 9.10). Hierbij is een correctie voor het Flynn-effect toegepast waarvan de grootte gebaseerd is op de uitkomsten van het Amerikaanse onderzoek met de Wechsler-tests. Voor de algemene scores bedroeg het effect 3.4 IQ-punten per tien jaar en voor de performale/non-verbale scores 4.3 IQ punten. Zonder correctie was het SON-IQ lager dan van de andere tests die jaren eerder waren genormeerd. Dit is in overeenstemming met hetgeen op grond van het Flynn-effect wordt verwacht. Gemiddeld over negen vergelijkingen bedroeg het verschil -2.4 IQ-punten. Nadat de correctie voor het Flynn-effect was toegepast lag het verschil in gemiddelden echter niet dicht bij nul maar was het bij zeven van de negen vergelijkingen duidelijk positief met een gemiddelde van 2.8. Dit betekent dat de gemiddelde correctie van 3.8 punten die was gehanteerd te sterk was. De correctie zou 1.75 IQ-punt per tien jaar geweest moeten zijn om de verschillen tussen de testgemiddelden te compenseren.

### Vergelijking WISC-III-NL en WISC-R-NL

In de handleiding en verantwoording van de WISC-III NL (Wechsler, 2005a, hoofdstuk 4) wordt een onderzoek beschreven dat in Vlaanderen is uitgevoerd en betrekking heeft op de testcores van de WISC-III en van de WISC-R die eerder was afgenomen. Deze gegevens waren bij 308 kinderen verzameld door revalidatiecentra en centra voor leerlingen begeleiding. De periode tussen de afname van beide tests bedroeg gemiddeld drie jaar. De gemiddelde score voor het Totaal IQ (TIQ) van de WISC-III was 79.1 met een standaarddeviatie van 8.1. De periode tussen de normeringen van de WISC-R en de WISC-III bedroeg ongeveer twintig jaar zodat bij een Flynn-effect van drie punten per tien jaar een verschil van zes IQ-punten verwacht kon worden tussen het TIQ van de WISC-R en het TIQ van de WISC-III. Het verschil bedroeg echter slechts 2.8 punten hetgeen correspondeert met een Flynn-effect van 1.4 IQ-punt per tien jaar. Opmerkelijk was verder dat het verschil voor het Verbaal IQ (VIQ) 4.9 punten bedroeg en het verschil voor het Performaal IQ (PIQ) zeer klein was (.30) en bovendien tegengesteld aan de verwachte richting. In het algemeen wordt aangenomen dat het Flynn-effect juist sterk is voor tests die *fluid intelligence* en het performaal IQ meten en veel minder sterk voor *crystallized intelligence* en het verbale deel van de Wechsler tests.

### Vergelijking tussen de WNV en de SON-R 2½-7 en de SON-R 5½-17

In de handleiding van de WNV worden een aantal vergelijkende onderzoeken tussen de WNV en de SON-tests beschreven die in het kader van het normerings- en valideringsonderzoek van de WNV zijn uitgevoerd (Wechsler & Naglieri, 2008, hoofdstuk 5). Hierbij wordt niet ingegaan op de verschillen in gemiddelde scores maar vanuit het gezichtspunt van het Flynn-effect zijn deze juist interessant. Bij benadering is het jaar van de normering voor de WNV 2007, voor de SON-R 2½-7 1993 en voor de SON-R 5½-17 1984. Dit betekent dat bij een Flynn-effect van drie punten per tien jaar de scores op de WNV gemiddeld ruim vier punten lager zullen zijn dan op de SON-R 2½-7 en gemiddeld bijna zeven punten lager dan op de SON-R 5½-17.

Bij het normeringsonderzoek van de WNV is in wisselende volgorde bij 47 kinderen ook de SON-R 2½-7 afgenomen (gemiddelde leeftijd 5.4 jaar). De gemiddelde score op de WNV bedroeg 96.1 en het gemiddelde SON-IQ was 97.7. Het verschil in scores van 1.6 punten is dus minder dan de helft van het te verwachten verschil.

Bij zeventig kinderen van het normeringsonderzoek van de WNV is ook de SON-R 5½-17 afgenomen (gemiddelde leeftijd 14.2 jaar). De gemiddelde WNV-score was 102.7 en het gemiddelde SON-IQ dat gebaseerd was op een normering van 23 jaar eerder was niet zeven IQ-punten hoger maar met een gemiddelde van 98.7 juist vier punten lager. Bij een aantal speciale groepen zijn bij 97 kinderen in de leeftijd vanaf acht jaar aanvullende testgegevens verzameld die gemiddeld drie jaar eerder

waren verzameld. De andere test betrof voornamelijk de SON-R 5½-17. Bij deze speciale groepen was de score op de andere test met een gemiddeld IQ van 76.2 wel iets hoger dan het gemiddeld IQ van 74.1 op de WNV. Indien we deze gegevens combineren met de groep van zeventig kinderen uit het normeringsonderzoek dan is het resultaat dat bij 167 personen het gemiddelde WNV-IQ gelijk is aan 86.1 en het gemiddeld SON-IQ met 85.1 een halve punt lager. In een periode van 23 jaar zou er geen sprake zijn van enig Flynn-effect.

## 10.2 VERGELIJKINGSONDERZOEK TUSSEN DE SON-R 6-40 EN SON-R 5½-17

De bovengenoemde onderzoeken geven aanwijzingen voor een verminderd Flynn-effect maar het is moeilijk om op grond van deze uitkomsten tot een goede nieuwe schatting te komen die voor de SON-tests toepasbaar is. Voor een deel komt dit door de methodologische problemen bij de verschillende onderzoeken. Zo was het bij de vergelijking met de SON-R 2½-7 niet altijd duidelijk wanneer de criteriumtests waren genormeerd; bij het onderzoek met de WISC-III speelden problemen met selectie van personen en volgorde van afname een rol en bij de WNV is er reden tot twijfel of de normen van de WNV wel juist zijn.

Omdat het voor de interpretatie van de uitkomsten op intelligentietests van groot belang is om op een juiste wijze met de veroudering van normen rekening te houden (Fletcher, Stuebing & Hughes, 2010), is na het normeringsonderzoek van de SON-R 6-40 nog een uitgebreid en methodologisch goed opgezet onderzoek uitgevoerd om de verschillen tussen de normen van de SON-R 6-40 en de SON-R 5½-17 te kunnen beoordelen.

### De opzet van het onderzoek

Op twee locaties van een basisschool in Groningen zijn bij honderd leerlingen van groep 5, 6 en 7 met een interval van enkele maanden in wisselende volgorde de SON-R 6-40 en de vier corresponderende subtests van de SON-R 5½-17 afgenomen (zie Tabel 10.1). Per groep waren de kinderen door de leerkracht op een drie-puntsschaal beoordeeld op intelligentie en binnen elke categorie werd *at random* bepaald welke test het eerst zou worden afgenomen. Dit garandeerde een gelijke verdeling binnen de groepen en binnen de intelligentieniveaus. Bij vijftig leerlingen is eerst de SON-R 6-40 afgenomen en bij de andere vijftig leerlingen eerst een verkorte versie van de SON-R 5½-17 bestaande uit Analogieën, Mozaïeken, Categorieën en Patronen. De eerste afname vond plaats in maart 2011 en de tweede afname in mei of juni van dat jaar. Gemiddeld bedroeg de periode tussen de afname's tweeënhalve maand. De tests zijn door drie testleiders afgenomen waarbij ieder kind beide keren door dezelfde testleider werd onderzocht (Matthijssen & Geertsema, 2011).

De honderd leerlingen waren afkomstig uit vier klassen die bijna integraal aan het onderzoek hebben deelgenomen. Enkele ouders wilden geen toestemming geven, maar een weigering door leerlingen kwam niet voor. De gemiddelde leeftijd bij de eerste afname was 10.3 jaar met een standaarddeviatie van 1.0 jaar. Van 82 van de 100 leerlingen zijn achtergrondgegevens bekend. Hieruit blijkt dat het percentage allochtonen (tenminste één ouder buitenlands) met 27% enigszins hoger

**Tabel 10.1** Samenstelling van de onderzoeksgroep (N=100) met betrekking tot het Flynn-effect

Sekse	Groep	Leeftijd bij eerste afname	
54% man	23% groep 5	15% 8 jaar	16% 11 jaar
46% vrouw	26% groep 6	23% 9 jaar	3% 12 jaar
	51% groep 7	43% 10 jaar	

**Tabel 10.2** SON-IQ per testversie en per afname (N=100)

Volgorde	Eerste testafname	Tweede testafname
conditie 1 (N=50)	SON-R 6-40 (gem.=103.0; sd=16.5)	SON-R 5½-17 (gem.=116.0; sd=17.7)
conditie 2 (N=50)	SON-R 5½-17 (gem.=107.8; sd=15.2)	SON-R 6-40 (gem.=110.3; sd=16.2)
gemiddelde tweede afname = 113.1 (N=100)		gemiddelde SON-R 5½-17 = 111.9 (N=100)
gemiddelde eerste afname = 105.4 (N=100)		gemiddelde SON-R 6-40 = 106.6 (N=100)
verschil (leereffect) = 7.7 (N=100)		verschil (testversie-effect) = 5.3 (N=100)

**Tabel 10.3** Variantie-analyse van de SON-scores met testversie en afnamevolgorde als onafhankelijke variabelen (N=200)

	testversie			afnamevolgorde			interactie onafhankelijke variabelen		
	F	p	effect	F	p	effect	F	p	effect
Analogieën	3.79	.053	.858	7.29	.008	1.190	.23	.634	.210
Mozaïeken	2.01	.158	.692	6.87	.009	1.280	.03	.870	-.080
Categorieën	10.13	.002	1.392	8.46	.004	1.272	.04	.834	-.092
Patronen	4.52	.035	.980	7.42	.007	1.256	.81	.368	-.416
SON-IQ	5.16	.024	5.280	10.98	.001	7.700	.04	.837	-.480

is dan het landelijke percentage van 20% en dat het opleidingsniveau eveneens iets hoger is (66.5% van de ouders heeft tenminste havo/vwo niveau versus 50% in het normeringsonderzoek).

### De resultaten

In Tabel 10.2 zijn de gemiddelde IQ-scores van de SON-R 6-40 en de SON-R 5½-17 weergegeven per conditie van volgorde van afname. Gemiddeld over beide tests is het verschil tussen de eerste en de tweede afname 7.7 IQ-punten. Dit is het leereffect dat vergelijkbaar is met een hertest-effect indien dezelfde test twee keer wordt afgenomen.

De scores op de SON-R 6-40 en de SON-R 5½-17, gemiddeld over de afnamevolgordes, verschillen 5.3 punten. Zoals verwacht is het gemiddelde van de SON-R 6-40 lager en het verschil kan beschouwd worden als indicatie van het Flynn-effect.

Met een variantie-analyse met de test scores als afhankelijke variabelen en testversie en afnamevolgorde als hoofdeffecten is voor de genormeerde subtest scores en voor de IQ-score getoetst of de verschillen als gevolg van afnamevolgorde en testversie significant zijn en of er sprake is van interactie-effecten. Met een regressie-analyse met testversie, volgorde, en eerst afgenomen test als dummy-variabelen is de grootte van de verschillende effecten geschat. De N bedroeg bij deze analyses 200 (Tabel 10.3).

Geen van de interactie-effecten was significant. Het hoofdeffect effect voor volgorde was voor alle subtests en de totaalscore zeer significant. De effecten voor de subtests verschillen weinig van elkaar en variëren van 1.2 tot 1.3. Het leereffect voor de totaalscore is 7.7.

De effecten van de testversie die is afgenomen, zijn in verhouding minder groot en variëren voor de subtests van .7 (Mozaïeken) tot 1.4 (Categorieën). Voor de IQ-score is het verschil tussen de SON-R

**Tabel 10.4** Gepaarde t-toetsen tussen de testcores van beide testversies, na correctie afnamevolgorde (N=100)

testscore	SON-R 5½-17		SON-R 6-40		verschil		95%-interval	t	p	r
	gem.	(sd)	gem.	(sd)	gem.	(sd)				
Analogieën	11.4	(3.1)	10.5	(3.1)	.86	(2.4)	.38 - 1.33	3.6	.001	.70
Mozaïeken	11.2	(3.5)	10.5	(3.3)	.69	(2.0)	.29 - 1.09	3.5	.001	.83
Categorieën	12.4	(3.1)	11.0	(3.0)	1.39	(2.7)	.85 - 1.93	5.1	.000	.61
Patronen	10.7	(3.7)	9.7	(2.8)	.98	(2.0)	.58 - 1.38	4.9	.000	.84
SON-IQ	108.1	(16.4)	102.8	(16.3)	5.28	(8.2)	3.64 - 6.92	6.4	.000	.87

6-40 en de SON-R 5½-17 gelijk aan 5.3 IQ-punten met een significantie van  $p=.024$ . Voor enkele subtests zijn de effecten niet significant maar deze methode van variantie-analyse heeft relatief weinig power omdat de scores vertekend worden door de afnamevolgorde en omdat geen gebruik wordt gemaakt van het feit dat het hier de gepaarde waarnemingen betreft.

Voor de beoordeling van de grootte van het effect van de testversie, en daarmee van het Flynn-effect zijn toetsen voor gepaarde waarnemingen uitgevoerd waarbij de scores bij de tweede afname gecorrigeerd voor het effect van afnamevolgorde door van de scores bij de tweede afname dit effect (onafhankelijk van de testversie) van de scores af te trekken (bij de IQ-scores is dit dus 7.7 punten indien een test als tweede is afgenomen). Alle verschillen tussen de gemiddelden voor de testversies zijn met de t-toets zeer significant (Tabel 10.4). Hoewel er tussen de subtests verschillen zijn tussen de gemiddelde afwijkingen, blijken de intervallen voor het verschil elkaar voor een belangrijk deel te overlappen. Daarom kunnen geen uitspraken worden gedaan of de sterkte van het Flynn-effect voor de subtests verschillend is.

Het 95%-betrouwbaarheidsinterval voor de verschillscore van het SON-IQ varieert van 3.6 tot 6.9. Tussen de normering van de SON-R 5½-17 (in 1984) en de normering van de SON-R 6-40 (in 2010) ligt een periode van 26 jaar. Dit betekent dat het gemiddelde verschil van 5.28 punten correspondeert met een verandering van 2.0 punten per tien jaar. De grenzen van het 95%-interval corresponderen met een verschil dat varieert van 1.4 tot 2.7 punten per tien jaar.

### 10.3 CONCLUSIES

In het normeringsonderzoek van de SON-R 2½-7 is gebleken dat een correctie van 1.75 IQ-punt per tien jaar voor veroudering van de normen van een aantal criterium tests leidde tot overeenkomstige gemiddelde scores. Deze correctie komt goed overeen met de correctie van twee IQ-punten per tien jaar die als meest aannemelijke waarde uit het vergelijkend onderzoek tussen de SON-R 6-40 en de SON-R 5½-17 is gekomen.

Uit het Vlaamse onderzoek met de WISC-III en de WISC-R kwam een kleiner verschil van 1.4 IQ-punt per tien jaar maar de auteurs wijzen erop dat het feit dat de WISC-III in dit onderzoek steeds als tweede test is afgenomen mogelijk heeft bijgedragen tot het relatief kleine verschil.

Het ontbreken van een duidelijk verschil in gemiddelde scores tussen de WNV en de SON-R 5½-17 is opmerkelijk. Gezien de verschillen tussen de normen van de WISC-R en de WISC-III en tussen de normen van de SON-R 5½-17 en de SON-R 6-40 is het niet aannemelijk dat in de periode van 23 jaar tussen de normering van de SON-R 5½-17 en de SON-R 6-40 geen Flynn-effect zou hebben plaatsgevonden. Mogelijk zijn de genormeerde scores van de WNV systematisch te hoog en verklaart dit waarom er geen verschil is met de normen van de SON-R 5½-17. Indien het Flynn-effect in deze

**Tabel 10.5** Veroudering van de normen van het SON-IQ

SON-R 2½-7	SON-R 5½-17	SON-R 6-40	
2011 - 2015: 4 IQ-punten	2011: 4 IQ-punten	2011 - 2012: 0 IQ-punt	2023 - 2027: 3 IQ-punten
2016 - 2020: 5 IQ-punten	2012 - 2014: 5 IQ-punten	2013 - 2017: 1 IQ-punt	2028 - 2032: 4 IQ-punten
2021 - 2025: 6 IQ-punten	2015 - 2019: 6 IQ-punten	2018 - 2022: 2 IQ-punten	

De veroudering is gebaseerd op een Flynn-effect van twee IQ-punten per tien jaar

periode twee IQ-punten zou bedragen dan zouden de scores van de WNV gemiddeld 4.6 punten lager moeten zijn dan van de SON-R 5½-17. Ze zijn echter een halve punt hoger, hetgeen betekent dat de normen van de WNV in feite 5.1 punt te hoog zijn (hierbij worden mogelijke toevalsfluctuaties buiten beschouwing gelaten). In het kader van het normeringsonderzoek van de SON-R 6-40 is bij 55 personen de WNV afgenomen. Tussen de normeringen zit drie jaar, wat bij een Flynn-effect van twee IQ-punten per tien jaar tot een verschil van .6 IQ-punten zou moeten resulteren. De gemiddelde score op de WNV is echter 4.9 IQ-punten hoger hetgeen een indicatie is voor een systematische vertekening van de WNV-normen van 4.3 IQ-punt.

De aanwijzingen dat de normen van de WNV tot te hoge scores leiden hebben niet alleen betrekking op de vergelijkingen met de SON-tests. In de handleiding worden ook vergelijkingen gemaakt met de WISC-III-NL en met de WAIS-III-NL. Aangezien de normen van deze tests herzien zijn is het niet duidelijk wat de gemiddelde testdatum is geweest bij het normeringsonderzoek. Een gemiddeld verschil van één IQ-punt op grond van veroudering van de normen lijkt bij een verondersteld Flynn-effect een redelijke schatting ten opzichte van de normering van de WNV. De WISC-III-NL is in combinatie met de WNV bij 51 personen afgenomen. De verwachting is dat op grond van de veroudering van de normen van de WISC-III-NL de scores op de WNV één IQ-punt lager zullen zijn. De gemiddelde scores op de WNV zijn echter .4 hoger. Bij de vergelijking met de WAIS-III-NL die bij 45 personen in combinatie met de WNV is afgenomen is de gemiddelde score op de WNV 4.3 punten hoger. Gemiddeld over beide tests wijst dit op een systematische afwijking bij de WNV van vier IQ-punten. Bij een systematische vertekening van het WNV-IQ van ongeveer vier IQ-punten kan een Flynn-effect van twee IQ-punten per tien jaar de gevonden verschillen tussen de gemiddelde SON-scores en de WNV-scores goed verklaren.

Op grond van de uitkomsten van het door ons uitgevoerde onderzoek en van de analyses van de andere onderzoeken is de keuze gemaakt om met ingang van 1 september 2011, de datum van uitgave van de SON-R 6-40, bij de schatting van het Flynn-effect niet langer uit te gaan van één IQ-punt per drie jaar (3.3 punten per tien jaar) maar van één IQ-punt per vijf jaar (2.0 punten per tien jaar). In de nieuwe versie 5.1 van het computerprogramma van de SON-R tests dat tegelijkertijd zal verschijnen wordt deze verandering voor de berekening van IQ\* (het voor het Flynn-effect gecorrigeerde IQ) doorgevoerd voor de SON-R 2½-7, de SON-R 5½-17 en voor de SON-R 6-40. In Tabel 10.5 is weergegeven hoe groot deze correctie de komende jaren is voor de verschillende versies van de SON. Een minder sterk Flynn-effect dan in eerste instantie werd aangenomen is goed nieuws voor testgebruikers. Het betekent namelijk een verbetering van de onderlinge vergelijkbaarheid van IQ-scores op verschillende tests.