HET BETROUWBAARHEIDSINTERVAL IN INTELLIGENTIETESTS

KIMBERLEY LEK, WENNEKE VAN DE SCHOOT-HUBEEK, EVELYN KROESBERGEN & RENS VAN DE SCHOOT

Het 95% betrouwbaarheidsinterval in de verslaglegging van intelligentietests kan grofweg geïnterpreteerd worden als een soort 'foutenmarge' rondom de verkregen IQ-score van het kind. De exacte interpretatie ligt ingewikkelder. Genoemd interval staat er dan ook om bekend vaak verkeerd geïnterpreteerd te worden. Daarom bieden Kimberley Lek en collega's een 'opfrisser': hoe wordt een betrouwbaarheidsinterval ook al weer opgesteld ? De WISC-III NL wordt gebruikt ter illustratie.

HOE ZAT HET OOK ALWEER?

HET BETROUWBAAR-HEIDSINTERVAL IN INTELLIGENTIETESTS

De IQ-score die een kind[1] op een bepaald moment haalt, hangt af van allerlei factoren. Zo kunnen IQ-scores beïnvloed zijn door werkhoudings- en/of aandachtsproblemen (Pameijer, 2014), testleider-effecten, zoek- en rekenfouten bij de verwerking, storende factoren bij de afname (Tellegen, 2004), vermoeidheid, stemming, faalangst (Schouws, 2015) et cetera. Daarom kan de IQ-score niet té strikt geïnterpreteerd worden (Tellegen, 2004). Om al te absolute interpretatie tegen te gaan, is het raadzaam gebruik te maken van het bijbehorende betrouwbaarheidsinterval; een soort foutenmarge. Het interpreteren van dit betrouwbaarheidsinterval blijkt echter niet zo makkelijk te zijn.

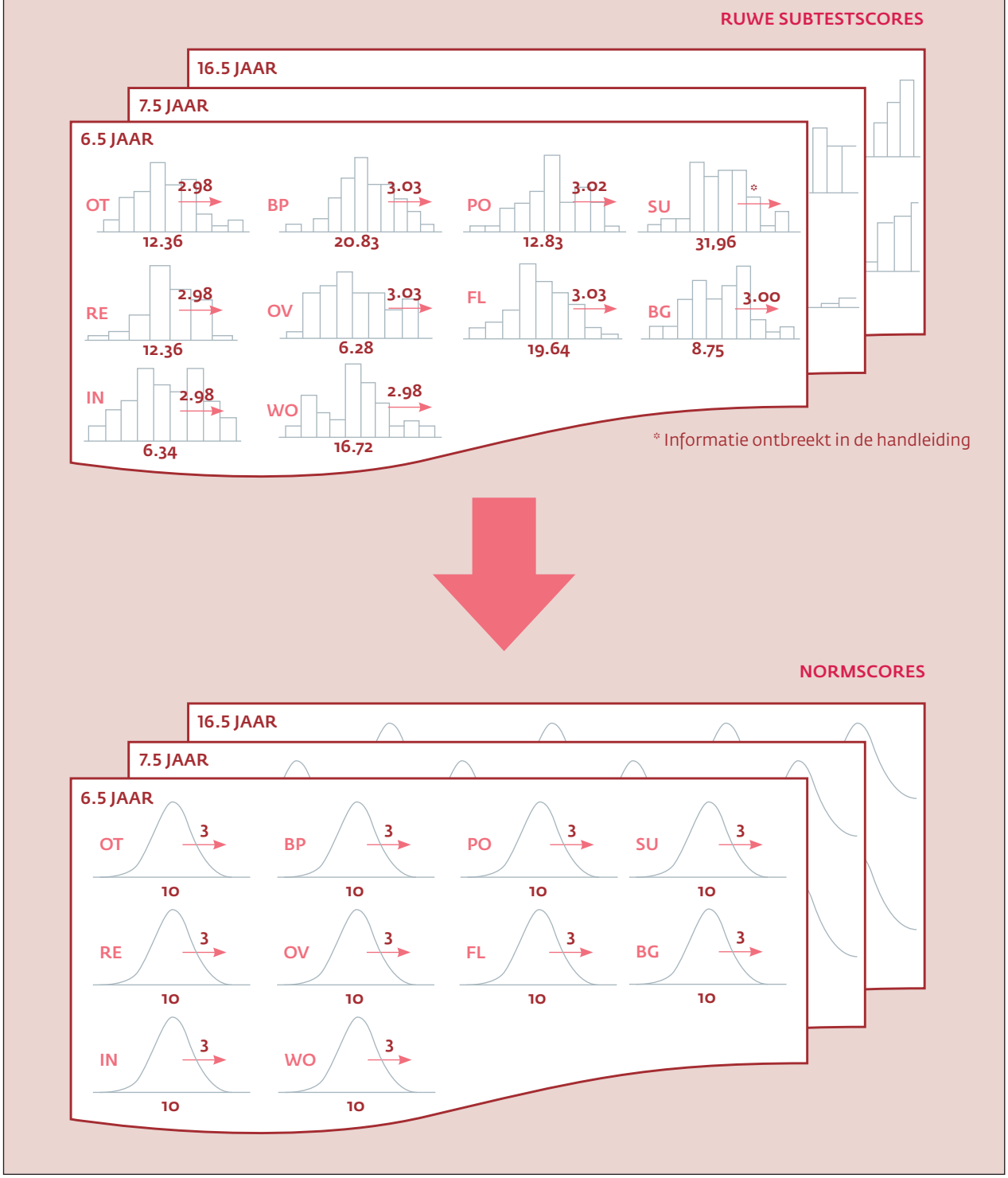
Stel, bijvoorbeeld, dat u net een WISC-III^{NL} afgenomen heeft bij Julia. Zij heeft een (T)IQ-score behaald van 97, met bijbehorend 95% betrouwbaarheidsinterval lopend van 90 tot 104. Hoe zou u dit betrouwbaarheidsinterval interpreteren? Eind 2014 hebben wij deze vraag gesteld aan 293 psychologen verbonden aan het NIP. Analyse van de reacties toonde een variëteit aan van mogelijke interpretaties van het 95% betrouwbaarheidsinterval. Dit is niet zo vreemd, want de handleiding van de WISC-III^{NL} legt niets uit over de totstandkoming van de intervallen en een concrete interpretatie wordt ook niet gegeven (Tellegen, 2002). Daarnaast staan betrouwbaarheidsintervallen erom bekend vaak verkeerd geïnterpreteerd te worden (bijv. Hoekstra, Morey, Rouder et al., 2014).

Vorig jaar hebben wij een presentatie gegeven op een bijeenkomst van het NIP – sectie Schoolpsychologen (22 januari 2016) over betrouwbaarheidsintervallen in intelligentietests. Dit artikel borduurt voort op deze bijeenkomst en is bedoeld als opfrisser: hoe zit het ook alweer met de constructie van betrouwbaarheidsintervallen en de interpretatie ervan? We geven een uitleg van de stappen die nodig zijn om 95% betrouwbaarheidsintervallen te construeren. Deze constructie is hetzelfde voor het merendeel van de huidige intelligentietests en wordt besproken in de tweede sectie van dit artikel[2]. De eerste sectie staat stil bij de stappen voorafgaand aan de constructie van het betrouwbaarheidsinterval. Deze stappen zijn vergelijkbaar voor de meeste intelligentietests. Ter illustratie bespreken wij deze stappen in detail voor de WISC-III^{NL}. We vervolgen met de interpreta-

1 Of andere deelnemer aan een testonderzoek

2 Specifieker gezegd, alle intelligentietests gebaseerd op de Klassieke Test Theorie (toegelicht onder "Hoe kunnen WISC-III^{NL} -betrouwbaarheidsintervallen geïnterpreteerd worden?") construeren hun betrouwbaarheidsintervallen op eenzelfde wijze.

FIGUUR 1. DE WISC-III^{NL}-PROCEDURE VOOR HET VERTALEN VAN RUWE SUBTESTSCORES (BOVEN) NAAR NORMSCORES (ONDER)[3]



tie van het betrouwbaarheidsinterval, waarbij we stil staan bij de aannames die ten grondslag liggen aan het betrouwbaarheidsinterval en een vergelijking maken met veelvoorkomende (foutieve) interpretaties uit de survey van eind 2014. Het artikel eindigt met een praktische discussie: wat betekent dit artikel voor de 95% betrouwbaarheidsintervallen van tests en vragenlijsten die u als psycholoog gebruikt?

Als aanvulling op dit artikel hebben wij een interactieve applicatie ontwikkeld die de interpretatie van betrouwbaarheidsintervallen illustreert en (gratis) gebruikt kan worden voor onderwijsdoeleinden (zie https://osf.io/7gz4q/). Deze illustratie is niet beperkt tot de WISC-III^{NL}; ook andere intelligentietests kunnen als uitgangspunt genomen worden.

STAPPEN VOORAFGAAND AAN CONSTRUCTIE BETROUWBAARHEIDSINTERVAL (WISC-III^{NL})

VAN ITEMSCORE NAAR RUWE SUBTESTSCORE Allereerst worden de scores op items behorende bij een bepaalde subtest opgeteld om tot een ruwe subtestscore te komen (zie tabel 3.1, p. 45, WISC-III^{NL}-handleiding).

VAN RUWE SCORE NAAR NORMSCORE Vervolgens worden de ruwe subtestscores vertaald naar normscores om (1) de (ruwe) subtestscores onderling te kunnen vergelijken en (2) ervoor te zorgen dat een bepaalde subtestscore dezelfde *relatieve* betekenis heeft voor verschillende leeftijdsgroepen. Hiervoor wordt gebruikgemaakt van de verzamelde gegevens uit de normpopulatie, die bestaat uit 1,239 kinderen in de leeftijdscategorieën 6-16. De omzetting van ruwe scores naar normscores (met gemiddelde tien en standaarddeviatie drie) gebeurt voor elk van deze leeftijdscategorieën (zie de 'slides' in figuur 1) én iedere subtest (de tien verdelingen binnen de 'slides') afzonderlijk.

Allereerst wordt in elk van de leeftijdscategorieën en voor elk van de subtests bepaald wat de verdeling is van ruwe subtestscores die door de normpopulatie kinderen in die leeftijdscategorie zijn behaald. Dit is te zien in het bovenste gedeelte van Figuur 1. In de leeftijdsgroep 6.5 jaar, bijvoorbeeld, is de gemiddelde ruwe subtestscore op Onvolledige Tekeningen (OT) 12.36, met een standaarddeviatie van 2.98. Zoals te zien in figuur 1 kunnen de ruwe scoreverdelingen in principe elke vorm hebben: (niet) symmetrisch, met/ zonder duidelijke piek, et cetera. Wat echter aangenomen wordt door de WISC-III^{NL} is dat de ruwe scores een normaalverdeling *zouden* volgen op het moment dat alle mogelijke kinderen binnen een bepaalde leeftijdscategorie getest zouden zijn (Glutting & Oakland, 1993). Daarom worden de ruwe scoreverdelingen omgezet in normaalverdelingen, ongeacht de vorm van de ruwe scoreverdeling. Dit is te zien in het onderste gedeelte van Figuur 1.

IQ-scores zijn vaak onbetrouwbaarder en instabieler dan gedacht

VAN NORMSCORE NAAR TOTAALSCORE De normscores op de 10 subtests worden bij elkaar opgeteld om tot één totaalscore per kind te komen, zie pijl 1 in Figuur 3.

VAN TOTAALSCORE NAAR IQ-SCORE Om te komen tot één intelligentiemaat die vergelijkbaar is met andere intelligentietests, wordt de verdeling van totaalscores omgezet naar een IQ-verdeling, zie pijl 2 in Figuur 3.[4]

HOE WORDEN BETROUWBAARHEIDSINTERVALLEN GECONSTRUEERD?

Een betrouwbaarheidsinterval ontstaat door bij een schatting van de IQ-score van een kind 'z-maal' een schatting van meetfout op te tellen en af te trekken:

$$\text{Schatting IQ} \pm z * \text{schatting meetfout (Formule 1)}$$

De z in Formule 1 kan verschillende waarden aannemen. Vervangen we z door 1.96, dan ontstaat bijvoorbeeld een 95% betrouwbaarheidsinterval.

De meetfout waarop het betrouwbaarheidsinterval is

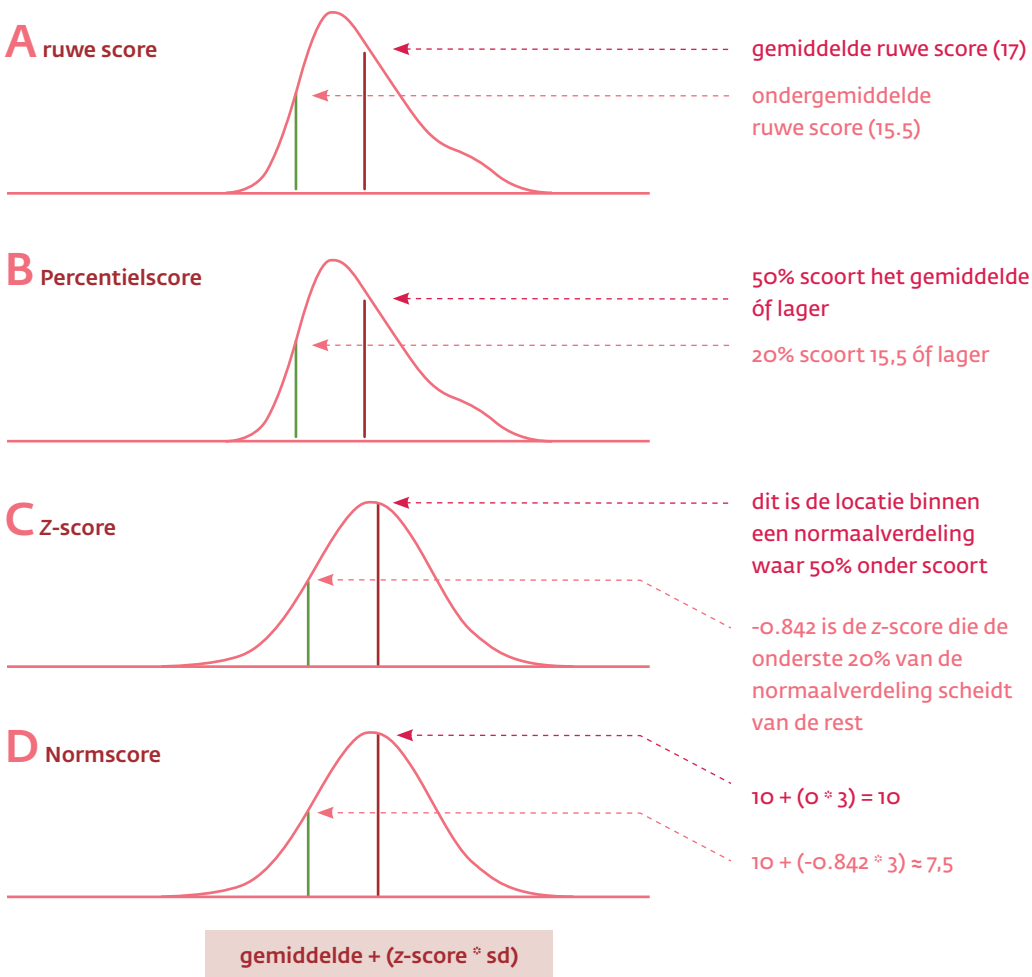
3 Merk op dat de standaarddeviaties van de ruwe subtestscore verdelingen niet zijn weergegeven in één van de WISC-III^{NL} tabellen. Echter, deze waarde kan eenvoudig berekend worden door: Standaardmeetfout/$\sqrt{(1-\alpha)}$, waar α de betrouwbaarheidscoëfficiënt is voor de specifieke subtest en specifieke leeftijdsgroep. Voor de leeftijdsgroep 6.5 en de subtest overeenkomsten geldt: 1.42 (tabel 3.7) / $\sqrt{(1-0,78)}$ (tabel 3.5) = 3.03.

4 In eerste instantie heeft de WISC-III^{NL} deze omzetting van totaalscore naar IQ-score per leeftijdscategorie apart uitgevoerd. Er bleken echter geen leeftijdseffecten te zijn (de totaalscores zijn immers al gebaseerd op gestandaardiseerde normscores).

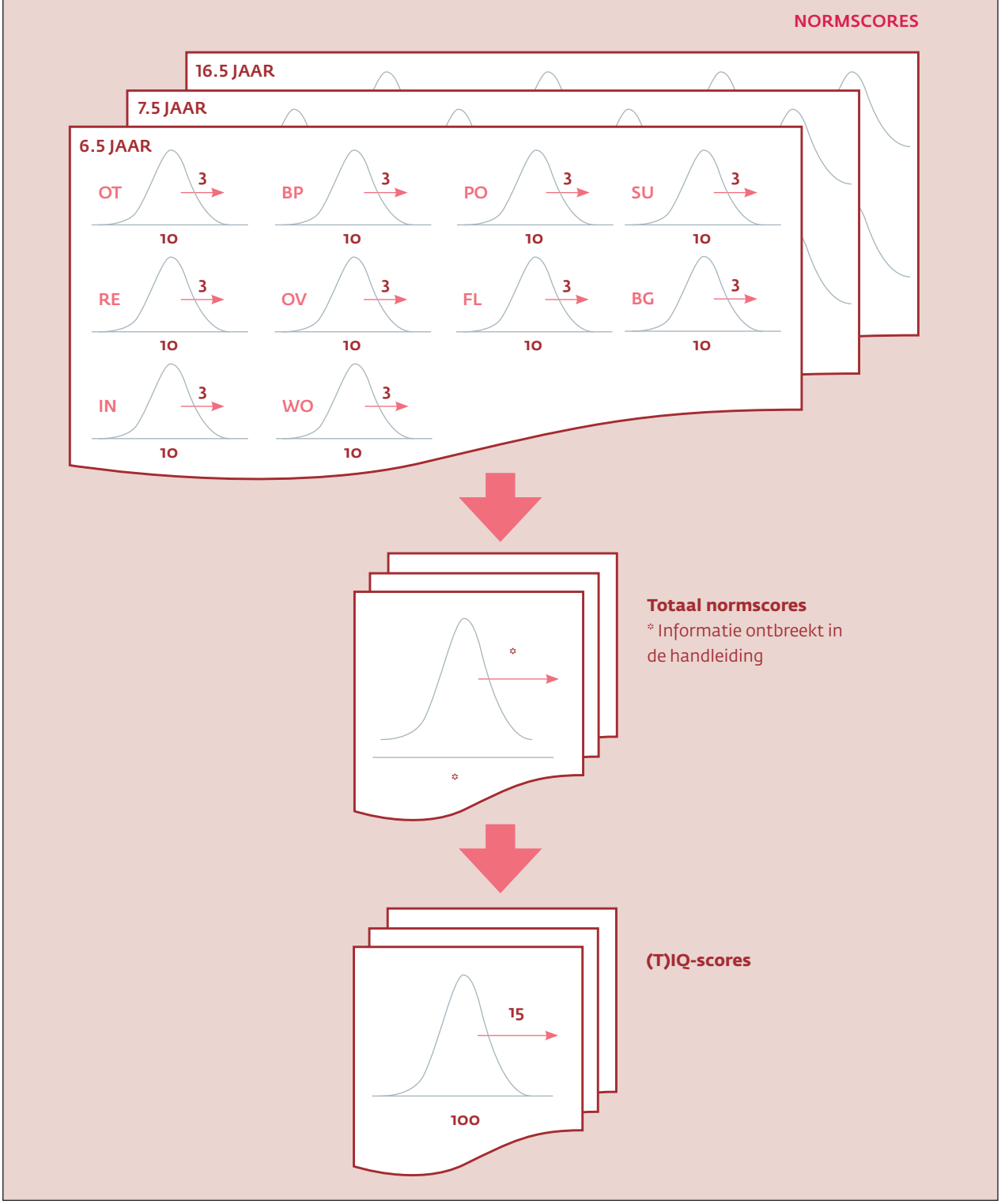
TECHNISCH INTERMEZZO 1. HOE KUNNEN RUWE SUBTESTSCORES VERTAALD WORDEN NAAR NORMSCORES?

Het omzetten of standaardiseren van ruwe subtestscores naar normscores gaat als volgt (zie Figuur 2). Na het bepalen van de ruwe scoreverdeling voor een leeftijdscategorie en subtest (zie deel A Figuur 2), worden allereerst percentielscores bepaald (Crawford, 2004). Figuur 2B toont deze percentielscores voor twee fictieve kinderen met ruwe subtestscore 15.5 en 17. Percentielscores drukken uit hoeveel procent van de kinderen in een bepaalde leeftijdscategorie een bepaalde ruwe subtestscore heeft behaald óf lager. Vervolgens worden deze percentielscores gematcht met z-scores binnen de normaalverdeling voor de normscores (deel C Figuur 2). Een z-score die aangeeft waar 90% van de scores in de normaalverdeling van normscores valt wordt bijvoorbeeld gematcht met de ruwe score die hoort bij de percentielscore 90. Als we de z-scores vervolgens vermenigvuldigen met de gewenste standaarddeviatie 3 en optellen bij het gewenste normscore gemiddelde 10, dan weten we welke normscore hoort bij een bepaalde z-score en dus, indirect, bij een bepaalde ruwe score (deel D Figuur 2).

FIGUUR 2. ILLUSTRATIE VAN DE OMZETTING VAN RUWE SCORES (A) NAAR PERCENTIELSCORES (B), Z-SCORES (C) EN UITEINDELIJK NORMSCORES (D). MERK OP DAT IN A EN B DE VERDELING VAN DE DATA WORDT AANGEHOUDEN, TERWIJL IN C EN D EEN NORMAALVERDELING AANGENOMEN WORDT



FIGUUR 3. DE WISC-III^{NL} -PROCEDURE VOOR HET VERTALEN VAN NORMSCORES NAAR IQ-SCORES, VIA TOTAALSCORES[3]



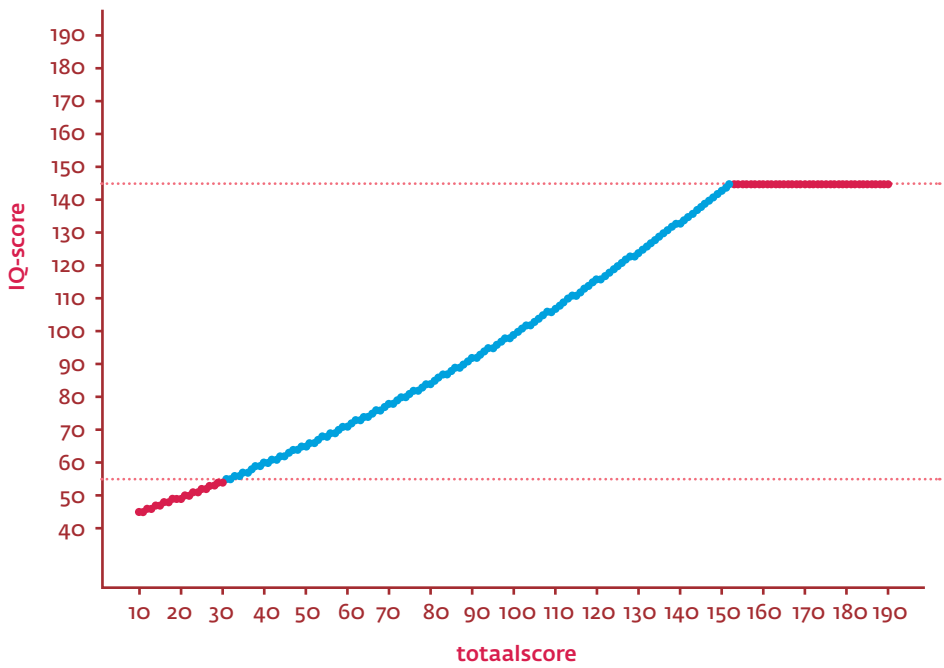
gebaseerd kan met twee verschillende formules geschat worden (zie COTAN, p. 27). Het resultaat van de ene formule wordt 'standaardmeetfout' genoemd; de ander de 'standaardschattingsfout' (zie technisch intermezzo 3). Welke formule gekozen wordt, heeft invloed op de interpretatie van het resulterende betrouwbaarheidsinterval (zie volgende secties).

Naast een schatting van de meetfout bevat formule 1 ook een schatting voor de IQ-score. Hiervoor kán de verkregen IQ-score van het kind gebruikt worden. Een andere optie is het gebruik van Kelley's formule om te controleren voor *regressie naar het gemiddelde* (Kelley, 1947). 'Regressie naar het gemiddelde' kan als volgt worden uitgelegd. Als een kind een relatief hoge of relatief lage IQ-score heeft gehaald bij een eerste intelligentietestafname, dan is de kans statistisch gezien groot dat dit kind bij (theoretische) volgende intelligentietests een minder extreme IQ-score behaalt,

TECHNISCH INTERMEZZO 2: HOE KUNNEN TOTAALSCORES VERTAALD WORDEN NAAR IQ-SCORES?

Figuur 4 visualiseert de omzetting van totaalscores (x-as) naar bijbehorende IQ-scores (y-as). De grote range van totaalscores (10 tot 190) wordt, met behulp van de normpopulatie, teruggebracht naar de toegestane range van WISC-III^{NL}-IQ-scores (45-145). De kleine verspringingen in de lijn worden veroorzaakt door afronding. Merk op dat aan de onderkant van de WISC-III^{NL} -schaal een even groot verschil in totaalscores leidt tot een kleiner verschil in IQ-scores dan aan de bovenkant van de WISC-III^{NL} -schaal. De rode punten laten zien hoe de vertaling van totaalscores naar IQ-scores eruitziet voor extreem lage (< 3 sd's) en extreem hoge (> 3 sd's) totaalscores. Opvallend is vooral dat kinderen met een totaalscore hoger dan 152 allemaal dezelfde IQ-score krijgen: 145. Dus alle kinderen met een totaalscore tussen 152 en 190 kunnen op basis van het IQ van de WISC-III^{NL} niet onderscheiden worden.

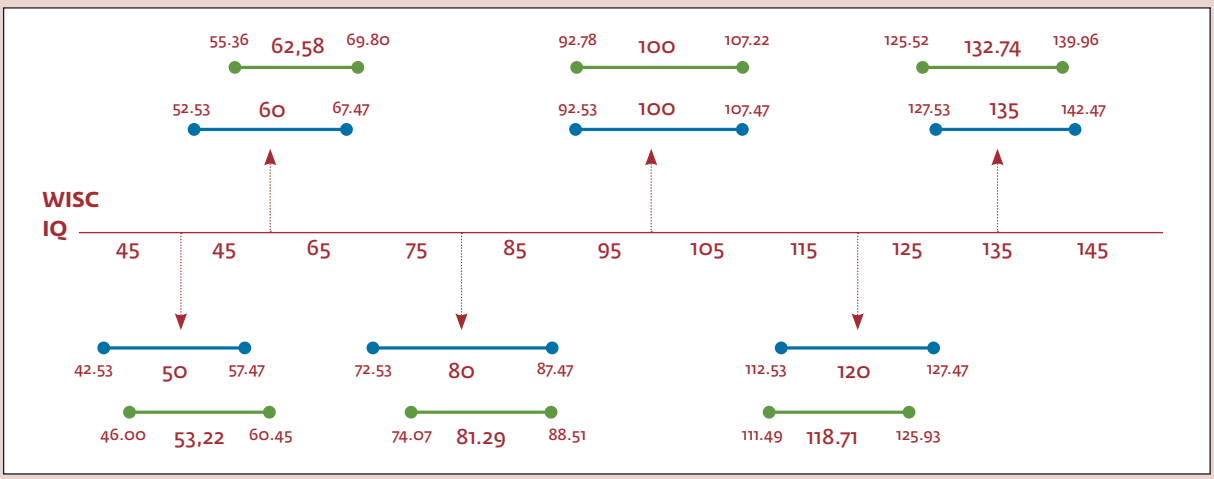
FIGUUR 4. ILLUSTRATIE OMZETTING TOTAALSCORES (X-AS) NAAR IQ-SCORES (Y-AS)



oftewel meer richting het normpopulatiegemiddelde scoort. Het risico is aanwezig dat we daardoor bij kinderen met relatief hoge IQ-scores de 'ware' IQ-score overschatten terwijl we bij een relatief lage IQ-score het risico lopen op onderschatting (Charter & Feldt, 2001). In zulke gevallen zal de 'ware' IQ-score *gemiddeld gezien* dus meer in de richting van het (norm)populatiegemiddelde van 100 liggen dan de geobserveerde IQ-score impliceert (Barnett, van der Pols & Dobson, 2005). Kelley's formule corrigeert de geobserveerde IQ-score richting het (norm)populatiegemiddelde om dit effect tegen te gaan.

In intelligentietests wordt doorgaans óf de standaardmeetfout en de verkregen IQ-score gebruikt om het betrouwbaarheidsinterval op te stellen (optie 1) óf de standaardschattingsfout in combinatie met Kelley's formule (optie 2). In de WISC-III^{NL} wordt bijvoorbeeld optie 2 gebruikt.

FIGUUR 5. VERGELIJKING VAN BETROUWBAARHEIDSINTERVALLEN GEBASEERD OP OPTIE 1 (BLAUW) EN 2 (GROEN) VOOR VERSCHILLENDE IQ-WAARDEN VAN DE WISC-III^{NL} SCHAAL



TECHNISCH INTERMEZZO 3: DE TWEE FORMULES VOOR HET SCHATTEN VAN 'MEETFOUT'

De formules voor het schatten van 'meetfout' bevatten allebei twee ingrediënten:
1) de betrouwbaarheidscoëfficiënt (betr)
2) de standaarddeviatie van de IQ-scoreverdeling (SD)

De standaarddeviatie (ingrediënt 2) laat zien hoeveel de kinderen in de normpopulatie verschillen in IQ-score. De betrouwbaarheidscoëfficiënt (ingrediënt 1) wordt gebruikt om te schatten welk deel van de verschillen tussen IQ-scores van kinderen in de normpopulatie (ingrediënt 2) veroorzaakt wordt door 'ware' IQ-verschillen[6]. In de WISC-III^{NL} is de betrouwbaarheidscoëfficiënt (ingrediënt 1) bijvoorbeeld geschat op ongeveer 0.94. Verwacht wordt dus dat 94% van de variantie in IQ-scores veroorzaakt wordt door 'ware' IQ-verschillen. De overige 6% zegt iets over het verwachtte meetfoutdeel in de normpopulatie.

De WISC-III^{NL} verwacht dat als in de normpopulatie 6% van IQ-verschillen veroorzaakt wordt door meetfouten dit ook geldt voor één specifiek kind.

De formules voor het schatten van meetfout zijn (COTAN, 2010; McManus, 2012):
Standaardmeetfout
SD * √(1-betr)
In de WISC-III^{NL} is de standaardmeetfout 3.81.

Standaardschattingsfout
SD * √(1-betr) * √betr
In de WISC-III^{NL} is de standaardschattingsfout ≈3.68.

Laten we ter illustratie voor de (fictieve) kinderen Joachim, met een IQ-score van 100, en Thomas, met een IQ-score van 70, de optie 1 en optie 2 betrouwbaarheidsintervallen opstellen, gebruikmakend van de gegevens van de WISC-III^{NL}.

Volgens optie 1 (standaardmeetfout + verkregen IQ) is het betrouwbaarheidsinterval van Joachim:

$$100 - (3.81 \times 1.96) = 92.53$$
$$100 + (3.81 \times 1.96) = 107.47$$

Ronden we deze waarden af, dan is het resulterende betrouwbaarheidsinterval voor Joachim: 93 – 107.

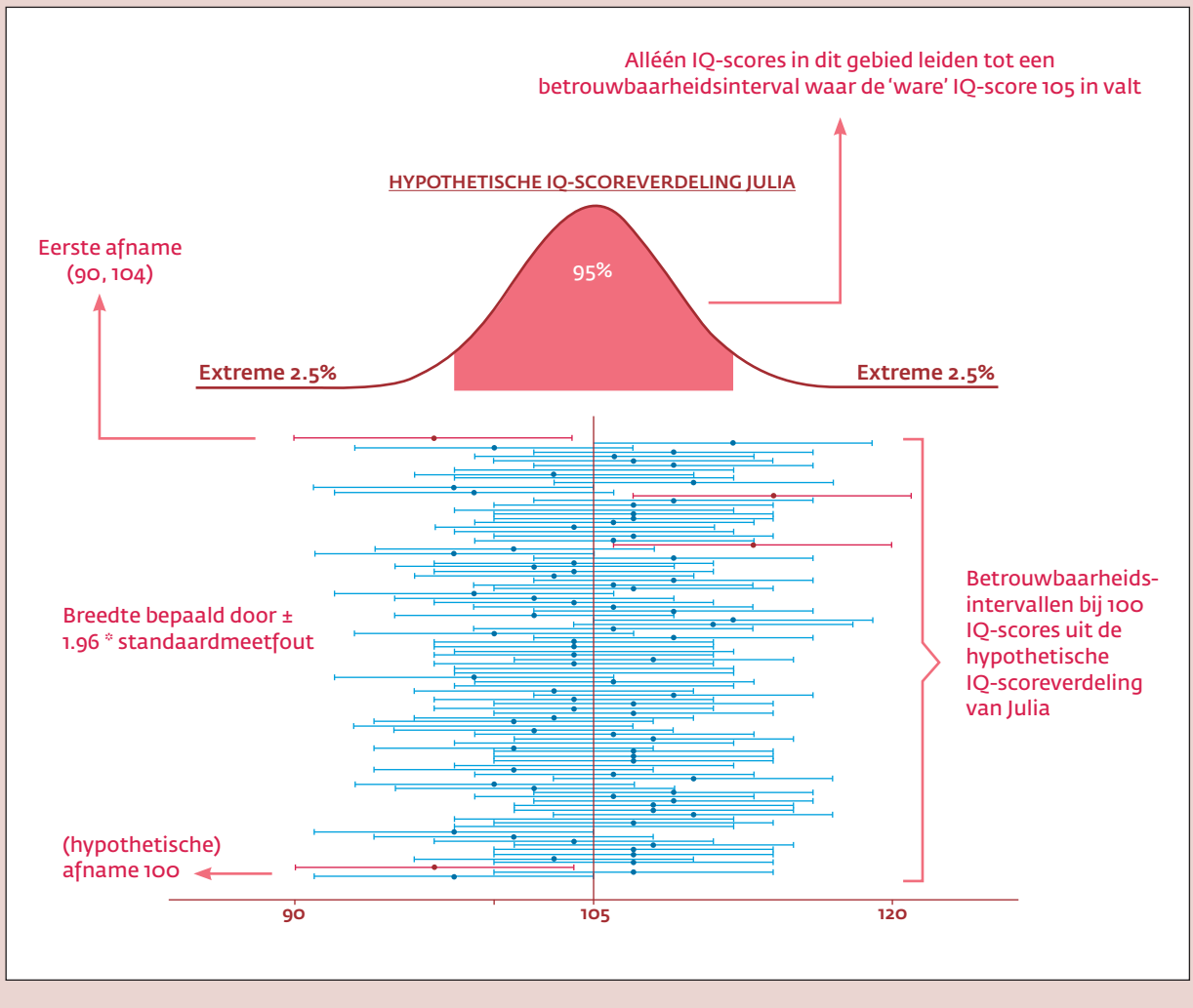
Voor Thomas leidt optie 1 tot het betrouwbaarheidsinterval:

$$70 - (3.81 \times 1.96) = 62.53$$
$$70 + (3.81 \times 1.96) = 77.47$$

Oftewel, het interval voor Thomas komt uit op: 63 – 77.

Nu optie 2 met de standaardschattingsfout in plaats van de standaardmeetfout, en gebruikmakend van Kelley's

FIGUUR 6. 100 HYPOTHETISCHE, OPTIE 1-BETROUWBAARHEIDSINTERVALLEN VOOR JULIA. DE VERTICALE LIJN GEEFT DE 'WARE' IQ-SCORE VAN JULIA AAN (105). BLAUWE BETROUWBAARHEIDSINTERVALLEN BEVATTEN DEZE 'WARE' IQ-SCORE, RODE BETROUWBAARHEIDSINTERVALLEN NIET



formule om te controleren voor regressie naar het gemiddelde. Allereerst wordt in Kelley's formule de verkregen IQ-score vermenigvuldigd met de eerder besproken betrouwbaarheidscoëfficiënt. Voor Joachim geldt dan:

100 (verkregen IQ-score) * 0.9355 (betrouwbaarheidscoëfficiënt) = 93.55

Voor Thomas geldt:

70 (verkregen IQ-score) * 0.9355 (betrouwbaarheidscoëfficiënt) = 65.48388

Vervolgens worden bij deze waardes één min de betrouwbaarheidscoëfficiënt maal 100 opgeteld.

Voor Joachim geldt:

(1 – 0.9355) * 100 = 6.452
93.55 + 6.45 = 100

Omdat zowel het normpopulatiegemiddelde als Joachim's verkregen IQ-score 100 zijn, komt zijn geschatte IQ-score met Kelley's formule óók op 100 uit.

Bij Thomas maakt de correctie met Kelley's formule echter wél uit:

(1 – 0.9354) * 100 = 6.45
65.48 + 6.45 = 71.94

Thomas' geschatte IQ-score (71.94) met Kelley's formule komt dus bijna twee IQ-punten hoger uit dan zijn verkregen IQ-score (70). Met de geschatte IQ-scores voor Joachim en Thomas en de standaardschattingsfout (3.69) kunnen we nu hun betrouwbaarheidsintervallen volgens optie 2 narekenen (deze komen overeen met de betrouwbaarheidsintervallen in tabel D.4 van de WISC-III^{NL}-handleiding):

Joachim

100 - (3.69 x 1.96) = 92.78 93
100 + (3.69 x 1.96) = 107.22 107

Thomas

71,94 - (3.69 x 1.96) = 64.71 65
71,94 + (3.69 x 1.96) = 79.15 79

Merk op dat het voor Joachim niet uitmaakt of zijn betrouwbaarheidsinterval volgens optie 1 of optie 2 wordt berekend. Voor Thomas maakt dit wél uit. Over het algemeen geldt dat hoe verder de verkregen IQ-score van 100 af ligt, des te groter de verschillen zijn tussen optie 1 en 2 (zie figuur 5).[5]

De Klassieke Testtheorie veronderstelt dat elke kind één 'ware' IQ-score heeft

HOE KUNNEN WISC-III^{NL} -BETROUWBAARHEIDSINTERVALLEN GEÏNTERPRETEERD WORDEN?

Zowel het betrouwbaarheidsinterval van optie 1 (standaardmeetfout en verkregen IQ; zie voorgaande sectie) als 2 (standaardschattingsfout en Kelley's gecorrigeerde IQ) is gebaseerd op de Klassieke Test Theorie (KTT; Guilford, 1936; Lord & Novick, 1986). KTT gaat uit van de aanname dat elk kind één 'ware' IQ-score heeft. De IQ-score die een kind verkrijgt, kan in meer of mindere mate afwijken van deze 'ware' IQ-score, wat in de KTT beschouwd wordt als meetfout. In een realistische testsituatie weten we nooit de 'ware' IQ-score van een kind en dus ook niet hoe groot of klein de meetfout is die we met één IQ-resultaat begaan. In de KTT wordt echter aangenomen dat *als* we in staat zouden zijn *oneindig veel* intelligentietests af te nemen bij een kind, de resulterende IQ-scores een normaalverdeling volgen (Wang & Osterlind, 2013). De piek van deze normaalverdeling is gelijk aan de 'ware' IQ-score van het kind, zodat gemiddeld gezien de verkregen IQ-score de 'ware' IQ-score goed inschat. Omdat de normaalverdeling symmetrisch is, wordt aangenomen net zo vaak een onderschatting te maken van de IQ-score als een overschatting. Het betrouwbaarheidsinterval van optie 1 en optie 2 kan geïnterpreteerd worden vanuit dit gedachtegoed van de KTT.

OPTIE 1 Stel dat een kind vele intelligentietests zou voltooien. Dan verwachten we dat dit kind relatief vaak een IQ-score

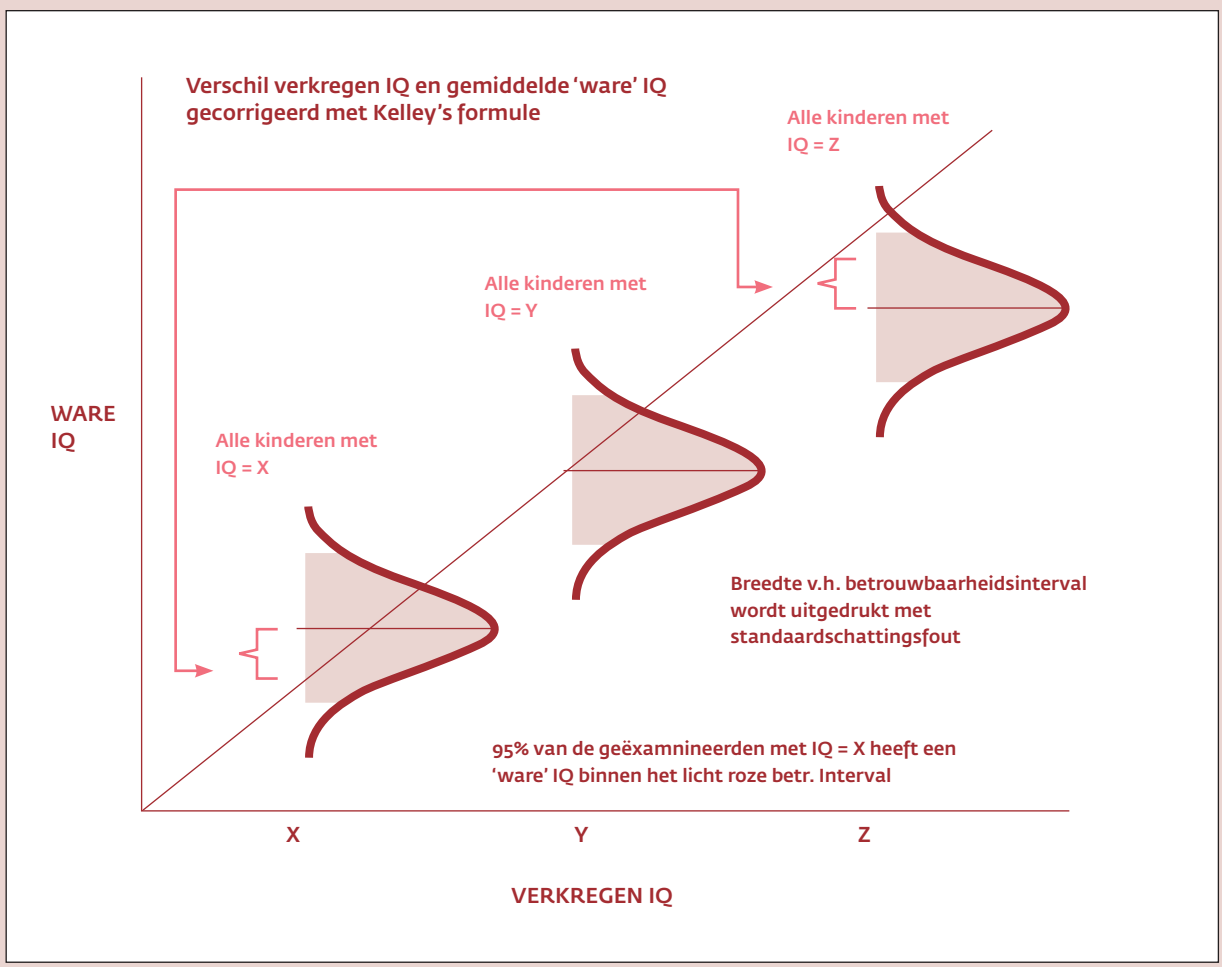
5 Voor Julia (IQ = 97) geldt dat de optie 1 en optie 2 betrouwbaarheidsintervallen na afronding precies hetzelfde zijn: 90 tot 104.

zal halen dichtbij zijn/haar 'ware' IQ-score (zijn 'piek'). Dit betekent dat als wij een betrouwbaarheidsinterval zouden opstellen rondom deze verkregen IQ-scores, deze relatief vaak de 'ware' IQ-score van het kind zal omvatten. Hoe vaak 'relatief vaak' is, hangt af van de breedte van het betrouwbaarheidsinterval. Het idee van optie 1 is de breedte van het betrouwbaarheidsinterval zó te kiezen dat als we oneindig vaak een intelligentietest zouden afnemen (theoretisch gezien) en het bijbehorende 95% betrouwbaarheidsinterval zouden berekenen, deze in 95% van de gevallen de 'ware' IQ-score van het kind zou bevatten. Specifiek wordt deze breedte bepaald met de eerder geïntroduceerde *standaardmeetfout*.

Ter illustratie toont figuur 6 de 95% betrouwbaarheidsintervallen van 100 (hypothetische) intelligentietestafnamen bij Julia. Te zien is dat ongeveer 95% van deze betrouwbaarheidsintervallen de 'ware' IQ-score van Julia – 105 - bevat (blauwe intervallen).

Samengevat is de definitie van optie 1: 'Bij vele herhalingen van intelligentietests *verwachten* we dat 95% van de bijbehorende betrouwbaarheidsintervallen de 'ware' IQ-score van het kind bevatten. Voorafgaand aan een intelligentietest hebben we dus 95% kans dat het resulterende 95% betrouwbaarheidsinterval de 'ware' IQ-score van het kind zal bevatten.'

FIGUUR 7. ILLUSTRATIE VAN DE INTERPRETATIE VAN OPTIE 2-BETROUWBAARHEIDSINTERVALLEN (ZOALS DIE IN DE WISC-III^{NL})

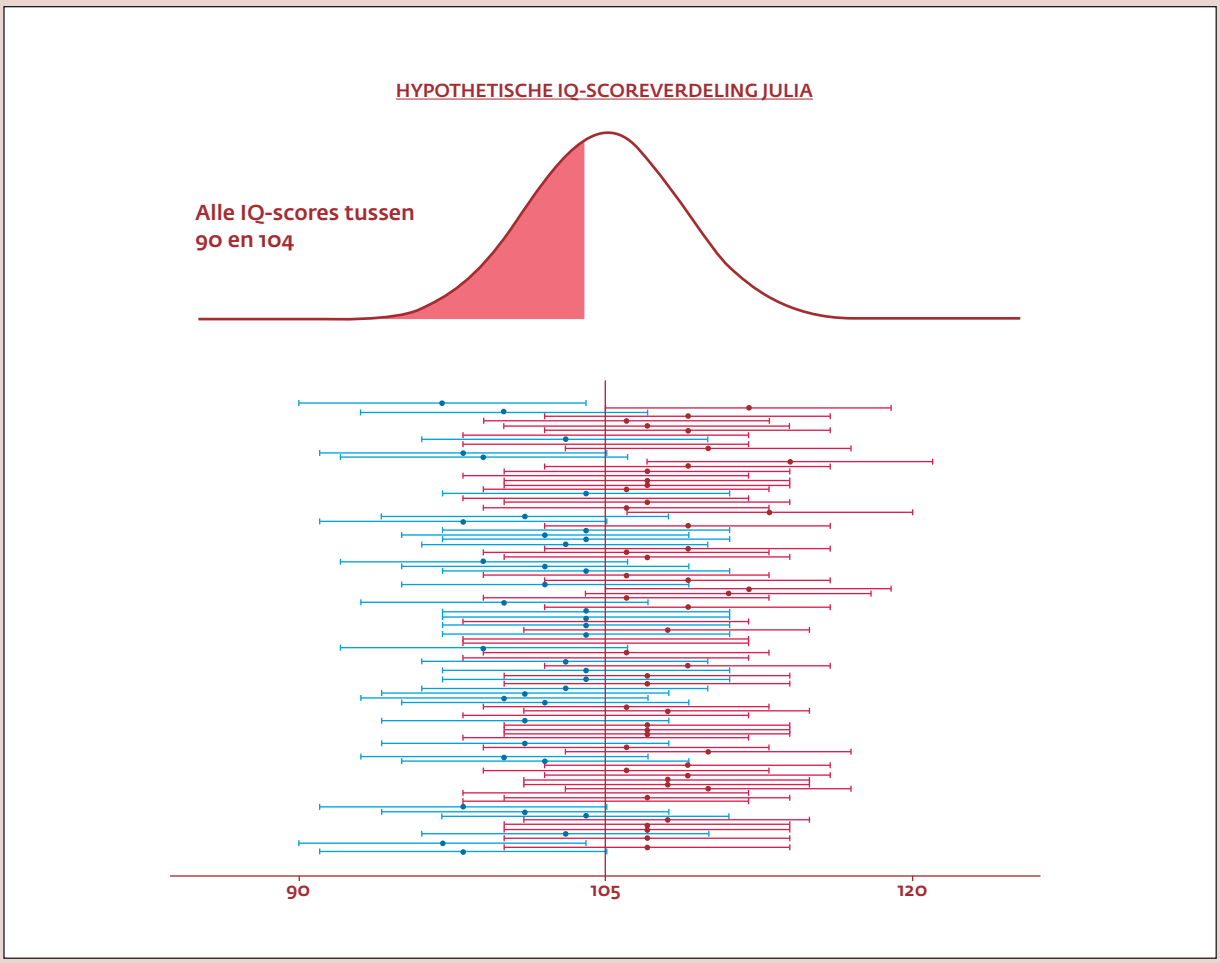


OPTIE 2 Stel nu dat we niet oneindig vaak een intelligentietest afnemen bij één kind maar slechts één keer bij oneindig veel kinderen. In dat geval zouden we kinderen kunnen groeperen op basis van hun verkregen IQ-score. Dit is geïllustreerd in figuur 7 voor kinderen met score 'X', 'Y' en 'Z'. De optie 2 betrouwbaarheidsintervallen zijn gebaseerd op de vraag: welke 'ware' IQ-scores zullen de kinderen met dezelfde verkregen IQ-score hebben? Omdat de verkregen IQ-score in meer of mindere mate kan afwijken van de 'ware' IQ-scores van elk van de kinderen in één groep, verwachten we niet één 'ware' IQ-score maar een verdeling zoals de drie normaalverdelingen in figuur 7.

Op basis van de KTT kunnen we het gemiddelde en de standaarddeviatie van deze verdeling voor een specifieke verkregen IQ-score schatten (zie Charter & Feldt, 2001). Dit gemiddelde is uitgelegd onder het voorgaande kopje onder de naam *Kelley's formule* en de standaarddeviatie als *standaardschattingsfout*. Op basis van de geschatte verdeling van 'ware' IQ-scores voor één verkregen IQ-score kan een gebied geselecteerd worden waarvan verwacht wordt dat deze 95% van de 'ware' IQ-scores van de kinderen bevat. Dit is te zien in Figuur 7 als de drie licht roze gearceerde gebieden in elk van de normaalverdelingen voor score 'X', 'Y' en 'Z'.

Samengevat is de definitie van optie 2: 'Voor kinderen

FIGUUR 8. ILLUSTRATIE VAN DE (FOUTIEVE) INTERPRETATIE "WANNEER IK BIJ JULIA 100 KEER DE IQ TEST ZOU AFNEMEN, ZOU HET IN 95% VAN DIE AFNAMEN EEN IQ TUSSEN 90 EN 104 OPLEVEREN".



Als een kind een relatief hoge of relatief lage IQ-score heeft gehaald bij een eerste intelligentietest, dan is de kans groot dat dit kind bij volgende afnames meer richting het normpopulatie-gemiddelde scoort

met een IQ-score gelijk aan die van het geteste kind, verwachten we dat 95% van hen eigenlijk een IQ-score heeft tussen de linker- en rechtergrens van het betrouwbaarheidsinterval (Charter & Feldt, 2001).'

AANNAMES Wat belangrijk is om in gedachten te houden – zowel voor optie 1 als 2 –, is dat de interpretatie van betrouwbaarheidsintervallen zoals hierboven beschreven *enkel* opgaat wanneer alle aannames kloppen. Zo neemt zowel optie 1 als optie 2 aan dat ieder kind gemeten is met evenveel meet(on)zekerheid. Er zijn meerdere redenen te noemen waarom deze aanname onrealistisch is (Sijtsma, 2009; Molenaar, 2004). Wanneer de items, bijvoorbeeld, (veel) te makkelijk of (veel) te moeilijk zijn voor een kind, dan kan zijn of haar intelligentie slecht ingeschat worden door de desbetreffende intelligentietest. De test maakt daardoor relatief gezien méér meetfouten dan bij kinderen bij wie het intelligentieniveau beter aansluit bij de moeilijkheid van de items. Ook kunnen er andere redenen zijn waarom er méér meet(on)zekerheid is bij het ene kind dan bij het ander. Denk aam zaken als concentratie, faalangst, meertaligheid en (ervaring van) testafnemer.

In optie 2 wordt daarnaast aangenomen dat er geen fundamentele verschillen in gemiddelde (sub)testscores zijn tussen bijvoorbeeld jongens en meisjes of tussen kinderen met een westerse en niet-westerse achtergrond. Wanneer dit wel het geval is, geeft Kelley's formule een verkeerd beeld van de verwachte 'ware' IQ-score, omdat de formule deze verschillen niet in ogenschouw neemt. Dit wordt Kelley's paradox genoemd (Wainer, 2000; Borsboom, Romeijn & Wicherts, 2008).

Kortom, het is goed in gedachte te houden dat voor een specifiek kind of groep met kinderen meetfouten groter kunnen zijn en de betrouwbaarheidsintervallen een onderschatting van meetonzekerheid kunnen geven. Of – zoals een van onze reviewers het verwoordde – het is belangrijk bewust te zijn van de 'zachtheid' van testdata. Dàt het betrouwbaarheidsinterval nauwkeurig berekend kan worden tot vele cijfers achter de komma (zie sectie 'Hoe worden betrouwbaarheidsintervallen geconstrueerd?'), betekent dus niet dat we met zoveel nauwkeurigheid IQ-scores kunnen uitsluiten bij kinderen (ook wel 'valse precisie' genoemd; zie Huff, 2010).

VERGELIJKING MET VEELVOORKOMENDE (FOUTIEVE) INTERPRETATIES

In november 2014 hebben 293 NIP-psychologen deelgenomen aan onze survey. Daarin vroegen wij onder meer hoe de deelnemers 95% betrouwbaarheidsintervallen interpreteren en of zij deze ook rapporteren in IQ-verslagen. Opvallend is dat 65% van de NIP-psychologen in onze survey definities gaven van het WISC-III^{NL}-betrouwbaarheidsinterval van Julia (zie inleiding) die lijken op de definitie van optie 1. Deze definitie geldt echter niet voor het WISC-III^{NL}-betrouwbaarheidsinterval omdat deze aansluit bij optie 2. Daarnaast wijzen sommige definities uit de survey op enkele valkuilen in de interpretatie van optie 1. Zo gaven 14 deelnemers de definitie: 'Wanneer ik bij Julia 100 keer de IQ-test zou afnemen, zou het in 95% van die afnamen een IQ tussen 90 en 104 opleveren.' Deze definitie wijkt op twee belangrijke punten af van de definitie van optie 1-betrouwbaarheidsintervallen hierboven. Allereerst focust deze definitie op het verkregen IQ en niet op het 'ware' IQ. Ten tweede wordt gerefereerd aan het specifieke interval 90-104 en niet aan betrouwbaarheidsintervallen in het algemeen.

Vergelijk figuur 6, behorend bij de definitie van het optie 1-betrouwbaarheidsinterval, met figuur 8. In zowel figuur 6 als figuur 8 betreft het bovenste betrouwbaarheidsinterval het interval behorend bij de eerste testafname[6] (90-104); de daaropvolgende 99 betrouwbaarheidsintervallen zijn van hypothetische vervolgafnamen, bij een 'ware' IQ-score van 105. In figuur 6 zijn alle betrouwbaarheidsintervallen

6 De betrouwbaarheidscoëfficiënt is direct geschat op basis van de normpopulatie. De normpopulatie heeft hiermee een directe invloed op de breedte van het betrouwbaarheidsinterval (zowel optie 1 als 2). Wanneer de normpopulatie niet representatief is voor kinderen die een WISC-III^{NL}-intelligentietest maken kan dit dus verstrekkende gevolgen hebben voor de interpretatie van het betrouwbaarheidsinterval.

roodgekleurd die niet de 'ware' IQ-score 105 bevatten; waaronder de huidige afname. Dit betreft *ongeveer* 5% van de betrouwbaarheidsintervallen. 'Ongeveer' omdat de 100 betrouwbaarheidsintervallen gebaseerd zijn op random IQ-scores uit de hypothetische IQ-verdeling van Julia. Zouden we random een andere set van 100 IQ-scores selecteren, dan kunnen er toevallig 3,4,5,6 et cetera betrouwbaarheidsintervallen niet de 'ware' IQ-score bevatten. Figuur 8 is een kopie van figuur 6, alleen nu zijn de betrouwbaarheidsintervallen roodgekleurd waarvan de verkregen IQ-score niet tussen 90 en 104 ligt, naar de definitie uit de survey 'Wanneer ik bij Julia 100 keer de IQ-test zou afnemen, zou het in 95% van die afnamen een IQ tussen 90 en 104 opleveren.' Omdat 90-104 toevallig bij de 5% behoort die niet Julia's 'ware' IQ-score bevat (zie figuur 6), ligt het percentage roodgekleurde intervallen in figuur 8 vele malen hoger dan 5%, om precies te zijn is dit 61%. Alléén als de verkregen IQ-score precies overeenkomt met de 'ware' IQ-score komen de definitie van optie 1 en de definitie uit de survey overeen.

Een definitie die ook vaak gegeven werd (41% van de antwoorden), was: 'De kans dat Julia's 'ware' IQ-score tussen de 90 en 104 valt, is 95%.' Deze definitie is wat kort door de bocht. Wanneer een intelligentietest is afgerond ligt de 'ware' IQ-score in het resulterende betrouwbaarheidsinterval (kans = 1) óf niet (kans = 0). Een purist zou daarom zeggen dat de kans 95% alleen geldt vóór aanvang van de test, en niet wanneer het gerealiseerde interval bekend is.

PRAKTISCHE CONSEQUENTIES

Het belangrijkste, praktische verschil tussen de beide soorten betrouwbaarheidsintervallen is dat in het geval van 'optie 2' de grenzen van het betrouwbaarheidsinterval – bijvoorbeeld 90 en 104 – *direct* geïnterpreteerd kunnen worden. We kunnen dus zeggen dat er 95% kans is dat het 'ware' IQ van Julia tussen de linker- en rechtergrens van het interval ligt. Immers, 95% van alle kinderen met een vergelijkbare verkregen IQ-score hebben een 'ware' IQ-score tussen de grenzen van het betrouwbaarheidsinterval (als aan alle aannames is voldaan). In de optie 1-betrouwbaarheidsintervallen zijn de grenzen van het betrouwbaarheidsinterval *niet* direct te interpreteren. De breedte van het betrouwbaarheidsinterval is zó bepaald dat op de *lange termijn* 95% van de betrouwbaarheidsintervallen de 'ware' IQ-score bevat; de grenzen van het huidige interval zijn hier een artefact van. We *verwachten* dat het 'ware' IQ van het kind tussen de grenzen van het huidige betrouwbaarheidsinterval ligt (we hebben immers vooraf 95% kans dat dit het geval is), maar zeker weten we het niet. Hoeveel nadruk er in de praktijk kan worden gelegd op de grenzen van het betrouwbaarheidsinterval hangt dus af van het 'type' betrouwbaarheidsinterval (optie 1 of 2).

Er zijn verschillende manieren om erachter te komen of een (intelligentie)test type 1 of type 2 betrouwbaarheidsintervallen rapporteert. Soms staat het in de handleiding vermeld (zoek bijvoorbeeld op de termen 'standaardmeetfout', 'standaardschattingsfout' en 'Kelley's formule' of check of het betrouwbaarheidsinterval in formulevorm gegeven wordt – zie technisch intermezzo 3). Zo niet, dan kan dit eenvoudig gecheckt worden door naar het betrouwbaarheidsinterval te kijken van de hoogst en de laagst mogelijke IQ-score. Als het midden van dit betrouwbaarheidsinterval na afronding hoger (laagst mogelijke score) of lager (hoogst mogelijke score) ligt dan de verkregen IQ-score is (hoogstwaarschijnlijk) gebruik gemaakt van Kelley's formule. Ten slotte is het mogelijk zelf de betrouwbaarheidsintervallen voor optie 1 en 2 uit te rekenen en te vergelijken met de betrouwbaarheidsintervallen die in de handleiding staan gerapporteerd (rekening houdend met eventuele afrondingsverschillen). Om het zo makkelijk mogelijk te maken, kan onze gratis app (zie wederom https://osf.io/7gz4q/) voor u de optie 1 en 2 betrouwbaarheidsintervallen uitrekenen als u de benodigde ingrediënten invult.

Hoe zit het ook alweer met de constructie van betrouwbaarheidsintervallen en de interpretatie ervan?

TAKE HOME MESSAGE

IQ-scores zijn maar in zekere mate stabiel en vaak onbetrouwbaarder en instabieler dan gedacht wordt. Daarom is het belangrijk niet enkel naar de IQ-score te kijken. Dit geldt in zekere mate ook voor classificatie van IQ-scores (bijv. Resing & Blok, 2002), omdat hier een term als 'beneden gemiddeld' wordt toegekend op basis van de verkregen IQ-score (Tellegen, 2004). Mits correct geïnterpreteerd en met inachtneming van aannames en beperkingen kunnen de optie 1 en 2 betrouwbaarheidsintervallen als beschreven in dit artikel helpen de IQ-score in perspectief te plaatsen.

OVER DE AUTEURS

Kimberley Lek, Msc., is PhD-kandidaat bij Methodenleer en Statistiek bij de Universiteit Utrecht. Wenneke van de Schoot-Hubeek, Msc., is gedragswetenschapper en K&J Schoolpsycholoog i.o. bij het Schreuder College van Horizon Jeugdzorg & Onderwijs. Dr. Evelyn Kroesbergen is associate professor Educatie en Pedagogiek bij de Universiteit Utrecht. Prof. dr. Rens van de Schoot is hoogleraar bij Methodenleer en Statistiek bij de Universiteit Utrecht. Correspondentie aangaande dit artikel via Kimberley Lek, e-mail: k.m.lek@uu.nl.

De auteurs danken dr. Helen Bakker, dr. Esmee Verhulp, Noëlle Pameijer, dr. Jelte Wicherts en prof. dr. Rob Meijer voor hun waardevolle commentaar op een eerdere versie van dit artikel. Ook danken zij het sectiebestuur Schoolpyschologen van het NIP en de deelnemers aan hun survey in 2014. De eerste auteur wordt ondersteund met een NWO-talent beurs (406-15-062), de laatste auteur met een NWO-VIDI (452-14-006).

Summary

WHAT TO DO WITH CONFIDENCE INTERVALS IN INTELLIGENCE TESTS?

K. LEK, W. VAN DE SCHOOT-HUBEEK, E. KROESBERGEN & R. VAN DE SCHOOT

The IQ-score of a child at a certain moment in time depends on many factors, including his/her attention and motivation, the test environment and the test leader. This makes that the IQ-score is not a perfect reflection of the intelligence of a child. In the interpretation of the IQ-score it is therefore advisable to take the accompanying 95% confidence interval into account, which acts as a 'margin of error'. Problematic with this confidence interval, however, is that its meaning is often misunderstood. In this paper, we discuss the exact meaning and interpretation of the confidence interval in intelligence tests.

Literatuur

Barnett, A.G., van der Pols, J.C. & Dobson, A. J. (2005). Regression to the mean: what it is and how to deal with it. *International Journal of Epidemiology, 34*, 215-220.

Borsboom, D., Romeijn, J.-W. & Wicherts, J.M. (2008). Measurement invariance versus selection invariance: is fair selection possible? *Psychological Methods, 13*(2), 75-98.

Charter, R.A. & Feldt, L.S. (2001). Confidence intervals for true scores: is there a correct approach? *Journal of Psychoeducational Assessment, 19*, 350-364.

Crawford, J.R. (2004). Psychometric foundations of neuropsychological assessment. In L.H. Goldstein & J.E. McNeil (Eds.), *Clinical neuropsychology: a practical guide to assessment and management for clinicians* (p. 121–140). John Wiley & Sons, Ltd.

Embretson, S.E. (1996). The new rules of measurement. *Psychological Assessment, 8*(4), 341-349.

Evers, A., Lucassen, W., Meijer, R. & Sijtsma, K. (2010). *COTAN beoordelingssysteem voor de kwaliteit van tests (geheel herziene versie)*. FMG: Psychology Research Institute.

Glutting, J. & Oakland, T. (1993). *GATSB Guide to the Assessment of Test Session Behavior for the WISC-III and the WIAT.* Psychological Corporation.

Guilford, J.P. (1936). *Psychometric methods*. New York: McGraw-Hill.

Hoekstra, R., Morey, R.D., Rouder, J.N. & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review, 21*(5), 1157-1164.

Huff, D. (2010). *How to lie with statistics*. New York: W. W. Norton & Company.

Kelley, T.L. (1947). *Fundamentals of statistics*. Cambridge: Harvard University Press.

Lord, F.M. & Novick, M.R. (1986). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

McManus, I.C. (2012). The misinterpretation of the standard error of measurement in medical education: a primer on the problems, pitfalls and peculiarities of the three different standard errors of measurement. *Medical Teacher, 34* (7), 569-576.

Molenaar, P.C.M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology – This time forever. *Measurement, 2*, 201-218.

Pameijer, N. (2014). Waarom een ontwikkelingsperspectief meer is dan IQ en leerrendement. Geraadpleegd op 13-02-2017, van http://wij-leren.nl/intelligentietest-passend-onderwijs.php

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

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's Alpha. *Psychometrika, 74*(1), 107-120.

Schouws, S. (2015). Zin en onzin van het meten van intelligentie. *Psycho-Praktijk, 3*, 34-36.

Tellegen, P. (2002). De handleiding van de WISC-III^{NL}: correcties, opmerkingen en suggesties. Verkregen via: http://www.testresearch.nl/wisc/wiscopm.html,16 februari 2016.

Tellegen, P. (2004). De waan van "het" IQ. Verkregen via: http://www.testresearch.nl/tstdiagn/waaniq.html, 3 februari 2017.

Van Ravenzwaaij, D. & Hamel, R. (2006). De Nederlandstalige WAIS-III na hernormering. *De Psycholoog*, 268–271.

Wainer, H. (2000). Kelley's paradox. *Chance, 13*, 47-48.

Wang, Z. & Osterlind, S. J. (2013). Classical test theory. In T. Teo (ed.), *Handbook of Quantitative Methods for Educational Research* (p. 31-44). SensePublishers.

Zachary, R.A. & Gorsuch, R.L. (1985). Continuous norming: implications for the WAIS-R. *Journal of Clinical Psychology*, 41, 86-94.