N=1: nauwkeurige en sensitieve behandelevaluatie op individueel niveau

Eerste versie van verhandeling over de N=1-methode voor dataverwerking

Arnold Bartels, Marinus Spreen, Erwin Schuringa & Vasthi Teeken[1]

Utrecht/Groningen/Doorwerth:
Expertisecentrum Forensische Psychiatrie (EFP) /
FPC Dr. S. van Mesdag / Dr. Leo Kannerhuis,
16 september 2008

[1] Dr. Arnold A.J. Bartels is klinisch psycholoog-psychotherapeut (gedragstherapeut) en senior-onderzoeker aan het *Dr. Leo Kannerhuis* in Doorwerth en Oosterbeek. Hij was tot 1 juli 2007 manager *Psychodiagnostisch & Wetenschappelijk Onderzoek* en coördinerend hoofdbehandelaar van het patiëntencluster *Training & Vaardigheden* van het Forensisch-Psychiatrisch Centrum (FPC) de *Dr. S. van Mesdag*. Thans is hij nog programmaleider van het zorgprogramma autismespectrumstoornissen en consulent onderzoek, hoofdbehandelaar van de sectie forensische psychiatrie Forint van Lentis in Groningen. Dr. Marinus Spreen is socioloog, methodoloog, was tot 1 juli 2007 senior wetenschappelijk onderzoeker aan die kliniek en daarna hoofd wetenschappelijk onderzoek. Drs. Erwin Schuringa is junior-onderzoeker in het FPC Van Mesdag en onder andere bezig met N=1-onderzoek. Mevr. Vasthi Teeken is werkzaam bij het Autismeteam *GGZ Buitenamstel* te Amsterdam. De gegevens van de in dit artikel als voorbeeld gebruikte casus zijn afkomstig van haar N=1-studie in het kader van het traject tot haar lidmaatschap van de *Vereniging voor Gedrags- en Cognitieve therapie* (VGCt).

Inhoud

Inleiding

De bestaande evaluatiemethoden voor het bepalen van vooruitgang en effect van interventies als trainingen en therapieën zijn vooral gebaseerd op het vergelijken van groepen: één of enkele interventiegroepen en idem controlegroepen. De toewijzing van personen aan de interventie- en controlegroep is daarbij bij voorkeur willekeurig (random). Daarmee heeft men een onderzoeksopzet met de hoogste interne validiteit: met de grootste waarschijnlijkheid dat een gevonden statistisch significant verschil werkelijk een effect betekent. Zo'n onderzoeksopzet heet een 'randomized control trial' of 'random clinical trial' (RCT).
Een RCT is de 'gouden standaard', maar een RCT is doorgaans een forse investering in menskracht, organisatie, middelen (financiën vooral) en tijd. De uitslag laat enkele jaren op zich wachten en de betekenis nog weer enkele jaren als reacties in de wetenschappelijke tijdschriften zijn verwerkt.
Deze tijd werkt vertragend. Het betekent ook dat de feedback van wat werkt en wat niet werkt of anders werkt dan verwacht, pas laat tot stand komt (na de ontwikkeling van een nieuwe methode die men in een RCT toetst) en ook vaak van globale aard is. De ontwikkeling van nieuwe methoden verloopt daarmee ook traag. Deels is dit een gegeven en inherent aan het proces van valide kennis verwerven.

Er zijn alternatieven voor het RCT zoals quasi-experimenteel onderzoek (interventie- en controlegroep worden daarbij zoveel mogelijk gelijk samengesteld, maar niet op basis van willekeurige toewijzing), maar deze vragen vaak nauwelijks minder investering en de resultaten ervan zijn methodologisch minder overtuigend. Deze alternatieven komen alleen eigenlijk in aanmerking als een RCT werkelijk niet te realiseren is of, als ze duidelijk minder investering vragen, en dienen dan als voorbereiding op een RCT.

Het loont echter zeer de moeite als voorbereiding van en opmaat tot een RCT onderzoek te doen dat een redelijke indicatie geeft over effect en dat tevens directe feedback kan bieden voor de ontwikkeling van behandelmethoden. Aldus heeft men meer zekerheid als het RCT gedaan gaat worden, dat het iets zal opleveren: nogal belangrijk gezien de genoemde grote investering bij een RCT. En een evaluatiemethode die sneller feedback geeft, in de ontwikkelfase van behandelmethoden, is ook buitengewoon aantrekkelijk.

Zulk een methode bestaat in principe, maar de methodologie ervan is nog in ontwikkeling: het 'single case design' of 'single case research design', in het Nederlands ook wel N=1 geheten. Naar het aantal proefpersonen (N) dat 1 is.

N=1 stond eigenlijk aan de wieg van de empirische psychologie, de onderzoeken in de 19e en het begin van de 20e eeuw waren vaak onderzoeken bij één persoon of dier. Nadat in de jaren dertig van de 20e eeuw goede methoden voor statistisch vergelijken van groepen beschikbaar kwamen, kreeg onderzoek bij groepen een sterke stimulans. N=1-designs bleven bestaan, vooral bij de ontwikkelaars en theoretici op het gebied van de operante leertheorieën en de gedragstherapie.
Het design dat daarbij gebruikt werd, had het nadeel dat men, om de werkzaamheid van de interventie doorslaggevend aan te tonen, na de interventie terug moest naar de vóór-behandelfase, de baseline-fase. Dit door naar de condities van de baseline-fase terug te gaan. Als dat lukte, was dat een sterke aanwijzing dat de interventie de oorzakelijke factor was van de verandering (verbetering). Maar in de GGZ en bij behandelingen die ernaar streven om mensen 'als persoon' (in manieren van denken en attituden) te veranderen, is dat gewoonlijk een pro-

bleem. Effecten van psychotherapie en een behandelprogramma zijn bedoeld om te beklijven en men wil ze zo maken dat ze moeilijk teruggedraaid kunnen worden.

Er werden voorts wel methoden ontwikkeld om bij een individuele behandeling een voorzichtig antwoord te kunnen geven op de vraag of er vooruitgang is en of deze het waarschijnlijke gevolg is van de interventie (behandeling). Zoals doel-realisatieschema's en Goal Attainment Scaling (GAS). En procedures zoals het nagaan van verbeteringen inzake normklassen. Dat wil zeggen iemand gaat bijvoorbeeld vooruit van de categorie 'beneden-gemiddeld' naar 'gemiddeld'.
Tot dusver ontberen deze methoden een goed werkbare methodologisch-statistische strategie en voldoende overtuigingskracht om op basis van in de GGZ-praktijk gebruikte psychometrische instrumenten en observaties met redelijke waarschijnlijkheid aan te kunnen geven of er werkelijk vooruitgang is en of deze het waarschijnlijke gevolg is van de behandeling.
Er bestaan in principe twee methoden om individuele vooruitgang statistisch te evalueren. Deze zijn gebaseerd op de bestaande statistische methoden voor groepen. Het gaat om het werken met de standaardmeetfout en de standaardmeetfout van een verschil. Elke meting heeft een meetfout; op basis van de test-hertestbetrouwbaarheid uit de schaalconstructie bij een proefgroep kan men deze berekenen. Een verschil tussen twee metingen moet een bepaalde grootte van de standaardmeetfout bedragen, wil er met een vooraf bepaalde waarschijnlijkheid werkelijk sprake van een verschil zijn. Ditzelfde kan men ook doen met de standaardmeetfout van een verschil tussen twee scores.
Het nadeel van deze procedures is dat ze om erg grote verschillen vragen, vaak veel groter dan in de praktijk bij individuele vooruitgang optreedt. Dat is begrijpelijk: de toets corrigeert als het ware voor toevalsinvloeden en die zijn bij één persoon uiteraard veel groter dan in een groep. Maar het nadeel is dat duidelijke vooruitgangssignalen voor de behandelaar heel vaak niet worden teruggevonden bij deze individuele evaluatie met behulp van meetinstrumenten. En dat dus de informatie waarop bijstelling van de behandeling plaatsvindt, nog veel subjectiever is dan schaalscores van gevalideerde en genormeerde instumenten. Namelijk op basis van intuïtie van en subjectieve taxatie door de behandelaar.
Er is dus een procedure nodig die sensitiever is voor kleine verschillen, maar voldoende kritisch ten opzichte van verschillen die geen inhoudelijke of therapeutische betekenis hebben. De ontwikkeling van een methodologie waarin dit wordt nagestreefd ontkomt er niet aan zich met enkele fundamentele statistische concepten bezig te houden zoals de onafhankelijkheid van waarnemingen.

Deze verhandeling wil zo'n methodologie bieden. Deze werd geboren uit de behoefte sneller feedback te kunnen geven over het resultaat van individuele behandelingen in de intramurale, gesloten forensische psychiatrie, de justitiële jeugdzorg en de ambulante GGZ, de laatste met name de GGZ ten behoeve van mensen met autismespectrumstoornissen.
De ervaringen ermee zijn positief. Als we behandelaren vragen hun verwachtingen vooraf te specificeren en te 'deponeren' (laten vastleggen) en hen later de feedback van de resultaten te geven, levert dit steeds een patroon van verwachtingen op dat in grote lijnen overeenkomt met de uitslagen. En wordt hun specificering van resultaten in de vervolgfasen en bij andere behandelingen toenemend preciezer.
De versie die hier vóór U ligt is nog een eerste versie, bestemd om reacties te peilen. Om daarmee concepten verder te doordenken en uit te werken. En om aan de hand daarvan de methode in definitiever vorm te gieten en voor praktijkgebruik verder uit te werken. Commentaar is zeer welkom.

Arnold Bartels, Marinus Spreen, Erwin Schuringa en Vasthi Teeken,
Groningen: Forensisch-Psychiatrisch Centrum (FPC) Dr, S. van Mesdag;
Doorwerth: Dr. Leo Kannerhuis, Centrum voor Autisme
Utrecht: Expertisecentrum Forensische Psychiatrie (EFP),
9 juni 2008

I Probleemstelling en doel

1 Onderzoeksdesigns, interne en externe validiteit, beslissingsfouten, N=1

De methodologie en statistiek bij de evaluatie van vooruitgang van cliënten of patiënten in de GGZ na behandeling dient ervoor aan te kunnen geven of vooruitgang werkelijk vooruitgang is (en niet een gevolg van toevalsfluctuaties of spontaan herstel) en of vooruitgang werkelijk effect is (een effect is een vooruitgang die het gevolg is van de uitgevoerde interventie).
Dit wordt in overgrote meerderheid gedaan door vergelijkingen van groepen. Gegevens van de vóór- en nameting worden vergeleken, deze gegevens worden vergeleken met zulke gegevens van niet of anders behandelde controlegroepen.
Sinds het overzicht van Campbell en Stanley (1963; zie ook Shadisch, Cook & Campbell, 2002) worden de volgende vier onderzoeksdesigns onderscheiden (zijzelf onderscheiden pre-, quasi- en werkelijk experimentele designs).
Ten eerste het *niet-experimentele design*. Men onderzoekt verbanden in een steekproef. Bijvoorbeeld wat de kenmerken zijn van jongeren die zich agressief gedragen. Of wat de opvoedingskarakteristieken zijn voor mensen met autismespectrumstoornissen in vergelijking met mensen zonder deze stoornissen. Het design signaleert alleen optredende verbanden, maar staat zelf geen toetsen op causale verbanden toe. Overigens zijn er wel methoden ontwikkeld om in niet-experimentele designs enkele hypothesen over causale verbanden te toetsen.
Het tweede design is het *pre-experimentele design*. Dit onderzoekt of er verandering (vooruitgang) is van een vóórmeting naar een nameting. Of deze gebleken verandering het gevolg was van een eventueel uitgevoerde interventie, is er niet mee te meten.
Het derde designtype is het *quasi-experimentele design*. Men vergelijkt daarbij een interventiegroep met een vergelijkingsgroep zonder interventie of met een andere interventie. Interventie- en vergelijkingsgroep zijn zo vergelijkbaar mogelijk samengesteld, maar niet door random (willekeurige) toewijzing. Er zijn dus altijd checks nodig op de vergelijkbaarheid van de groepen en ondanks deze checks blijft het vaak de vraag of de groepen werkelijk in alle relevante opzichten vergelijkbaar waren.
Het vierde designtype is het *experimentele design,* ook wel 'random clinical trial' of 'random control trial' genoemd (RCT). Men heeft hier één of meer interventiegroepen en één of meer controlegroepen. Toewijzing van personen aan de groepen is random (willekeurig). RCT's zijn het meest 'harde' design en geven de meeste waarborg dat een gevonden verschil (effect) ook een werkelijk effect is. RCT's zijn de 'gouden standaard'.

Er zijn twee begrippen van belang bij de beoordeling van designs. Ten eerste de interne validiteit en ten tweede de externe validiteit (Campbell & Stanley, 1963; Shadish, Cook & Campbell, 2002).
De *interne validiteit* betreft de vraag of het design kan doen wat het beoogt. Als men alleen wil vaststellen of er vooruitgang is, maar men (nog) niet geïnteresseerd is in de vraag of deze vooruitgang effect is (het gevolg van een interventie), heeft een pre-experimenteel een design voldoende interne validiteit. Als men echter geïnteresseerd is in de vraag of een eventuele vooruitgang een werkelijk effect is, heeft een pre-experimenteel design onvoldoende interne validiteit. Een quasi-experimenteel design heeft dan meer interne validiteit, maar aangezien daarbij altijd de vraag open blijft of de groepen werkelijk vergelijkbaar waren, komt men uit op een RCT dat de hoogste interne validiteit heeft. RCT's (en soms gedegen quasi-experimentele designs) geven 'evidence based' resultaten.
De *externe validiteit* betreft de kwestie waarnaar men resultaten mag generaliseren. Stel men onderzoekt het effect van een combinatie van gedragstherapie en medicatie bij jeugdigen met

ADHD. Het onderzoeksdesign is een RCT. Men selecteert jeugdigen met ADHD bij wie men verwacht dat de te onderzoeken combinatie goed zal werken. Men verdeelt de jeugdigen random in een interventie- en een controlegroep. De interne validiteit van dit onderzoeksdesign is hoog. Het blijkt nu dat de interventiegroep duidelijk vooruit gaat, de controlegroep niet. Dan mag besloten worden tot effect van de combinatietherapie. Deze is 'evidence based'.
Maar vervolgens is de vraag aan de orde waarnaar men de resultaten mag generaliseren, wat de resultaten betekenen. Men heeft de combinatietherapie gegeven aan jeugdigen van wie men verwachtte dat ze er goed op zouden reageren. Zou de therapie ook werken bij de 'doorsneejeugdige met ADHD die naar de jeugd-GGZ verwezen wordt'? Dat is maar de vraag. Het vergelijken van kenmerken van jeugdigen uit het onderzoek met de genoemde 'doorsneejeugdigen' kan daar wellicht iets over verduidelijken, maar zal niet voldoende uitsluitsel geven. Het was dus achteraf beter geweest als ook jeugdigen met ADHD waren opgenomen in het onderzoek bij wie men minder goede verwachtingen over de therapie had. En als men dan binnen de interventie- en controlegroep ook gekeken had naar jeugdigen van wie men effect van de combinatietherapie verwachtte en naar jeugdigen bij wie dat niet het geval was.
Zonder voldoende interne validiteit weet men niet of er überhaupt sprake was van vooruitgang of effect. Maar de externe validiteit raakt het hart van het vak. Want die gaat over de vraag: wat betekenen de resultaten (van een intern valide onderzoek)? Als men zich alleen richt op de interne validiteit, wordt het vak uiteindelijk tot een verzameling van (intern valide) onderzochte procedures die zo precies mogelijk moeten worden nagevolgd, een verzameling gewaarmerkte procedures. Maar men weet onvoldoende waarom die procedures in omstandigheden die zoveel mogelijk lijken op de omstandigheden van het onderzoek, werken. Men kent niet goed de afzonderlijke werkzame principes die aan het effect ten grondslag liggen; vakinhoudelijk en theoretisch heeft men niet veel geleerd. Dit kan leiden tot behandelingen waarin cliënten behandelactiviteiten worden aangeboden, omdat deze onderdeel uitmaken van het 'evidence based' behandelprotocol, maar waarvan de behandelaar zich op inhoudelijk goede gronden afvraagt of zijn individuele cliënt er wel baat bij heeft (en zelfs of het in dit individuele geval wellicht gecontraïndiceerd is).

Vooral bij de interne validiteit spelen twee soorten statistische fouten die men zo klein mogelijk wil houden. Er is de *type I fout:* men concludeert ten onrechte tot een verband, tot een vooruitgang of effect. Men concludeert bijvoorbeeld tot een effect dat er in werkelijkheid evenwel niet is. Deze fout wil men zo klein mogelijk houden. In statistische procedures wordt de grootte van de type I fout bij vergelijkingen aangegeven met α. Bij $\alpha = 0.05$ is deze kans 5% ofwel 1 op 20. Deze $\alpha = 0.05$ is een gebruikelijke waarde. Soms $\alpha = 0.01$ genomen als men erg zeker wil zijn.
De *type II fout* is het omgekeerde: men concludeert dat er geen vooruitgang of effect is, maar in werkelijkheid (al weet men dat niet, want daarvoor is juist het onderzoek), is dat effect er wel. De type II fout wordt vaak aangegeven met β. Met deze β hangt de statistische 'power' van een design of onderzoeksprocedure samen, ook wel het *onderscheidingsvermogen* $(1 - \beta)$ genoemd: hoe groot is de kans dat een werkelijk effect onderkend wordt, dat de toetsingsprocedure een werkelijk effect onderscheidt?
Beide fouten hebben een relatie met elkaar. Als men de één verkleint, wordt doorgaans de ander vergroot. Als men de kans dat iemand ten onrechte veroordeeld wordt, erg klein wil houden en dus zeer strikte eisen stelt aan bewijs, dan zal de kans dat iemand wordt vrijgesproken terwijl hij in werkelijkheid wel de dader is, groter worden.
In de onderzoeksmethodologie wordt vooral gestreefd naar een kleine type I fout, meestal, zoals genoemd, 5% of 1%. Het wordt bedenkelijker geacht tot effect te besluiten als dat er niet is (type I fout), dan een werkelijk effect niet te onderkennen (type II fout). Want, zo wordt gedacht, bij later onderzoek zal een werkelijk effect wel naar voren komen. De kansen op een

type II fout zijn in veel onderzoek dan ook veel groter dan die voor de type I fout, en liggen vaak boven de 30% tot soms meer dan 80%! Het onderscheidingsvermogen (de kans een werkelijk bestaand effect te detecteren) is dan respectievelijk rond de 70% en 20%.
Men kan het onderscheidingsvermogen vergroten door

1. een grotere α (een grotere kans op type I fout, men bereikt dan sneller significantie),
2. door te proberen de variatie in scores te beperken (zo strikt mogelijke onderzoeks- en uitvoeringsprocedures, goede meetinstrumenten met een kleine standaardmeetfout) en
3. vooral door het aantal proefpersonen te verhogen. Dat aantal is sterk gerelateerd aan het onderscheidingsvermogen.

Overigens toetst men hypothesen door eerst een nulhypothese (H_0) te poneren van geen verschil of vooruitgang en deze eventueel te verwerpen ten gunste van een alternatieve hypothese H_1. H_1 is de hypothese waarin men geïnteresseerd is.
Stel men wil toetsen dat er na een behandeling een verbetering (verlaging) is van angstscores, dan is H_0 dat er geen verandering is (de beide gemiddelden vóór en na de therapie, m_1 en m_2, verschillen niet). Men bekijkt of men aan de hand van de data H_0 kan verwerpen ten gunste van H_1 dat er wel verschillen zijn.
Daarbij kan men H_0 ten onrechte verwerpen (en H_1 ten onrechte aannemen) met een kans van maximaal een gekozen α. Dit is de type I fout. Of men kan H_0 ten onrechte handhaven (en H_1 ten onrechte niet accepteren). Dat is de type II fout, β.

Box 1
Enkele technische kwesties: 1 hypothesen vooraf stellen, 2 meer hypothesen over dezelfde onderzoeksgegevens toetsen (kanskapitalisatie) en 3 het theoretische raamwerk waarin hypothesen passen

Vaak wordt vergeten dat hypothesen vooraf gesteld moeten worden. Significantieniveaus van bijvoorbeeld 5% gelden voor vooraf gestelde hypothesen. Blijken uit de resultaten verschillen die significant zouden zijn indien het verschil vooraf werd gehypothetiseerd, maar die niet gehypothetiseerd zijn, dan mag men die niet op de gebruikelijke wijze toetsen.

Een technische kwestie die van belang kan zijn betreft het aantal hypothesen dat men toetst over dezelfde onderzoeksgegevens. Als men één hypothese toetst bij $\alpha = 0.05$ is er 5% kans op een onterecht significant resultaat. Bij twee (onafhankelijke) hypothesen die men over dezelfde onderzoeksgegevens toetst met $\alpha = 0.05$, is de kans dat minstens één ervan ten onrechte een significant resultaat oplevert 10%.[2] Toetst men 20 (onafhankelijke) hypothesen over dezelfde onderzoeksgegevens dan is de kans dat minstens één daarvan ten onrechte significant is 64,2%. Dit verschijnsel wordt kanskapitalisatie genoemd.
De enige manier om eraan te ontkomen is voor elke hypothese een afzonderlijke steekproef (en dus onderzoek), wat praktisch niet realiseerbaar is. Er zijn procedures om de nadelige effecten van deze kanskapitalisatie te beperken. Ten eerste is dat door de hypothesen onderling in een goed theoretisch raamwerk in te bedden. Hoe beter en consistenter dat is, hoe beter onderzoeksresultaten (significante verschillen) geïnterpreteerd kunnen worden. Een tweede manier om kanskapitalisatie te verminderen is door op α een correctie toe te passen. Dit komt aan de orde in § 11.4.

Verwijderd: ¶

[2] De precieze waarde is 9,75%. De formule is: $1 (1 - \alpha)^k$. Daarin is k het aantal (onafhankelijke) hypothesen. Bij twee hypothesen is dat 9,75%; bij 20 is dat: $1 - (1 - 0{,}05)^{20} = 64{,}2\%$.

Dit alles gaat over groepen. Maar behandelingen in de GGZ zijn in overgrote meerderheid individueel. Het gaat om een individuele cliënt of patiënt en/of diens partner of gezin. Zelfs al neemt de cliënt deel aan een groepstherapie, gaat het om de individuele vooruitgang die er bij en voor hem bereikt wordt.
In de praktijk van de GGZ gaat het dus om evaluatie van individuele behandelingen zelfs als die in een groep gegeven worden. Het gaat om beslissingen van een behandelaar inzake het vervolg van een individuele behandeling: wordt bijvoorbeeld gekoerst op vergroten van vaardigheden, het onderkennen en bespreken van overtuigingen (cognities) of het versterken van de structuur van het gezinssysteem, op het eerst verminderen van angst of op het vergroten van het gedragsrepertoire en angst daarbij voorlopig accepteren?
Met psychometrische instrumenten en methoden die voor groepen voldoende sensitief zijn om vooruitgang en effect te signaleren (die voor groepen voldoende 'power' hebben, dus bij groepen een redelijke type II fout), komt men in het individuele geval vaak niet uit. Het onderscheidingsvermogen is dan te gering, dat wil zeggen: de type II fout te groot.
Eenzelfde scoreverschil van een pre- naar een postmeting dat bij een groep significant is, is dat bij één individu vaak niet. Dat is op zichzelf begrijpelijk, want in het individuele geval spelen toevalsinvloeden een veel grotere rol en de toets corrigeert daar als het ware voor. Dit gegeven vloeit voort uit het beschouwen van individuele verandering en vooruitgang vanuit de optiek van groepsvergelijkingen. De kern is: verandering wordt gezien tegen de achtergrond van groepsgegevens waarbij de groep veelal beschouwd wordt als steekproef uit de populatie. Het individu wordt daarbij niet op zichzelf bezien, niet als een populatie op zichzelf, niet als het statische 'universum' waarover men uitspraken wil doen, ongeacht wat die voor een bestaande of denkbeeldige groep of populatie betekenen.
De behandelaar kan als gevolg hiervan niet afgaan op de betekenis van kleine veranderingen bij metingen (bijvoorbeeld psychometrische instrumenten als vragenlijsten), want dat staan de statistische evaluatieprocedures die ontwikkeld werden voor groepsvergelijkingen niet toe. Dit leidt er dan wel toe dat beslissingen over het vervolg van de behandeling dus vooral of alleen plaatsvinden op basis van indrukken van de behandelaar en behandelinhoudelijke overwegingen. En niet op basis van de uitgevoerde metingen. Dus er blijft potentieel onbenut.

Er zijn echter designs voor individuele behandelingen, 'single case designs', ook wel N=1-studies geheten (N, het aantal subjecten in het onderzoek, is 1).
Historisch gezien begon het psychologisch onderzoek eigenlijk met N=1-designs ('single case designs') al werden die toen niet zo genoemd. Het eerste psychologisch onderzoek, van bijvoorbeeld Fechner, Ebbinghaus en Wundt in de 19e eeuw en Pavlovs conditionering berust voornamelijk of geheel op gegevens van één proefpersoon (vaak de onderzoeker zelf: Ebbinghaus deed geheugenexperimenten met zichzelf). Pavlov gebruikte voor enkele onderzoeken individuele honden. Om valide uitkomsten te krijgen, werd gewerkt met zeer grote aantallen observaties. Want de moderne statistiek, vooral variantieanalyse en t-toetsen, ontstond pas in de jaren dertig van de 20e eeuw. Die statistiek werkt, als aangegeven, vooral met groepen (die steekproeven zijn uit populaties) en maakt conclusies mogelijk over oorzakelijke factoren ('inferential statistics').
Als gevolg daarvan werd na de jaren dertig in onderzoeken vooral met groepen gewerkt en werden conclusies gebaseerd op de genoemde en verwante statistische procedures. Overigens bleven sommige onderzoekers deze methodologie die zich op statistische besluitvorming baseert, principieel afwijzen en ze bleven werken met onderzoeksgegevens van afzonderlijke personen (of dieren in het onderzoek). Een bekend voorbeeld hiervan is de theoreticus en onderzoeker op gebied van de operante leertheorie, Skinner (Kerlinger & Lee, 2000, p. 547). Onderzoek in de door hem geïnspireerde 'operant approach' geschiedt doorgaans nog steeds met individuele personen of dieren.

Dit leidde vanaf de jaren zestig van de vorige eeuw tot een N=1-methodologie ('single case experimental designs'), maar tot dusver bleef het gebruik ervan beperkt en werd de bewijskracht ervan (de interne validiteit) te gering gevonden.
N=1-studies hebben evenwel een groot potentieel. Waar in onderzoek over groepen gemiddeldes naar voren komen met spreidingen rondom de gemiddeldes (sommigen in de therapiegroep of de interventiegroep gaan vooruit, anderen niet of gaan zelfs achteruit), heeft met bij één behandeling te maken met gegevens of scores van één persoon op een bepaald moment. En heeft men te maken met de zeer individuele constellatie en configuratie van aanleg (constitutie), persoonlijkheid, verleden en de sociale context van die ene persoon, heeft men dus te maken met diens biopsychosociale werkelijkheid. De vraag waarom sommige deelnemers aan het onderzoek vooruitgaan en andere niet, is door behandelaars die de interventie kennen, vaak moeilijk te beantwoorden (het zijn 'achterafverklaringen') en is in deze individuele gevallen vanuit het (groeps)design doorgaans niet of hoogstens in de vorm van veronderstellingen te beantwoorden. Bij N=1-studies (N=1-behandelingen) kan men zeer specifieke hypothesen stellen, specifieker dan bij een groepsonderzoek omdat men alleen maar te maken heeft met die ene persoon.
Geleidelijk wordt wel wat vaker met N=1-studies gewerkt. Aan herhaalde N=1-studies die duidelijke vooruitgang laten zien, wordt zelfs een grote waarde toegekend door de American psychological Association, vergelijkbaar zelfs met RCT's (Task Force APA, 1995; Van Yperen en Bijl, 2006, p. 28)! Van Yperen en Bijl (2006) plaatsen herhaalde N=1-studies qua interne validiteit dan ook meteen na RCT's. Kerlinger en Lee (2000, p. 546) achten N=1-studies van gelijke interne validiteit als quasi-experimentele designs. Van Gageldonk en Bartels (1990, 1991) deden dit ook. De Beurs en Barendregt (2008, p. 47-58) nemen een ook door ons bepleite middenpositie in: N=1-studies als vóórstudie voor een RCT, een cumulatie of aggregatie van N=1-studies als 'next best' strategie.
Met een goede N=1-methodologie is dus veel winst te halen, zowel voor het evalueren van behandelingen als voor het plannen en ontwikkelen ervan. *De N=1-benadering dient dan als 'opmaat' voor een RCT. Een RCT blijft wenselijk,* het is de gouden standaard, maar het is altijd een behoorlijke investering. Door als voorbereiding al een aantal N=1-studies te hebben uitgevoerd, heeft men de kans op succes aanmerkelijk verhoogd.
Bovendien is het wenselijk bij of in een RCT de afzonderlijke interventies als N=1-studies uit te voeren als aanvullende gegevensbron. Men kan dan meer zeggen over individuele patronen van verandering.

Resumerend is het belang van N=1-studies dus het volgende.

1 Opmaat, voorbereiding, voor RCT.
  Onderdeel hiervan is:
  - 1.1 voorlopige indicatie voor de effectiviteit van de (individuele) interventies. En:
  - 1.2 de interventiestrategie (behandelstrategie) mede baseren op de tussentijdse metingen tijdens de N=1-studie. Men begrijpt dan beter waarom de interventie verlopen is zoals ze verliep.

2 Uitvoering tijdens RCT om extra gegevens te verschaffen voor de interpretatie van de RCT-resultaten.

De dataverwerking van N=1-data is thans echter nog steeds gestoeld op procedures die hun oorsprong vinden in het analyseren van groepsvergelijkingen en nog maar in beperkte mate toegespitst op werkelijk individuele gegevens. Dit betekent dat de type I fout laag blijft, maar de type II fout hoog, vaak erg hoog. En dus dat werkelijke vooruitgang, en de betekenis die deze heeft voor behandelbeslissingen en het ontwikkelen van behandelvormen, beperkt gesignaleerd wordt.

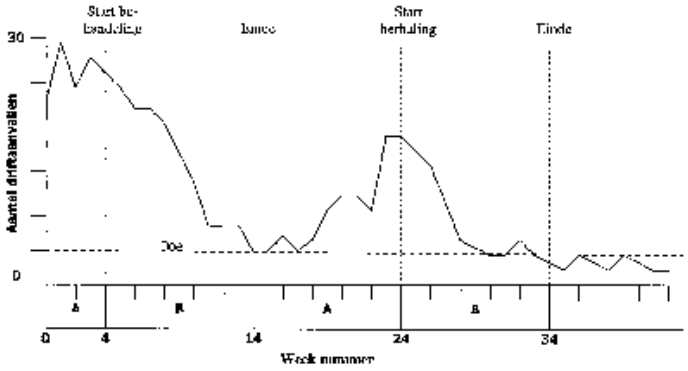
Deze verhandeling wil manieren van dataverwerking van N=1-data aangeven die sensitief zijn voor het individuele geval en bij de gebruikelijke type I fout voldoende power hebben (dus kleine type II fout) om er in het individuele geval behandelbeslissingen op te kunnen baseren en om op basis van de gegevens behandelvormen te kunnen ontwikkelen.

2 N=1, single case designs, ABA-designs

De Vereniging voor Gedrags- en Cognitieve therapie (VGCt) vraagt, als onderdeel van het traject naar het lidmaatschap, het schrijven van een N=1-studie over de behandeling van één patiënt/cliënt (of groep). Een N=1-studie is niet alleen een N=1-rapport, maar ook een N=1-behandeling. Want de behandeling moet vanaf het begin zodanig zijn opgezet dat het een N=1-studie kan worden (Hakenscheid, Kuipers & Marinkelle, 1998; Hermans, Eelen & Orlemans, 2007, p. 225-238). Gedragstherapie is leertheoretisch geïnspireerd en gebaseerd, en het vragen van een N=1-studie sluit nauw aan bij de door Skinner geïnspireerde operante leertheoretische benadering.
Gebruikelijk is om bij gebruikte meetinstrumenten in het kader van deze N=1, aan te geven waarom ze gekozen zijn, wat ze representeren en welke conclusies er aan vooruitgang of gelijk blijven verbonden zullen worden.
Wanneer over een reeks meetmomenten observaties worden uitgevoerd en in een grafiek uitgezet, zijn de gegevens vaak in aantal te gering voor statistische toetsen. En dus komt het aan op een beoordeling aan de hand van de grafiek in combinatie met het design, bijvoorbeeld een ABA-design. Dat is een design met een baseline-fase zonder behandeling A, een (eerste) interventiefase B, en een terugkeer tot A dan wel een met A erg vergelijkbare situatie A'. Deze laatste fase heet de *reversal-fase*. Daarna kan eventueel een interventiefase C volgen (of B'), wederom gevolgd door A, A' of A".
Als het gedrag tijdens B verandert in de gewenste richting en daarna bij A (of A') weer terugkeert tot de eerdere A-waarde, is dat een sterke bevestiging van het feit dat de interventie de oorzakelijke factor was. In een volgende fase B of B' wordt dan de interventie opnieuw ingezet en eventueel continu gemaakt (Hermans, Eelen & Orlemans, 2007, p. 225-238).
De vooruitgang in de B-fase (het verschil met de A-fasen) moet daarbij dermate duidelijk zijn dat statistische toetsen niet meer echt nodig zijn. Als de vooruitgang alleen met een statistische toets zou zijn vast te stellen, maar rechtstreeks niet goed zichtbaar is voor de betrokken behandelaars en cliënt(en) dan wordt de therapie toch als niet of minder geslaagd beschouwd. Aan een therapie tegen angst die blijkens scores op een vragenlijst en een statistische toets blijkt te werken, maar waaraan de cliënt geen verbetering ervaart, heeft cliënt noch therapeut iets.
In deze lijn zijn ook ABAC-, ABACA-, ABACADAB-designs etc. mogelijk (Hermans, Eelen & Orlemans, 2007, p. 225-238).
Dit sluit aan bij de op Skinner geïnspireerde operante benadering van de *Association for Applied Behavior Analysis* (AABA) en het tijdschrift *Journal of Applied Behavior Analysis* (JABA). Bij strikt operante programma's waarbij het gedrag bepaald blijft worden door de omgevingscontengenties (bij verandering daarvan, verandert het gedrag mee) is dit zinvol en uitvoerbaar (Bartels, 1993). In de praktijk van de GGZ evenwel zijn zulke strikte designs vaak niet goed bruikbaar, omdat een reversal-fase zich gewoonlijk niet goed verhoudt met de dynamiek van de behandeling (Swanborn, 1999, 2003). Veel cognitief gedragstherapeutische behandelvormen laten zich ook niet zo goed verenigen met een reversal-fase. Immers wanneer basiscognities als bijvoorbeeld kernovertuigingen en gehanteerde leefregels adequater geworden zijn, is een terugkeer naar de eerdere, inadequate niet wat men als therapeut graag ziet en ook vaak niet realistisch.
In de praktijk van de GGZ kan vaak wel met een 'quasi reversal-fase' worden gewerkt. De behandeling (fase B) laat bijvoorbeeld een duidelijke afname van het probleemgedrag zien en wordt afgesloten. De cliënt neemt de draad van zijn gewone leven weer op, een A'-fase in termen van het ABA-design. Het probleemgedrag blijft eerst nog enige tijd op een laag niveau, maar gaat vervolgens geleidelijk stijgen. De cliënt vervoegt zich weer bij zijn therapeut die dezelfde interventie als in de B-fase uitvoert, waarna het probleemgedrag terugkeert naar een

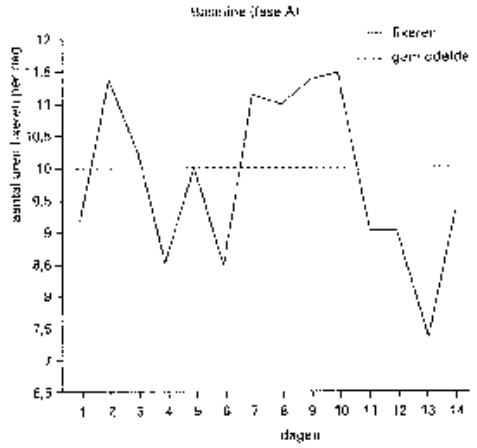
laag niveau, zelfs nog lager dan in de voorgaande B-fase. De cliënt pakt de draad van zijn leven weer op (tweede quasi reversal-fase, A") en nu stijgt het probleemgedrag iets, maar blijft op een laag niveau.
Een voorbeeld geven Van Yperen en Bijl (2006, p. 27), zie grafiek 1.

Grafiek 1: N=1-studie over individuele behandeling (naar Van Yperen & Bijl, 2006)

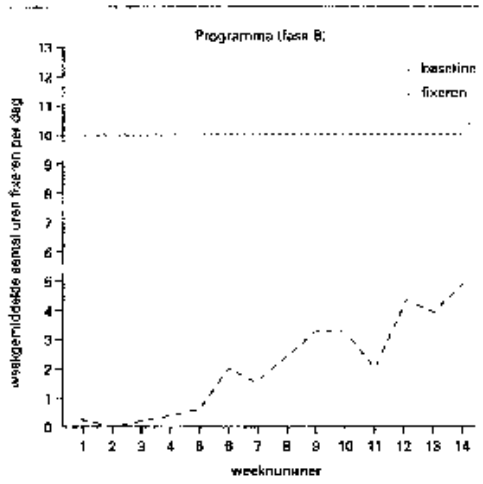


Een voorbeeld van een verstandelijk gehandicapte jongvolwassene die wegens ernstig agressief en destructief probleemgedrag al ruim vijf jaar circa 20 uur per dag op bed was vastgebonden, wat afnam tot alleen en lichter vastgebonden 's nachts, geeft Bartels (1993). De grafieken 2.1, 2.2 en 2.3 geven respectievelijk de baseline-fase (fase A), fase B ('interventie', bestaande uit een andere, ruimere, rustiger en wat verder van de groep afgelegen kamer voor de persoon; het probleemgedrag verdwijnt aanvankelijk vrijwel geheel maar neemt later steeds meer toe) en fase C (een gerichte operant-gedragstherapeutische benadering in de nieuwe kamer).

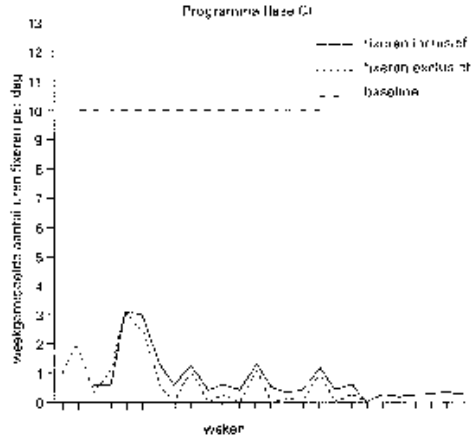
Grafiek 2.1: baseline-fase (fase A) verstandelijk gehandicapte man met ernstig probleemgedrag (agressie). Gegeven zijn de tijden dat hij overdag op bed was vastgebonden, per dag in uren.



Grafiek 2.2: Interventiefase 1 (fase B): nieuwe, ruimer en rustiger kamer
Gegeven zijn weer de tijden, nu weekgemiddeldes, van overdag fixeren (vastgebonden zijn) op bed, in uren



Grafiek 2.3: Interventiefase 2 (fase C): operant-gedragstherapeutisch programma
Gegeven worden de tijden (weekgemiddeldes) van overdag fixeren op bed in uren: stippellijn (het fixeren exclusief, dunne onderbroken lijn). Toen dit minder of niet meer nodig was werd een alternatieve gedragsmaat gevonden bestaand uit fixeren op bed op eigen verzoek, time-out (op eigen kamer of in de douche) en 10 minuten voor elke keer dat hij wegens gespannenheid en dreiging moest douchen met gesloten deur (fixeren inclusief, dikkere lijn).



N=1-studies volgens de JABA-traditie zijn er al sinds de jaren zestig van de vorige eeuw (Hersen & Barlow, 1976; Hermans, Eelen & Orlemans, 2007, p. 225-238; Kazdin, 1981; zie ook Van Yperen & Bijl, 2006, p. 27-29).
De waarde van dit soort N=1-studies wordt, zoals aangegeven in hoofdstuk 1, vaak onderschat. Van Yperen en Bijl (2006, p. 28) schrijven daarover: 'Bij een voldoende aantal casestudies op een rij (meer dan acht), vormt deze bewijsvoering volgens de American Psychological

Association (APA) een volwaardig alternatief voor het uitvoeren van een aantal randomized clinical trials' en ze geven verwijzingen (onder andere Task Force APA, 1995)!
Zijzelf zien hierin reden herhaalde N=1-studies qua interne validiteit direct achter RCT's te plaatsen. De door ons bepleite en ook door De Beurs en Barendregt (2008, p. 47-58) onderschreven strategie is N=1-studies als aanloop naar een RCT te gebruiken. Een RCT vraagt veel investering en met herhaalde N=1-studies als voorbereiding is het rendement in termen van resultaat gegeven de investering, optimaal.
Men zal daarbij N=1-designs wel breder moeten maken dan het ABA-design. Want met het ABA-design kan men in de praktijk van de GGZ meestal niet uit de voeten. Juist niet wanneer een therapie zo succesvol is dat in de (eerste) B-fase een aanzienlijke verbetering wordt bereikt en de behandeling wordt afgesloten. Er is dan mogelijk wel een quasi reversal-fase, maar er vindt geen terugval van gedrag plaats waarna een nieuwe B-fase nodig is, een nieuwe B-fase die de werkzaamheid van de therapie daarmee duidelijk toont. Bij geslaagde (praktijk)therapieën is er geen tweede B-fase. En juist bij geslaagde therapieën blijft dus methodologisch de vraag open of de therapie echt de oorzakelijke factor was in de verbetering.

Zoals al aangegeven komt het in de GGZ heel veel voor dat men niet terug kan of wil naar het oorspronkelijke gedrag in een reversal-fase. Vandaar dat dan gezocht wordt naar andere manieren van evaluatie die het causale verband tussen de behandeling (interventie) en opgetreden vooruitgang aannemelijk moet maken.
Eén wezenlijk verschil met de behoefte individuele vooruitgang te evalueren bij veel cognitieve gedragstherapieën (of andere interventies) in vergelijking met het ABA-design is dus dat men niet terug wil of kan naar het oorspronkelijke gedrag in een reversal-fase.
Een tweede verschil is dat in ABA-designs vaak gewerkt wordt met 'harde' gedragsobservatiematen als bijvoorbeeld aantal agressie-incidenten, tijd dat iemand onder speciale controle staat, tijd doorgebracht met vrienden, aantal dagen dat de angst niet boven een bepaald niveau komt etc. Bartels (1993, grafieken 2) werkte met de tijd dat de persoon wegens ernstige agressie overdag op bed was vastgebonden. Dit soort maten hebben veel voordelen. Men kan zulke maten verhoudingsgewijs tamelijk 'hard' kan maken, met een relatief geringe foutmarge. Als elke dag wordt bijgehouden hoe lang de persoon wegens agressie op bed wordt vastgebonden, kan men bij dagtotalen zich wel eens een paar minuten of een kwartier vergissen, maar dat maakt op een totaal van zo'n zo'n 20 uur per etmaal weinig verschil.
Het aantal waarnemingen is beperkt, maar doorgaans voldoende. Er is bij de genoemde agressiemaat van op bed vastbinden één score per dag: de totaaltijd dat de persoon was vastgebonden. Wil men dan statistische vergelijkingen maken tussen fasen dan zijn er minimaal enkele weken per fase nodig. Nu werd al aangegeven dat het klassieke ABA-design het niet moet hebben van statistische toetsen, maar dat de verbetering op het oog evident moet zijn. Voor het klassieke ABA-design is het relatief gezien geringe aantal waarnemingen daarom minder een probleem.
Maar als men in de praktijk van de GGZ met een quasi reversal-fase werkt, kan het relatief geringe aantal waarnemingen wel een probleem worden. Bovendien heeft men in de GGZ doorgaans dit soort 'harde' gegevens niet, maar heeft men scores van cliënten of patiënten op schalen van psychometrische instrumenten. Zulke schaalscores hebben meetfouten. Bij het bijvoorbeeld bekende instrument de SCL90 (Arrindell & Ettema, 2003; Arrindell, Boosma, Ettema & Stewart, 2004) heeft bij elke afname de antwoorden op 90 vragen en scores op 10 schalen. Stel men neemt de SCL90 driemaandelijks af en men heeft daarmee dan driemaandelijks tien schaalscores. Dat is een beperkt aantal gegevens. De vraag is wat kan men daarmee? Welke vooruitgang op schaalscores van een individu kan als een waarschijnlijk werkelijke vooruitgang gezien worden?

Om zonder reversal-fase op basis van psychometrische instrumenten een individuele behandeling sensitief en nauwkeurig te evalueren zijn de hierna te beschrijven N=1-evaluatiemethoden ontwikkeld.

3 Welke verandering, welk verschil of vooruitgang is betekenisvol?

De vraag welke vooruitgang op schaalscores van een individu als werkelijke vooruitgang gezien kan worden is om drie, onderling gerelateerde redenen van belang.
Ten eerste is het van belang bij behandelevaluatie om te kunnen bepalen of er in werkelijkheid (waarschijnlijk) vooruitgang is.
Ten tweede is het van belang om, tijdens een lopende behandeling, individuele vooruitgang te kunnen taxeren om te bepalen wat bij die individuele therapie, gezien de resultaten tot dusver, het beste vervolg is. Men formuleert de behandelstrategie bijvoorbeeld deels in termen van de gebruikte meet- en evaluatie-instrumenten. De aanvangsklachten zijn bijvoorbeeld angst en depressieve belevingen (SCL90-schalen). De behandeling focust eerst op insufficiëntie van denken en handelen (ook een SCL90-schaal); men verwacht dat daarop eerst een verbetering zich zal manifesteren, en pas later op angst en depressie. Geschiedt dit en in de aangegeven volgorde, dan is dat een bevestiging van de individuele behandeltheorie.
Ten derde is de vraag of een vooruitgang betekenisvol is, zeer relevant bij de ontwikkeling van nieuwe methoden. Zolang men daarbij voor de evaluatie afhankelijk blijft van vergelijking van groepen, bijvoorbeeld een interventie- en een controlegroep, komt de feedback of een methode waarschijnlijk werkt, vaak laat, pas na een onderzoek dat doorgaans behoorlijk veel tijd kost. Men is dan gebaat bij een nauwkeurige methode die in het individuele geval kan aangeven of een methode waarschijnlijk gewerkt heeft of niet. Aan de hand daarvan stelt men de methode bij en ontwikkelt deze verder. Vervolgens wil men deze individuele evaluaties sommeren en als volgende stap kan een experimenteel onderzoek worden uitgevoerd, een random clinical (control) trial (RCT) als uiteindelijke check.

Voor het tweede en derde doel zijn individuele vooruitgangsmetingen nodig die erg sensitief zijn, voldoende power hebben en dus een kleine type II fout. Ze moeten in staat zijn kleine verschillen op betekenis te kunnen beoordelen, want in een individuele therapie en bij de ontwikkeling van een methode is vooruitgang, zeker in het begin, doorgaans klein. Om zulke kleine verbetering vast te stellen, zijn sensitieve instrumenten en methoden nodig. Een instrument en evaluatiemethode die voldoen bij groepsvergelijkingen, waar relatief kleine verschillen significant kunnen zijn (niet te grote type II fout), zijn vaak onvoldoende sensitief voor de kleine verschillen bij individuele behandelingen.
Therapeuten gaan bij de beslissing over de te volgen strategie te werk volgens een individuele behandeltheorie zoals de gedragstherapeutische functieanalyse. Bevestiging van die theorie geschiedt door wat de cliënt vertelt over gebeurtenissen die zich na het opstellen van de behandeltheorie voordeden en die aldus tot bevestiging of weerlegging van de theorie kunnen leiden. Die gebeurtenissen zijn soms kwantitatief te registreren, maar meestal worden ze kwalitatief beoordeeld. Een cliënt was bijvoorbeeld agressief nadat hij zich veronachtzaamd voelde (tevoren is een relatie tussen insufficiëntiegevoelens, zelfbeeld, impulsiviteit en agressieregulatie geformuleerd). Dit wordt gezien als bevestiging van de functieanalyse. De neiging tot agressief reageren zal verminderen als de cliënt minder insufficiëntiegevoelens zal hebben en meer het gevoel dat hij invloed heeft op de meeste gebeurtenissen in zijn leven. Voor de beoordeling van de vervolgstrategie van de behandeling inclusief de bevestiging van de behandeltheorie, moeten het instrument en de methode voldoende sensitief zijn om verbetering van insufficiëntie van denken en handelen bij de individuele cliënt te kunnen vaststellen.
De ontwikkeling van gedifferentieerde behandeltheorieën is beter mogelijk met de feedback van gegevens van individuele behandelingen. Zonder die feedback blijft men afhankelijk van de gegevens van groepsvergelijkingen in RCT's die veel later beschikbaar komen en ook minder geïndividualiseerd zijn. Hier wordt het woord behandeltheorie gebruikt (Van Gageldonck

& Bartels, 1991). Er circuleren verschillende benamingen: praktijktheorie, N=1-theorie (Van Strien, 1986), impact theory (Rossi, Lipsey & Freeman, 2004), behandelingstheorieën (Veerman, Damen & Ten Brink, 2004), interventietheorie (Van Yperen, Bijl & Veerman, 2006).

De meerderheid van getalsmatige, statistische evaluaties bestaat uit het vergelijken en doen van analyses over groepsgegevens, zoals in § 1 en § 2 werd aangegeven. Groepen worden vergeleken en binnen of over groepen wordt gekeken naar trends en verbanden. Het instrumentarium om individuele analyses te doen, analyses over het verloop van één individuele behandeling of een op één individu gerichte interventie, is veel beperkter.
Op de vraag hoe schaalscoreverschillen van één individu te waarderen is op verschillende manieren (vanuit analyses voor groepen) geantwoord, maar aan elke manier kleven veel bezwaren. Deze verhandeling wil de voornaamste manieren en beperkingen daarvan bespreken en met een goed werkbare alternatieven komen. Een en ander zal worden geadstrueerd aan de hand van twee voorbeelden.
Ten eerste een N=1-studie die door de vierde auteur (de eerste auteur was supervisor) in het kader van het traject naar het lidmaatschap van de VGCt, inmiddels gerealiseerd, werd ondernomen. Het gaat daarbij om zelfrapportage-gegevens (vier invullingen door de cliënt van de SCL90 met steeds drie maanden tussentijd). Betrouwbaarheidsvragen liggen hier vooral rond de betrouwbaarheid van individuele item- en schaalscores en de conclusies die aan scoreverschillen kunnen worden toegekend. Deze kwestie van voortgangsmetingen aan de hand van zelfrapportage (maar één meting per gebeurtenis) gaan de hoofdstukken 4 t/m 17. Daar worden drie methoden van voortgangsevaluatie.besproken. De derde daarvan, wordt in § 18.2 en in de bijlage daarbij nader toegelicht.
Het tweede voorbeeld betreft waarderingen van het gedrag van een forensisch-psychiatrische patiënt door sociotherapeuten. De betrouwbaarheidsvragen liggen hier vooral op verschillen tussen verschillende beoordelaars. Een schets daarvan wordt gegeven in hoofdstuk 18.

II Hoe individuele verandering (vooruitgang) te beoordelen?

4 Normklassen

De meest gebruikte manier om individuele scoreverschillen te beoordelen is het werken met klassen als een 'klinische range'. Het gaat er dan om of een schaalscore van een individu vanuit de 'klinische' range naar de 'normale range' is gegaan of van bijvoorbeeld de klasse 'bovengemiddeld' (problematisch) naar 'gemiddeld'.
Het kan daarbij gebeuren dat een grote vooruitgang binnen dezelfde klasse niet als vooruitgang wordt geduid en een kleine vooruitgang net over een klassegrens heen, wel als vooruitgang wordt gezien. Bovendien zijn de klassen vaak gespecificeerd per normgroep. Er is dan bijvoorbeeld de normale populatie als normgroep (met daarbinnen bijvoorbeeld de klassen hoog, bovengemiddeld, gemiddeld etc.) en er is de populatie psychiatrische patiënten met idem klassen erbinnen. Waar vergelijkt men schaalscores van een individu dan mee? Het bekende instrument de SCL90 (Arrindell & Ettema, 2003; Arrindell, Boosma, Ettema & Stewart, 2004) heeft *zes* normgroepen.
Bij normgroepen vergelijkt men een individu bovendien met een normgroep en niet met zichzelf. *De voornaamste vraag is echter niet: gaat deze persoon vooruit in vergelijking met de normgroep, maar: gaat de persoon vooruit bij vergelijking met zichzelf?*

5 Standaardmeetfout en Reliable Change Index (RCI)

Een in principe betere manier om verschillen vóór en na een behandeling te evalueren is werken met de standaardmeetfout of, enigszins vergelijkbaar met en mede gebaseerd op de standaardmeetfout, de *Reliable Change Index,* RCI (Drenth & Sijtsma, 2006, p. 190-245; Maassen, 2003; Veerman, 2006). Er zijn overigens verschillende RCI's, hier houden we de meest gangbare aan. Het is een betere manier, omdat uitgegaan wordt van bij het individu geconstateerde vooruitgang die niet gerelateerd wordt aan normklassen.
De nadelen ervan zijn echter ten eerste dat er om behoorlijke verschillen gevraagd wordt, willen ze significant zijn bij de gangbare α-niveau van 0.05. Een verschil van vóór- naar náméting dat bij een groep zonder problemen significant is, is dat bij één individu vaak niet (de statistische power is beperkt). En ten tweede dat men ook nu afhankelijk blijft van gegevens van groepsgegevens uit test- of vragenlijstconstructie zoals betrouwbaarheidscoëfficiënten en standaarddeviaties.

Als er een gerichte hypothese is, bijvoorbeeld de score van een cliënt in behandeling zal verbeteren, dan toetst men éénzijdig: er wordt een specifiek verschil verwacht. *Werken met de standaardmeetfout* komt in dat geval neer op 1,645 maal de standaardmeetfout nemen tussen twee individuele scoreverschillen (éénzijdig toetsen). De kans op het ten onrechte besluiten tot een verbetering (besluiten tot verbetering terwijl die er in werkelijkheid niet is), is dan maximaal 5% ($\alpha = 0.05$), de type I fout. Men heeft dan 95% kans dat een gevonden verbetering werkelijk een verbetering is.
Is er geen specifieke hypothese, maar alleen de verwachting van een verschil, ongeacht of de tweede score hoger of lager is, dan moet 1,96 maal de standaardmeetfout genomen worden (tweezijdig toetsen).
Een testscore bestaat uit een component een 'ware score' en een component 'error-score'. Elke testafname wordt beïnvloed door toevalsinvloeden, in elk geval variabele niet-systematische invloeden. Bijvoorbeeld vermoeidheid van de onderzochte persoon, verlichting, duide-

lijkheid van de instructies etc. De standaardmeetfout (standard error of measurement, S_e) is een schatter van deze toevalsinvloeden. De toevalsinvloeden rondom de (hypothetische) ware score zorgen voor een normaalverdeling van waarden rondom deze ware score. Binnen het interval van $+1{,}96S_e$ en $-1{,}96S_e$ ligt, met 95% waarschijnlijkheid, de ware score. Stel voor een intelligentietest is de S_e genormaliseerd op 6 als meetfout. De werkelijke score, het werkelijke IQ, van een geteste persoon ligt dus tussen +11,76 en –11,76 (afronden op 12) van zijn gebleken score.
Twee ware scores verschillen werkelijk als het verschil tussen twee gebleken scores (IQ's) tenminste 12 bedraagt (= $1{,}96S_e$ afgerond).
Voor bepaling van de S_e neemt men als uitgangspunt een betrouwbaarheidscoëfficient (Veerman, 2006, p. 82), bijvoorbeeld de test-hertestbetrouwbaarheid r_{xx} (eventueel een gemiddelde van enkele bekende betrouwbaarheidscoëfficiënten) en de standaarddeviatie (SD), eventueel het gemiddelde van enkele bekende SD's. Men berekent de standaardmeetfout S_e als volgt. $S_e = SD\sqrt{(1 - r_{xx})}$.
Verbetert voor een persoon bijvoorbeeld de score op de angstschaal van de SCL90 van 21 naar 17 (de schaal heeft 10 items waarop de antwoorden kunnen variëren van 1 t/m 5, schaalwaarden kunnen dus variëren van 10 t/m 50), is dat dan een wezenlijke vooruitgang? $S_e = 6{,}07$ zoals te berekenen is uit de tabellen van Arrindell & Ettema (2003, p. 35). Een verschil moet dus $1{,}645 \times 6{,}07 = 9{,}99$ groot zijn, bij gerichte verwachting, om met 95% waarschijnlijkheid tot een werkelijk verschil te besluiten, dat wil zeggen om met 95% waarschijnlijkheid te besluiten dat het gevonden verschil een werkelijk verschil is. Een waarde van 4 ($21 - 17 = 4$) is daarvoor te klein.

Voor de *Reliable Change Index* RCI berekent men, op basis van standaardmeetfout S_e, de standaardmeetfout van het verschil tussen twee (schaal)scores: $S_{diff} = \sqrt{[2(S_e)^2]}$. $RCI = (VMS - NMS) / S_{diff}$ (Veerman, 2006, p. 82-83). Daarin is VMS de aanvangsscore (voormeting) van een persoon op een schaal en en NMS de eindscore (nametingsscore). De RCI wordt verondersteld statistisch normaal verdeeld te zijn met gemiddelde 0 en SD = 1.[3] In plaats van S_{diff} wordt ook wel het symbool $S_{(e)diff}$ gebruikt.
Een positieve RCI duidt op vooruitgang (als geldt: hoe hoger de schaalscore, hoe ongunstiger). Een RCI is te interpreteren als een z-waarde. Dat wil zeggen dat een waarde van 1,96 significant is op niveau $\alpha = 0.05$, bij tweezijdig toetsen (alleen voorspelling van verschil, niet van een specifiek verschil); bij éénzijdig toetsen (men had tevoren aangegeven dat men op die schaal vooruitgang verwachtte) moet een RCI tenminste 1,645 zijn.[4]
Verbetert voor een persoon bijvoorbeeld de score op de angstschaal van de SCL90 van 21 naar 17 (de schaal heeft, als al aangegeven, 10 items waarop de antwoorden kunnen variëren van 1 t/m 5, schaalwaarden kunnen dus variëren van 10 t/m 50), is dat dan een wezenlijke vooruitgang? $VMS - NMS = 21 - 17 = 4$. S_{diff} komt op 8,58 zoals is te berekenen uit de tabellen in Arrindell & Ettema (2003, p. 35). RCI komt daarmee op 0,47, een te geringe waarde om tot werkelijke vooruitgang te besluiten; 0,47 is duidelijk minder dan de genoemde 1,645 en blijft zelfs

---

[3] De procedure die Veerman (2006, p. 82-83) biedt, wordt gevolgd. Het gaat hier om de standaardmeetfout van een verschilscore, een verschil tussen twee scores resulterend uit verschillende gelegenheden, bij dezelfde persoon. Lange tijd, sinds 1970 toen een kritisch artikel van Cronbach en Furby verscheen over werken met verschilscores ('gain scores'), hebben deze een dubieuze reputatie gehad. Men zou ze niet mogen vergelijken met verschilscores tussen twee onafhankelijke personen, waarop de vergelijking wel gebaseerd was. Williams en Zimmerman (1996) maken aannemelijk dat met 'gain scores' wel op die manier gewerkt mag en kan worden.

[4] Men toetst éénzijdig als er een specifieke hypothese is, bijvoorbeeld test B is hoger dan test A. Men toetst tweezijdig zonder specifieke hypothese, maar alleen de vraag of er überhaupt sprake is van een verschil. Bijvoorbeeld: $A \neq B$.

binnen één standaardmeetfout van het verschilgegevens van de normgroep(en) uit de vragenlijstconstructie.
De RCI is iets kritischer dan de procedure met de standaardmeetfout. Dit komt omdat de RCI gebaseerd is op de standaardmeetfout van het verschil tussen twee metingen. Die standaardmeetfout is groter dan de standaardmeetfouten van de oorspronkelijke scores. Gewoonlijk is $S_{(e)diff}$ iets groter dan S_e aangezien een verschilscore te maken heeft met de S_e van de twee scores die bij de berekening gebruikt worden. De RCI berust dus op een iets ander principe, is niet beter of slechter, maar wel iets robuster. Met als nadeel een grotere β. Zie voor procedures ook Hafkenscheid, Kuipers en Marinkelle (1998), en Yperen en Bijl (2006).

6 Itemscores als afzonderlijke waarnemingen (bij zelfrapportage-instrumenten)

Gevraagd dus: een sensitieve, nauwkeurige evaluatiemethode voor individuele vooruitgang bij behandelvoortgangsevaluatie zoals zelfrapportagegegevens. De methode moet zo min mogelijk gebruik hoeven maken van bestaande normgegevens van de test- of vragenlijstconstructie. Men krijgt hiermee in een vroeg stadium van het ontwikkelen van een behandelmethode feedback inzake de waarschijnlijke werkzaamheid van de behandeling. Bij individuele behandelingen moeten de gegevens bruikbaar zijn voor het formuleren van de behandel- of interventietheorie voor de individuele cliënt of patiënt.
De oplossing werd gezocht in het beschouwen van afzonderlijke itemscores als waarnemingen. De invullingen van de 10 items van de angstschaal van de SCL90 bijvoorbeeld zijn dan 10 waarnemingen. Over de uitslagen van die waarnemingen kunnen de standaard statistische toetsen (parametrisch of non-parametrisch) worden gedaan.
Uiteraard is de vraag aan de orde in hoeverre het gerechtvaardigd is afzonderlijke itemscores van één persoon, herhaalde waarnemingen (verschillende metingen) als afzonderlijke waarnemingen te beschouwen. Die vraag wordt deels hierna behandeld en verder in hoofdstuk 11.
Het beschouwen van de afzonderlijke itemscores als waarnemingen vloeit voort uit het bezien van items vanuit de toetsingsvraag of de itemscores veranderd zijn of niet. Daarbij is de persoon de (hele) 'populatie'.
Bijvoorbeeld bij SCL90-schaal *2*, *Angst* (10 items), elk te scoren van 1 (geen probleem) naar 5 (veel problemen). Hierna in tabel 1 de items en de scores bij meting 1 (M1) en meting 2 (M2).

Tabel 1.1: itemscores SCL90-schaal *Angst* bij 2 metingen

SCL90-schaal *Angst*		**Meting**	
Item	**Inhoud**	**M1**	**M2**
2	Zenuwachtigheid, van binnen trillen	3	2
17	Trillen	2	2
23	Plotseling schrikken of zomaar bang worden	2	2
33	Je bang voelen	2	2
39	Hartkloppingen	2	2
57	Je gespannen voelen	3	2
72	Aanvallen van angst of paniek	3	2
78	Rusteloos voelen, niet stil kunnen blijven zitten	1	1
80	Gevoel dat je iets naars gaat overkomen.	1	1
86	Gedachten en voorstellingen van angstige aard	2	1
10 items	Totaal	21	17

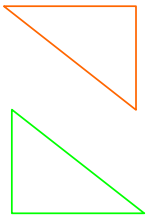
Deze tabel kan worden omgezet in de volgende matrix. In de linker kolom de 5 scoremogelijkheden per item (1 = geen problemen, 5 = veel problemen) bij M1. De bovenste rij geeft de scoremogelijkheden bij M2.

Er zijn 10 items. Gelijk blijven er 6 (4 van 2 naar 2, 2 van 1 naar 1): de diagonaal.Vooruitgang (driehoek links onder) is er op 4 items (1 van 2 naar 1, 3 van 3 naar 2). Als de items bij de tweede meting M2 alle gelijk zijn aan de eerste meting M1 dan liggen ze alle op de diagonaal.

Tabel 1.2: veranderingen van itemscores SCL90-schaal *Angst* bij 2 metingen

Waarden itemscores bij M2

Waarden itemscores bij M1	**1**	**2**	**3**	**4**	**5**
1	2				
2	1	4			
3		3			
4					
5					



Driehoek met items waarop achteruitgang naar voren komt.

Driehoek met items waarop vooruitgang naar voren komt.

Als er niets verandert, als beide metingen dezelfde uitslagen geven, liggen alle waarden op de diagonaal. Wat men hier in feite wil toetsen is de kans dat een afwijking van de waarden op de blauwe diagonaal naar de driehoek waarin sprake is van afname (vooruitgang) significant is. Anders gezegd: de nulhypothese is dat verandering van scores gelijk verdeeld is over vooruitgang (groene driekhoek) en achteruitgang (rode driehoek).
De verdeling van zo'n matrix valt te toetsen, met een McNemar-toets bijvoorbeeld. Ook pure kansberekening is mogelijk (randomisatietoets).
Maar men kan hetzelfde doen met het vergelijken van de afzonderlijke scores met een parametrische toets als de t-toets voor gepaarde waarnemingen of een niet-parametrische toets als de Wilcoxon.

In het vervolg nu een behandelevaluatiecasus waarin deze methode vergeleken wordt met het werken met normgroepen, de standaardmeetfout en de RCI. Daarna op basis van de methode om afzonderlijke itemscores als waarnemingen te zien, een uitwerking van deze gedachtengang.

III Casus

7 Casus: cliënt en (cognitieve gedrags)therapie

Aan de hand van een voorbeeld wordt de methode hierna uitgewerkt. In het voorbeeld gaat het om een 34-jarige man een autismespectrumstoornis. De diagnose was vóór de gedragstherapie nog niet eerder gesteld. Enkele details zijn omwille van privacy-redenen aangepast.
Cliënt heeft met veel hulp van zijn ouders en broer de HAVO gehaald. Hij woont alleen, heeft steunende contacten met beide ouders en een familielid dat veel voor hem betekent. Hij heeft ook veel ondersteuning uit al jaren bestaande contacten met een psychotherapeut (die overigens de diagnose ASS nooit opperde). Vanwege slaapproblemen, sterke angsten en spanningsklachten, zich depressief voelen, vaardigheidstekorten en planningsproblemen wordt hij aangemeld voor gedragstherapie, specifiek voor deze problemen. De lopende psychotherapie blijft doorgaan. De gedragstherapie komt neer op wekelijkse sessies van circa 1½ uur inclusief pauze en enkele telefonische, steunende contacten per week. De therapie duurt aldus tien maanden, waarna de frequentie verlaagd wordt naar tweewekelijks en maandelijks, en tot slotte beëindigd werd.
De gedragstherapie werd primair gericht op de slaapproblemen en spanningsklachten, maar ook op de andere genoemde problemen.
De diagnose autismespectrumstoornis was nog niet eerder gesteld en geeft cliënt de eerste anderhalve maand juist méér slaap- en spanningsklachten. De vragen over zijn slapen leiden de eerste twee maanden tot méér piekeren over slapen. Ook ervaart hij de therapiebijeenkomsten als nuttig maar wel stressvol. Hij wil het perfect doen, denkt lang na over zeer gedetailleerde antwoorden en komt er ook vaak op terug om ze te nuanceren of herformuleren. Hij ervaart de wereld als een verwarrende chaos waartegen zijn remedie is streven naar perfectie en controle. Voorwerpen in zijn woning hebben een vaste plaats, tot op de centimeter nauwkeurig.
Cliënt wil een beroemd schaker worden en verwacht dat zijn problemen dan over zijn aangezien men hem dan met alles wat hij nodig heeft, zal helpen. De behandelstrategie wordt mede geformuleerd in termen van de SCL90-schalen aangezien hij de SCL90 om de drie maanden invult. Cliënt gebruikt geen medicatie.
De gedragstherapie verloopt de eerste drie maanden redelijk voorspoedig. Daarna eigenlijk ook, maar in het laatste half jaar treden enkele erg stressvolle gebeurtenissen op. Het familielid dat veel voor hem betekent, wordt ernstig ziek en overlijdt; hij moet ook verhuizen.

8 Meetinstrument: de SCL90

Aan het begin van de therapie en telkens na drie maanden vulde de cliënt een zelfrapportagevragenlijst in over zijn psychisch functioneren (SCL90). De SCL90 heeft als schalen (Arrindell & Ettema, 2003; Arrindell, Boosma, Ettema & Stewart, 2004):

1. agorafobie,
2. angst,
3. depressie,
4. insufficiëntie van denken & handelen,
5. somatisch beleefde problemen,
6. wantrouwen & interpersoonlijke sensitiviteit,
7. vijandigheid (hostiliteit),
8. slaapproblemen,
9. overige problemen en

10 totaal.
Elke schaal bestaat uit een (per schaal variërend) aantal items (vragen). De antwoorden kunnen elk variëren van waarden 1 (totaal geen probleem) naar waarde 5 (veel problemen). Hoe hoger de score, hoe problematischer dus. De gemiddelde itemscores per schaal zijn dus qua hoogte onderling vergelijkbaar (1 als minimum, 5 als maximum).
Schaal 9 bevat items die niet konden worden ingedeeld op grond van de factoranalyses. Het zijn items die vooral psychotische kenmerken betreffen; ze komen uit de dimensie psychoticisme van de oorspronkelijke, Amerikaanse SCL90. Maar de Nederlandse factoranalyses leverden niet zo'n dimensie op (Arrindell & Ettema, 2003, p. 20).

9 Verwachtingen inzake resultaat, hypothesen

Conform de doelstelling van de therapie wordt primair verbetering nagestreefd inzake de SCL90-schalen 8 en ook inzake 2, 3 en 4. Dit moet mede resulteren in een verbetering van de totaalscore (schaal 10).[5] Op deze schalen wordt een significante verbetering verwacht. De behandelstrategie wordt zoveel mogelijk verwoord in termen van de te meten schalen. De gedragstherapeutische functieanalyse, betekenisanalyse en holistische theorie werden daartoe zoveel mogelijk geherformuleerd in termen van de SCL90-schalen.
De therapie richtte zich op

1 psycho-educatie (uitleg over autismespectrumstoornissen),
2 onderkennen en verbeteren van zelfredzaamheids- en sociale vaardigheden,
3 herkennen van belemmerende gedragspatronen (zoals streven naar perfectie) en cognities zowel inzake slapen als inzake het dagelijks leven in het algemeen
4 hanteren van angst en spanning (gedrag, ontspanning, cognities)
5 slaapeducatie.

Verbetering van slapen wordt verwacht van de slaapeducatie, maar ook indirect als gevolg van de therapeutische ingrediënten 1 t/m 4. Dit alles betekent dat verbetering veerwacht wordt op de genoemde schalen.

[5] De oorspronkelijke hypothese en verwachting inzake schaal 10 (Totaal) was dat er een niet-significante verbetering zou zijn, dat wil zeggen een verbetering bij het niveau $\alpha = 0.10$.

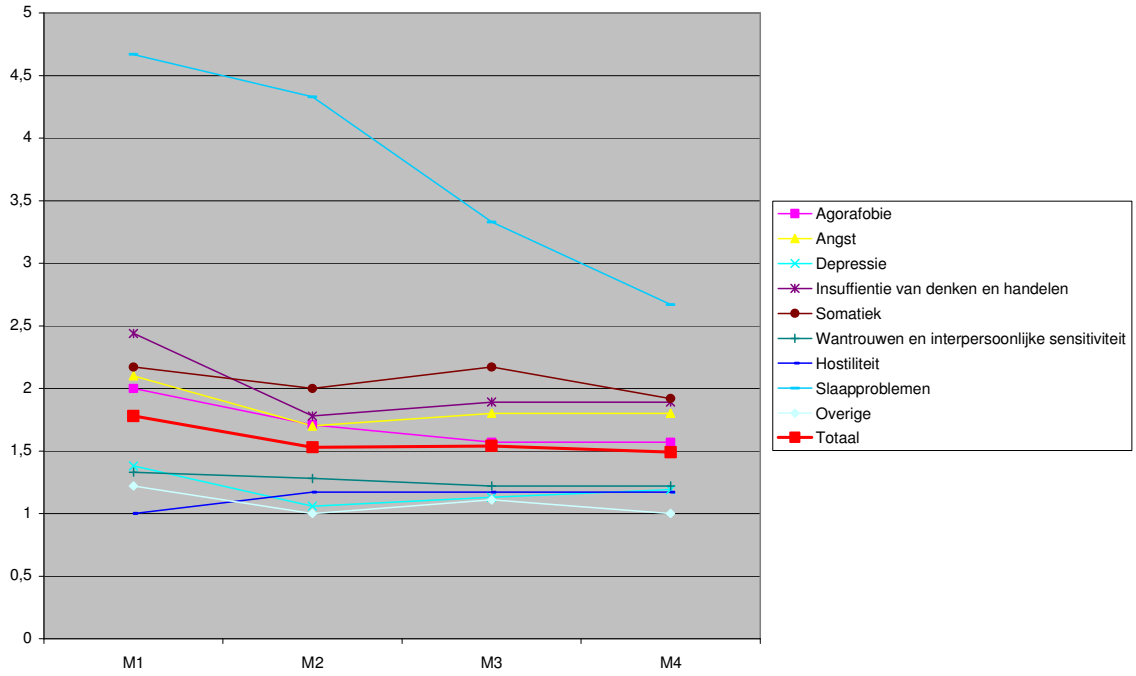
IV Analyses van de resultaten

10 Resultaten

10.1 Grafiek

Grafiek 3 toont de gemiddelde schaalscores bij de vier metingen, nog zonder toetsen. Men ziet dat het slapen verbetert (hoe lager het gemiddelde, hoe beter). Vooruitgang werd verwacht op angst, depressie, insufficiëntie van denken & handelen, en slaapproblemen. Men ziet dat er op die schalen inderdaad vooruitgang is, vooral van M1 naar M2 en min of meer gelijk blijven bij M3 en M4.

Grafiek 3: SCL90-waarden bij M1, M2, M3 en M4



10.2 Vergelijking met normgroepen

Tabel 2: gemiddelden, spreidingen en normgroepgegevens per schaal

SCL90-schalen	Metingen															
	M1				M2				M3				M4			
	M	SD	NG1	NG2	m	SD	NG1	NG2	M	SD	NG1	NG2	m	SD	NG1	NG2
1	2,00	1,15	4	7	1,71	0,76	4	7	1,57	0,79	3	7	1,57	0,79	3	7
2#	2,10	0,74	3	5/6	1,70	0,48	3	5/6	1,80	0,42	3	5/6	1,80	0,42	3	5/6
3#	1,38	0,50	2	4	1,06	0,25	1	2	1,13	0,34	1	2	1,19	0,54	1	2
4#	2,44	1,24	4	7	1,78	0,67	3	5/6	1,89	0,33	3	5/6	1,89	0,78	3	5/6
5	2,17	0,58	4	6	2,00	0,43	4	6	2,17	0,58	4	6	1,92	0,52	4	6
6	1,33	0,49	2	4	1,28	0,58	2	4	1,22	0,55	2	4	1,22	0,43	2	4
7	1,00	0,00	1/2	2	1,17	0,41	3	4	1,17	0,41	3	4	1,17	0,41	3	4
8#	4,67	0,57	6	7	4,33	0,58	6	7	3,33	0,58	5	7	2,67	0,58	4	5/6
9	1,22	0,44			1,00	0,00			1,11	0,33			1,00	0,00		
10#	1,78	0,97	3	6	1,53	0,78	2	6	1,54	0,69	2	6	1,49	0,64	2	6

Toelichting

[1] De itemscores per schaal zijn weergegeven in bijlage 1.

[2] De nummering van de SCL90-schalen is conform § 8. De schalen waarop vooruitgang werd verwacht (2, 3, 4, 8 en 10) zijn met # aangegeven.

[3] Metingen: M1 = meting 1, M2 = meting 2, M3 = meting 3, M4 = meting 4. M1 vond plaats aan het begin van de therapie, tussen elke meting lag 3 maanden.

[4] M: gemiddelde itemscore van de schaal, SD: standaarddeviatie (spreiding).

[5] NG1: normgroep 1 (poliklinische psychiatrische patiënten), NG2: normgroep 2 (gewone bevolking). Normklassen: 1 = zeer laag, 2 = laag, 3 = benedengemiddeld, 4 = gemiddeld, 5 = bovengemiddeld, 6 = hoog, 7 = zeer hoog. Voor schaal 9 (overig) zijn geen normgroepgegevens. Soms worden in de SCL90-handleiding twee normgroepen als één gegeven bij bepaalde scores (in de tabel is dat bij 5/6).

Men ziet (uiteraard) hetzelfde beeld als in grafiek 3. Afgezien van eventuele significanties: er is vooruitgang van M1 naar M2 op Agorafobie (1), Angst (2), Depressie (3), Insufficiëntie Denken & Handelen (4), Overige (9) en het Totaal (10). Van M2 naar M3 en M4 verandert er niet veel in vergelijking met het verschil M1-M2, behalve dat Slapen (8) steeds beter wordt. Heel kleine verschillen zijn er op Somatiek (5, M1-M4), Wantrouwen & Interpersoonlijke Sensitiviteit (6)

Wat kan op grond hiervan geconcludeerd worden?

De aanvangsscores voor de schalen waarop verbetering verwacht wordt (2, 3, 4, 8 en 10) liggen bij M1 binnen de normgroep poliklinische psychiatrische patiënten deels niet hoog (schalen 2, 3 en 4, respectievelijk benedengemiddeld, laag en gemiddeld), maar voor slapen (schaal 8) ligt het hoog. Dit beeld verandert iets als men als normgroep de totale bevolking neemt. Dan vallen de scores op de schalen 2, 3 en 4 respectievelijk in de categorieën bovengemiddeld/hoog, gemiddeld en zeer hoog. Slapen (schaal 8) valt in de categorie zeer hoog.

Is er vooruitgang? Het gemiddelde op Angst bijvoorbeeld daalt (van 2,10 naar 1,70), maar blijft binnen dezelfde normklasse bij zowel normgroep 1 als normgroep 2. Aan deze daling binnen dezelfde normklasse, 'heeft men niets'.

Op de schalen waarop verbetering verwacht werd (2, 3, 4, 8 en 10) kan men de som van de categorieën (normklassen) binnen een normgroep nemen (en waarden geven zoals genoemd in toelichting 5 bij tabel 2). Voor M1 is dat voor normgroep 1 (poliklinische psychiatrische pati-

ënten): 3 + 2 + 4 + 6 + 3 = 18. Bij M2, M3 en M4 zijn die totalen respectievelijk: 15, 14 en 13. Voor vergelijking met normgroep 2 (bevolking in z'n geheel) zijn de sommen bij M1 t/m M4 respectievelijk: 29½, 26, 26 en 24½.
Is dat verbetering? Afgezet tegen normgroep 1 is er (over de schalen waarop vooruitgang verwacht werd, zoals aangegeven in tabel 1) een vooruitgang van 18 naar 15, 14 en 13. Van 18 naar 15 lijkt niet veel, maar is procentueel een verbetering van 16,7%. Geen verwaarloosbaar percentage. Afgezet tegen normgroep 2 is de vooruitgang van M1 naar M4 een verbetering van 5 (29,5 naar 24,5) van de som van de categorieënscores, een verbetering van 16,9%.
Slapen, de voornaamste resultaatvariabele, verbetert van de categorieën 6 (normgroep 1) en 7 (normgroep 2) bij M1 naar respectievelijk 4 en 5/6 bij M4. Zo'n twee klassen vooruitgang dus (2 klassen bij vergelijking met normgroep 1, 1½ bij vergelijking met normgroep 2).
Bij de schalen waarop geen vooruitgang werd verwacht, treedt die niet of vrijwel niet op. Er is alleen bij agorafobie een verbetering van normklasse in vergelijking met normgroep 2 bij M3 en M4.
De vraag blijft of dit alles werkelijk vooruitgang representeert. Die vraag is niet goed te beantwoorden. Het beste antwoord is: er is vooruitgang te zien, mogelijk is dat werkelijke vooruitgang; in elk geval op schaal 8 (Slapen) mogen we concluderen tot werkelijke vooruitgang (verschil M1 naar M4 is een verbetering van 2 normklassen in vergelijking met normgroep 1 en 1½ normklasse in vergelijking met normgroep 2); er is geen achteruitgang.
Kan men met deze gegevens de behandelstrategie bijstellen? Hoogstens met behoorlijk wat slagen om de arm. Er kan niet tot duidelijk werkelijke vooruitgang besloten worden en de gegevens zijn erg globaal.

10.3 Standaardmeetfout

Met de methode van de standaardmeetfout vergelijken we een gebleken verschil op de schalen waarop vooruitgang verwacht werd, bij twee metingen met de waarde van 1,645 × de standaardmeetfout ($1{,}645S_e$). Zie tabel 3.
Als men één waarneming heeft (één schaalscore van één persoon, zoals hier) is er een vrij grote toevalsvariatie ($1{,}645S_e$). Bij SCL90-angstschaal dus 9,99. We besluiten pas tot een werkelijk verschil als een opgetreden verschil daaraan gelijk of groter is. Dat is een erg grote range. Dit verklaart waarom bij individuele scoringen zo moeilijk significante vooruitgang is aan te tonen, terwijl er in werkelijkheid heel goed vooruitgang kan zijn (grote type II fout).

Tabel 3: Verschillen tussen de metingen vergeleken met 2 × de standaardmeetfout, éénzijdige vergelijking

SCL90				Metingen									
				M1	M2			M3			M4		
Schaal	**N**	**S_e**	**$1{,}645S_e$**	**Score**	**Score**	**V**	**Sign.?**	**Score**	**V**	**Sign.?**	**Score**	**V**	**Sign.?**
2	10	6,07	9,99	21	17	4	Nee	18	3	Nee	18	3	Nee
3	16	8,35	13,74	22	17	5	Nee	18	4	Nee	19	3	Nee
4	9	5,51	9,06	22	16	6	Nee	17	5	Nee	17	5	Nee
8	3	2,16	3,55	14	13	1	Nee	10	4	***Ja***	8	6	***Ja***
10	90	25,31	41,63	160	138	22	Nee	139	21	Nee	134	26	Nee

Toelichting
- Van de SCL90 in de linkse kolom de schalen (schaalnummers, refererend aan § 8, waarop vooruitgang verwacht werd), gevolgd door het aantal items van de schaal (N), de standaardmeetfout S_e en $1{,}645S_e$ (een gevonden schaalverschil moet groter dan $1{,}645S_e$ zijn).

- M1, M2, M3 en M4: meting 1, meting 2, meting 3 en meting 4 (elke keer na 3 maanden)
- Onder 'Score' de schaalscore (totale schaalwaarde per meting, som van de itemscores). De schaalscores zijn de som van itemscores. Schalen hebben een verschillend aantal items, schaalscores zijn dus niet onderling vergelijkbaar.
- V is het verschil met M1 (M1-M2, M1-M3 en M1-M4).
- Er zijn voorafhypothesen over vooruitgang. Er is daarom éénzijdige toetsing. De vraag is of een verschil groter is dan $1{,}645S_e$ waarmee het significant is bij niveau $\alpha = 0.05$. Alleen de verschillen op schaal 8, Slapen, tussen M1 en M3, en M1 en M4 zijn dan significant.

Conclusie: alleen werkelijke vooruitgang op de schaal Slapen tussen M1 en M3, en tussen M1 en M4. Deze conclusie komt overeen met die uit § 10.2 na de vergelijking met de normklassen. Dit wil overigens niet zeggen dat er geen werkelijke vooruitgang wás op andere schalen 2, 3, 4 en 10. De type II fout speelt hier een rol. Het is heel goed mogelijk dat die er was. Alleen de kans dat ten onrechte tot vooruitgang geconcludeerd wordt die er in werkelijkheid niet is, is maximaal 5% (als het gaat om één hypothese, zie de Box 2 in hoofdstuk 1 en § 11.4). De kans dat een eventuele werkelijke vooruitgang niet gesignaleerd wordt, is aanmerkelijk groter dan 5%.
Als dit wordt berekend, blijkt het volgende: de statistische Power, het onderscheidingsvermogen (zie hoofdstuk 1) is heel gering, de type II fout groot ($\beta = 0.84$). Als dit verschil van M1 = 21 en M2 = 17 een 'werkelijk' verschil zou zijn (wat niet bekend is, want men voert juist de meting uit om dat te taxeren), is er slechts 16% kans dat het als een 'werkelijk' bestaand verschil zou worden gezien (21 en 17 als gevonden scores zijn benaderingen van onbekende 'werkelijke' scores). Zie Box 2.

Box 2: β en 1 – β

De kans op type II fout (β) en het onderscheidingsvermogen van de toets (1 – β) is niet makkelijk te berekenen (Van den Brink & Koele, 2005, p. 13-39; Ellis, 2007; Hays, 1970, p. 269-280; Lindgren, 1993, p. 328-335).
Ten eerste kan dit in principe alleen berekend worden als het gaat om het vergelijken van twee gemiddelden (van steekproeven of populaties). Maar we hebben hier twee scores van één persoon. Dit geeft weer aan dat alle procedures afgeleid zijn van methoden voor groepsvergelijkingen. Overigens kan aan de hand van de normaalverdeling in het geval van twee individuele scores waarvan de standaardmeetfout bekend is, wel worden aangegeven wat β kan zijn (Meerling, 1989).
Ten tweede betreft de type II fout de kans dat een werkelijk bestaand effect of verschil niet wordt onderkend. Dat kan alleen berekend worden, uitgaande van een specifiek, gehypothetiseerd werkelijk effect of verschil. Men vindt dus niet één β en (1 – β) maar afzonderlijke β's al naar gelang het werkelijk gedacht verschil dat men poneert of veronderstelt.
We gaan uit van de gegevens in tabel 3 over de SCL90-angstschaal (schaal 2). De score bij M1 was 21, die bij M2 was 17. Het verschil moet bij $\alpha = 0.05$ en éénzijdig toetsen groter zijn dan 9,99 (= $1{,}645S_e$) om significant te zijn. Het blijkt kleiner. De H_0 van geen verschil wordt niet verworpen, de H_1 van wel verschil, vooruitgang naar minder angst, wordt niet aangenomen.

Om elk van de schaalscores 21 en 17 ligt een verdeling van mogelijke scores. De verdeling wordt weergegeven met de standaardmeetfout (6,07 voor schaal 2). Want de cliënt scoort 21 bij M1, maar zijn 'werkelijke' score hoeft dat niet te zijn. Zijn 'werkelijke' score kennen we niet, de score die we vinden is een benadering. Rondom de werkelijke score ligt een verdeling van toevalsfluctuaties die wordt weergegeven met de standaardmeetfout. Wat we willen weten is of de scores 21 en 17 indicaties zijn van 'werkelijke' scores die wél verschillen.
Als de cliënt bij M1 een score heeft van 21 dan zal een andere score pas als significant lager worden beschouwd als deze 11 of kleiner is ($21 - 1{,}645S_e = 21 – 10 = 11$, afgerond). De gevonden score van 17 bij M2 was daarvoor te hoog.
Laten we er nu hpothetisch vanuit gaan dat de 'werkelijke' scores bij M1 en M2 inderdaad 21 en 17 zijn. De vraag is nu: hoeveel kans is er met de gehanteerde beslissingsregel van een verschil van minimaal $1{,}645S_e$ (= 9,99 afrondend op 10) tot een conclusie van een werkelijk verschil te komen? Dat is het onderscheidingsvermogen 1 – β van de toets. En hoeveel kans (β) is er dat we het 'werkelijke' verschil tussen 21 en 17 niet vinden? We volgen de procedure van Meerling (1989).
Beneden 11 beschouwt de toets een gevonden waarde van M2 als significant. We vonden 17, het verschil is 6. De z-score behorend bij 11 van een verdeling met m = 17 en s = 6,07 is 0,99 (= 6/6,07). Daarbij hoort in de tabel van de normaalverdeling een linkse overschrijdingskans van 0,1611 en een rechterdeel van de verdeling van 0,8389. Afgerond 16% en 84%. Derhalve $\beta = 0{,}84$ en $1 – \beta = 0{,}16$. Anders gezegd: áls 21 en 17 de 'werkelijke' scores zouden zijn geweest bij M1 en M2, dan is de kans dat we dit verschil niet onderkennen 84%. β is dus heel groot. De gebruikelijke weg om β te verkleinen, de groepsgrootte te vergroten, kunnen we niet volgen omdat het hier om één persoon gaat.

Concluderend: een verschil van angstscores van 21 en 17 bij M1 en M2 is niet significant en we besluiten dat er werkelijk geen verschil is, bij $\alpha = 0.05$. Zouden evenwel 'in werkelijkheid' de scores inderdaad 21 en 17 zijn (we weten dat niet, want daarvoor doet men onderzoek), dan is de kans dat we dat registreren (kans op significantie) slechts 16% ($1 – \beta = 0{,}16$).

10.4 RCI

Hierna, in de tabellen 4.1 en 4.2, de resultaten van de RCI.. Voor de berekeningen werd uitgegaan van de gegevens over de SCL90 van Arrindell en Ettema (2003, p. 35). RCI's werden berekend voor de schalen waar vooruitgang verwacht werd. Voor de berekening van een RCI wordt gewerkt met de schaalscores (som van de itemwaarden van een schaal), vandaar dat die hier worden gegeven. Bedacht moet worden dat de schalen onderling variëren in aantal items, waardoor de schaalscores onderling niet te vergelijken zijn.
In tabel 4.1 eerst de schalen waarop vooruitgang verwacht werd, het aantal items van een schaal en de somscore van de items (de schaalwaarde). In tabel 4.2 de schalen waarop vooruitgang verwacht werd, de standaardmeetfout van verschilscores, schaalscoreverschillen bij de metingen en nde RCI's.
De procedure van de standaardmeetfout ($1{,}645S_e$ vergelijken met een gebleken verschil) komt erg, hoewel niet volledig, overeen als de RCI. Met de RCI heeft men een coëfficiënt, waardoor RCI's (en daarmee verschillen) onderling vergeleken kunnen worden.

Tabel 4.1: Totaalscores (schaalscores) op de schalen waarop vooruitgang verwacht werd

SCL90		Metingen			
Schaal	**N (items)**	**M1**	**M2**	**M3**	**M4**
2	10	21	17	18	18
3	16	22	17	18	19
4	9	22	16	17	17
8	3	14	13	10	8
10	90	160	138	139	134

Toelichting
- M1, M2, M3 en M4: meting 1, meting 2, meting 3 en meting 4 (elke keer na 3 maanden)
- Gegeven worden de schaalscores, de totalen per schaal.

Tabel 4.2: Verschillen tussen meting 1 (M1) en de metingen 2, 3 en 4 (M2, M3 en M4)

SCL90		Verschillen tussen metingen					
		M1 – M2		M1 – M3		M1 – M4	
Schaal	**S_{diff}**	**Verschil**	**RCI**	**Verschil**	**RCI**	**Verschil**	**RCI**
2	8,58	4	0,47	3	0,35	3	0,35
3	11,81	5	0,42	4	0,34	3	0,25
4	7,79	6	0,77	5	0,64	5	0,64
8	3,05	1	0,33	4	1,31	6	1,97*
10	35,79	22	0,61	21	0,59	26	0,73

Toelichting
- M1, M2, M3 en M4: meting 1, meting 2, meting 3 en meting 4 (elke keer na 3 maanden)
- Gegeven worden de schaalscores, de totalen per schaal.
- S_{diff} (de standaardmeetfout van schaalverschilscores) werd berekend op basis van de gegevens van Arrindell en Ettema (2003, p. 35).
- RCI = V/ S_{diff}
- * = Significant op niveau $\alpha = 0.05$.

Conclusie op basis van de RCI's

Op basis van de RCI's moet besloten worden dat alleen de verbetering van Slapen (SCL90-schaal 8) van M1 naar M4 (RCI = 1,97 > 1,645) significant is en betekenis heeft. Dit verschil is significant ($\alpha = 0.05$, bij éénzijdig toetsen, maar is bij tweezijdig toetsen ook significant).
Men kan zeggen: de vooruitgang bij de schalen waarop vooruitgang verwacht werd, is er wel, maar is klein. Zo klein dat we er niet vanuit kunnen gaan dat het wérkelijke vooruitgang is, gezien de RCI's. Men kan daartegenin brengen dat de vooruitgang in termen van percentages behoorlijk oogt. Een vooruitgang van 21 (tabel 4.1, schaalscore bij M1 voor schaal 2, Angst) naar 17 (schaalscore M2 voor schaal 2) is 4/21 = 19%. Dat lijkt een substantieel percentage (als de vakbeweging zo'n salarisverhoging zou eisen of een regering zoveel wil bezuinigen, is er een maatschappelijke crisis).
Op basis van de RCI's besluiten we echter dat dit binnen de toevalsrange valt en dat we niet tot een werkelijke daling mogen besluiten. Therapeutisch is een schaalscoreverbetering van 21 naar 17 op de angstschaal (en een verbetering van het itemgemiddelde van 2,1 naar 1,7) van behoorlijke betekenis mits we er evenwel überhaupt betekenis aan mogen verlenen.
Men ziet dat het verschil op Angst van M1 naar M4 zowel op basis van de standaardmeetfout (tabel 3) als op basis van de RCI (tabel 4.2) significant is. Maar het verschil van M1 naar M3 is op basis van de standaardmeetfout wel significant (tabel 3), echter op basis van de RCI niet (tabel 4.2). Hoe kan dit, terwijl beide wel op dezelfde gegevens gebaseerd zijn?
Het antwoord werd al gegeven aan het eind van hoofdstuk 5: de RCI is gebaseerd op de standaardmeetfout van verschillen (verschilscores van M1 met M2, M3 of M4) terwijl de standaardmeetfoutprocedure gebaseerd is op de standaardmeetfouten van M1, M2, M3 en M4 zelf. De standaardmeetfout van een verschil is groter dan die van de oorspronkelijke scores waarvan het verschil is afgeleid.
Werken met de RCI is niet 'beter' dan de standaardmeetfout, het is een iets ander principe: de standaardmeetfout van een verschil is groter dan die van de afzonderlijke scores. De RCI is het meest robust, maar zal ook vaak niet significant zijn als er tóch (in werkelijkheid) wel een effect is.

Een verschil van het gemiddelde bij M1 voor schaal 2 (Angst) van 2,10 met het gemiddelde bij M2 van 1,70 (= 0,4 ofwel 19%; zie tabel 1.1 en tabel 2) is dus niet significant gebleken. Zoals al aangegeven in hoofdstuk 3 zijn deze manieren van individuele verschillen vergelijken afgeleid van vergelijkingsprocedures voor groepen. Zou men groepsgemiddelden (de schaalscores van een aantal personen bij een vóór- en námeting) vergelijken dan kan een kleiner verschil wel degelijk significant zijn (uiteraard afhankelijk van het aantal personen en de SD). Arrindell en Ettema (2003, p. 41-42) noemen een effectonderzoek over een semi-gestructureerde groepsprogramma van zeven tweewekelijkse sessies van twee uur met psycho-educatie, coping, seksualiteitsbeleving. Er was een verbetering op Angst van 16,5 - 13,4 = 3,1 ($s_{pre} = 6,7$; $s_{post} = 3.7$). Dat was een significante waarde.
Als de waarde van 0,4 wel significant geweest was, zou het gerechtvaardigd zijn de maat voor effectgrootte δ van Cohen (1988), ook wel ES: effect size genoemd (Cohen, 1988; Veerman, 2006) te gebruiken. Dit is het verschil van M1 min M2 gedeeld door de gepoolde spreiding (voor Angst zijn de spreidingen bij M1 en M2, zie tabel 2, respectievelijk 0,74 en 0,48).
De gepoolde spreiding is: $\sqrt{\{[(N_1-1) \times SD_1^2 + (N_2-1) \times SD_2^2]/(N_1+N_2-2)\}}$. Voor de SCL90-angstschaal (schaal 2) in tabel 2 levert dat een waarde op van 0,6237. ES (δ) zou dan worden: 0,4/0,6237 = 0,64. Dat is voor een ES een middelgrote, en dus behoorlijke waarde.

10.5 Resultaten van vergelijkingen en toetsen met normklassen, standaardmeetfouten en RCI vergeleken

In tabel 5 een vergelijking van de drie manieren van verschillen beoordelen.

Zoals we eerder zagen in dit hoofdstuk zijn er verbeteringen inzake normklassen al is niet duidelijk wat voor gewicht en betekenis daaraan gegeven kan worden. Gebruiken we de standaardmeetfout en de RCI als criterium dan zijn er respectievelijk slechts 2 en 1 significanties. Dit wordt samengevat in tabel 5, hierna.

Tabel 5: Conclusies op basis van normklassen, standaardmeetfout en RCI vergeleken

SCL90-schaal	Verschillen ten opzichte van M1											
	M1 – M2				M1 – M3				M1 – M4			
	NG		S_e	RCI	NG		S_e	RCI	NG		S_e	RCI
	1	2			1	2			1	2		
2												
3	1	2			1	2			1	2		
4	1	1½			1	1½			1	1½		
8					1		*		2	1½	*	*
10	1											

Toelichting

[1] De nummering van de SCL90-schalen is conform hoofdstuk 8. Het gaat hier alleen om de schalen waarop vooruitgang werd verwacht.

[2] Metingen: M1 = meting 1, M2 = meting 2, M3 = meting 3, M4 = meting 4. M1 vond plaats aan het begin van de therapie, tussen elke meting lag 3 maanden.

[3] Gegeven de verbetering van het aantal normklassen. 1 betekent derhalve vooruitgang van 1 normklasse. NG1: normgroep 1 (poliklinische psychiatrische patiënten), NG2: normgroep 2 (gewone bevolking). Normklassen: 1 = zeer laag, 2 = laag, 3 = benedengemiddeld, 4 = gemiddeld, 5 = bovengemiddeld, 6 = hoog, 7 = zeer hoog. Voor schaal 9 (overig) zijn geen normgroepgegevens. Soms worden in de SCL90-handleiding twee normgroepen als één gegeven bij bepaalde scores (in de tabel is dat bij 5/6). Verbetering ten opzichte daarvan kan dus resulteren in de verbetering van ½ normklasse.

[4] S_e: Standaardmeetfout

* = significant op niveau $\alpha = 0.05$.

[5] Het verschil M1-M3, getoetst met de standaardmeetfout is wel significant, maar met de RCI niet. Dit werd al uitgelegd aan het eind van hoofdstuk 5 en in § 10.4: de RCI berust op de standaardmeetfout van het verschil die groter is dan de standaarmeetfout van de oorspronkelijke score(s).

10.6 Toetsen op itemniveau

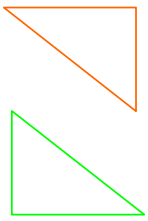
10.6.1 Op basis van de veranderingsmatrix, toets over de matrixgegevens met de McNemar-toets

In hoofdstuk 6, tabel 1.2 werd de veranderingsmatrix van de SCL90-angstschaal (schaal 2) gegeven voor meting 1 (M1). Zie tabel 6.

Tabel 6 (= tabel 1.2)

Waarden itemscores bij M2

Waarden itemscores bij M1	**1**	**2**	**3**	**4**	**5**
1	2				
2	1	4			
3		3			
4					
5					



Driehoek met items waarop achteruitgang naar voren komt.

Driehoek met items waarop vooruitgang naar voren komt.

Zes itemscores zijn gelijk gebleven, 4 zijn er gedaald (verbeterd). We toetsen dit met de McNemar-toets met als nulhypothese dat het aantal items met vooruitgang (nu 4) gelijk is aan het aantal items met achteruitgang (nu 0).[6] Dit kan getoetst worden met een χ^2-grootheid met df = 1. $\chi^2 = 4$, wat significant is op niveau $\alpha = 0.05$ ($\chi^2 \geq 3{,}841$ is significant op 0.05-niveau, éénzijdig toetsen). We concluderen derhalve dat de angstscore verlaagd (verbeterd) is van M1 naar M2. Deze McNemar-toets is non-parametrische toets (is niet gebaseerd op de statistische normaalverdeling). De toets maakt geen verschil tussen de mate van vooruitgang, alleen tussen al of niet vooruitgang. Indien er op een beperkt aantal items veel vooruitgang is, zal de toets dat lager waarderen dan beperkte vooruitgang op veel items.

Op dezelfde manier zijn ook de andere schalen waarop verbetering verwacht werd, te toetsen. Dat is gedaan in tabel 7.

Tabel 7: Verschillen tussen meting 1 (M1) en de metingen 2, 3 en 4 (M2, M3 en M4)

SCL90 schaal	**Verschillen ten opzichte van M1**								
	M1 – M2			**M1 – M3**			**M1 – M4**		
	Vooruit	**Achteruit**	χ^2	**Vooruit**	**Achteruit**	χ^2	**Vooruit**	**Achteruit**	χ^2
2	4	0	4,0*	4	1	1,8	4	1	1,8
3	5	0	5,0*	5	1	2,7	4	1	1,8
4	5	0	5.0*	5	2	1,3	4	0	4,0*
8	1	0	$1{,}0^{\#}$	3	0	$3{,}0^{\#}$	3	0	$3{,}0^{\#}$*
10	26	8	9,5*	29	9	10,5*	26	7	10,9*

[6] De McNemar-toets: Kwadraat van de aftrekking van het aantal items met vooruitgang min dat met achteruitgang, gedeeld door de som van beide. In formule: $\chi^2 = (n_{\text{aantal itemsvooruitgang}} - n_{\text{aantal items achteruitgang}})^2/(n_{\text{aantal items vooruitgang}} + n_{\text{aantal items achteruitgang}})$. In het voorbeeld: $\chi^2 = (4-0)^2/4 = 4$. Zie Bisshop, Fienberg, Holland, Light en Mosteller (1995) en Sheskin (2004, p. 633-664 en 725).

Toelichting

* Significante vooruitgang op niveau $\alpha = 0.05$.

\# Significante vooruitgang op niveau $\alpha = 0.05$ blijkt niet op schaal 8. Schaal 8 heeft 3 items. Met 3 items is op de McNemar-toets geen significante vooruitgang mogelijk. Daarom is een toets op basis van pure kansberekening uitgevoerd.

\#* Met de toets op basis van pure kansberekening is er significantie. Zie uitleg hierna en in tabel 8.

SCL90-schaal 8, Slapen, de meest wezenlijke schaal en dimensie voor de therapie heeft slechts 3 items. Met 3 items is met de McNemar-toets nooit een significante waarde te bereiken, zelfs niet als op alle drie de items de score van 5 naar 1 gaat. Want de toets levert bij vooruitgang op 3 items: $\chi^2 = (3\text{-}0)^2/(3+0) = 3$. De voor significantie benodigde waarde is 3,84. Met slechts 3 items is de toets dus niet te gebruiken, want de kans op significantie is 0 (de power, $1 - \beta = 0$). Een andere manier van toetsen in plaats van de McNemar-toets in dit geval is werken met exacte kansen.

Als er 3 items zijn, die elk bij bijvoorbeeld de laatste meting M4 vooruitgang, gelijk blijven of achteruitgang kunnen laten zien ten opzichte van M1, dan zijn er 27 mogelijkheden (één item toont vooruitgang, één gelijkblijven, één achteruitgang; twee vooruitgang, één achteruitgang etc.). Alle 3 de items kunnen vooruitgang laten zien. Dat is één van de 27 mogelijkheden. Die kans zou dus zijn: 1/27 = 0,037 = 3,7%. Eén item kan vooruitgang laten zien, de andere twee gelijk blijven. Dat zijn ook 3 van de 27 mogelijkheden. Die kans zou dus zijn: 3/27 = 1/9 = 0,11 = 11%.

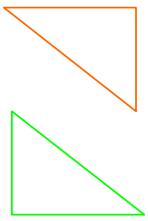
Deze manier van denken kan echter niet gevolgd worden, want het probleem hierbij is dat de kansen van vóórkomen voor vooruitgang of achteruitgang groter zijn dan voor gelijk blijven. Zie tabel 8.

Tabel 8: Verschil M1-M4 op SCL90-schaal Slapen (schaal 8, 3 items)

Waarden itemscores bij M4

Waarden itemscores bij M1

	1	**2**	**3**	**4**	**5**
1					
2					
3					
4		1			
5			2		



Driehoek met items waarop achteruitgang naar voren komt.

Driehoek met items waarop vooruitgang naar voren komt.

Puur op basis van kans heeft elk van de 25 cellen van het data-deel van de matrix evenveel kans. De diagnonaal is het aantal items dat gelijk blijft. De kans daarop bij een vervolgmeting, bijvoorbeeld M4, ten opzichte van de eerste meting (M1), is 5/25 = 0,20 of 20%. De kans op vooruitgang wordt gegeven door de 10 cellen van de (groene) linker driehoek, linksonder in de matrix, en is 10/25 = 0,4 of 40%. Dezelfde kans geldt voor achteruitgang de (rode) driehoek rechtsboven in de matrix. De kans op vooruitgang op alle drie de items, ongeacht de grootte van de vooruitgang, is $(10/25)^3 = 0{,}4^3 = 0{,}064$ of 6,4%. De kans op vooruitgang van 2 punten op alle drie de items, wordt gegeven door een nog kleinere driehoek linksonder is $(6/25)^3 = 0{,}014$ of 1,4%. Het laatste is het geval bij het verschil M1-M4 dat daarmee significant is op 0.05-niveau. Het verschil M1-M3 levert per item 1 punt verbetering op. De kans daarop is de genoem-

de 6,4%. Dat is niet significant op 0.05-niveau (wel op 0.10-niveau). Het verschil M1-M2 is 1 punt verbetering op één van de drie items en gelijk blijven van de twee andere items. De kans daarop is 0,267 of 26,7% wat uiteraard niet significant is.

10.6.2 Itemscores toetsen met gepaarde t-toets

Dezelfde redenering volgend als hiervoor, dat wil zeggen uitgaande van de matrix in tabel 5, dat is in feite uitgaand van tabel 1.1 en 1.2, kunnen de verschillen ook getoetst worden in een gepaarde t-toets.

Hierna in tabel 9.1 (en tabel 9.2) de gebleken waarden bij de metingen en de toetsen op verschillen. Eerst (in het linker deel van de tabel) de gemiddelden per meting, daarna de verschillen van M1 naar M2, M3 en M4 en de verschillen tussen M2, M3 en M4 onderling in het rechter deel. De verschillen zijn getoetst met de gepaarde t-toets. De gegeven waarden in de tabel zijn de p-waarden. Waarden, kleiner dan 0,05 zijn significant (en zijn gearceerd en cursief), waarden tussen 0,05 en 0,10 zijn een niet-significante trend (en lichter gearceerd aangegeven), waarden groter dan 0,10 staan voor gelijk blijven (niet-significant verschil). Bij een verwachte vooruitgang (schalen 2, 3, 4, 8 en 10) wordt éénzijdig getoetst, indien geen verwachtingen geformuleerd zijn (de overige schalen) tweezijdig.
Een (gepaarde) t-toets vraagt om normale verdelingen. Er zijn ook mogelijkheden voor gebruik van een verdelingsvrije toets of op basis van pure kansberekening zoals genoemd in hoofdstuk 6. In bijlage 2 worden dezelfde vergelijkingen met een non-parametrische toets uitgevoerd

Tabel 9.1: Resultaten (verschillen één- en tweezijdig getoetst), gepaarde t-toetsen over itemgemiddelden per schaal

	Gemiddelden				Significantie van verschillen (één- en tweezijdig)					
SCL90-schaal	**M1**	**M2**	**M3**	**M4**	**M1&M2**	**M1&M3**	**M1&M4**	**M2&M3**	**M2&M4**	**M3&M4**
1 Agorafobie	2,00	1,71	1,57	1,57	0,172	0,078°	0,078°	0,356	0,356	0,356
2 Angst#	2,10	1,70	1,80	1,80	0,019*	0,097°	0,097°	0,172	0,172	-
3 Depressie#	1,38	1,06	1,13	1,19	0,010*	0,052°	0,094°	0,167	0,082	0,290
4 Insuffiëntie van denken en handelen#	2,44	1,78	1,89	1,89	0,011*	0,090°	0,026*	0,297	0,174	0,500
5 Somatiek	2,17	2,00	2,17	1,92	0,337	1,000	0,275	0,339	0,586	0,191
6 Wantrouwen en interpersoonlijke sensitiviteit	1,33	1,28	1,22	1,22	0,668	0,430	0,331	0,331	0,579	1,000
7 Hostiliteit	1,00	1,17	1,17	1,17	0,363	0,363	0,363	-	-	-
8 Slaapproblemen#	4,67	4,33	3,33	2,67	0,212	0,029*	0,042*	0,042*	0,019*	0,092°
9 Overige	1,22	1,00	1,11	1,00	0,169	0,594	0,169	0,347	-	0,347
10 Totaal#	1,78	1,53	1,54	1,49	0,000*	0,002*	0,000*	0,798	0,374	0,278

Toelichting en bespreking van de toetsen
- M1, M2, M3 en M4: meting 1, 2, 3 en 4. Tussen elke meting lag circa 3 maanden.
- Cursief en met # in de meest linkse kolom de schalen waarop vooruitgang werd verwacht.
- In de 4 kolommen onder gemiddelden de gemiddelde itemwaarde per schaal bij de 4 metingen. Hier zijn dus gemiddelden per schaal gegeven, als in de tabel 2, waar in de tabellen 3, 4.1 en 4.2 er per schaal totalen vermeld.
- Als eerder aangegeven zijn gepaarde t-toetsen gebruikt.

- Gegeven zijn p-waarden. Significante verschillen op niveau $\alpha = 0.05$ zijn donker gearceerd en met * aangegeven, significante verschillen op niveau $\alpha = 0.10$ zijn licht gearceerd en met ° aangegeven (significante p-waarden tussen 0.05 en 0.10 worden als trend beschouwd).
- De p-waarde voor Slapen (schaal 8) bij M4 is hoger dan die bij M3 terwijl de gemiddelde itemscore is afgenomen (verbeterd van 3,33 naar 2,67). De reden is dat het verschil M1-M4 getoetst werd met een nonparametrische toets (Wilcoxon). De spreiding van de verschillen van M1 naar M4 is namelijk 0 (scores bij M1: 5, 4, 5; bij M4: 3, 2, 3), waardoor de toetsingsgrootheid t niet te berekenen is. Hetzelfde geldt voor het verschil van M2 met M3.
- Verschillen tussen M1 enerzijds en M2, M3 en M4 anderzijds zijn éénzijdig getoetst bij de schalen waarop vooruitgang verwacht werd: er werden verschillen verwacht, verbetering van M2 ten opzichte van M1, daarna voortgaande verbetering of gelijkblijven bij M3 en M4. Verschillen tussen M2, M3 en M4 onderling van deze schalen zijn ook éénzijdig getoetst bij de schalen waarop vooruitgang verwacht werd: áls er verschil zou zijn, werd dat geacht een verbetering te zijn of gelijk blijven.
  Bij schalen waarop geen vooruitgang verwacht werd, is tweezijdig getoetst.
  Er valt overigens iets voor te zeggen om de schalen waarop géén vooruitgang verwacht wordt, naast tweezijdig, ter vergelijking, ook éénzijdig te toetsen en te zien of dit veel verschil maakt. Het is correct om schalen waarop geen vooruitgang verwacht wordt, tweezijdig te toetsen. Maar aangezien bij éénzijdig toetsen sneller significantie bereikt wordt (de p-waarde bij eenzijdig toetsen is de helft van die bij tweezijdig toetsen), 'bevoordeelt' men op die manier de schalen waarop vooruitgang verwacht wordt. Als zou blijken dat de vooruitgang er alleen is bij de schalen waarop vooruitgang verwacht wordt, is dit een sterke bevestiging van de behandeling als oorzakelijke factor. Als evenwel blijkt dat bij éénzijdig toetsen meer schalen (ook enkele van die waarop geen vooruitgang verwacht werd) een significante vooruitgang laten zien, kan dat een aanwijzing zijn van 'kanskapitalisatie' (zie de box in hoofdstuk 1 en § 11.4) Alle verschillen éénzijdig toetsen ter vergelijking is gedaan in tabel 9.2.
  Mogelijk significante uitslagen op basis van kans komen aan de orde in § 11.4 bij de bespreking van de Bonferroni-correctie voor kanskapitalisatie.
- Over enkele verschilmetingen waren geen toetsen mogelijk, op grond van te weinig of te ongelijke spreiding in de scores. Deze verschillen zijn aangegeven met een liggend streepje (-).
- Toetsing kan ook geschieden op basis van een non-parametrische toets als aan vereisten van de normaalverdeling van de scores en gelijke spreiding niet voldaan wordt.
- Hoe staat het hier met β en $1 - \beta$? Immers in § 10.3 en tabel 3 werd β en $1 - \beta$ bepaald in het geval dat de 'werkelijke' scores op de SCL90-angstschaal 21 en 17 waren, en β kwam toen uit op de hoge waarde van 84% (0,84) die de type II foutkans aangeeft. Dat komt overeen met itemgemiddelden van 2,10 en 1,70 (de angstschaal heeft 10 items), zoals ook te zien is in tabel 8.1.
  De spreidingen (SD) bij $m = 2{,}10$ en $m = 1{,}70$ zijn respectievelijk 0,74 en 0,48 (zie het slot van § 10.4 en tabel 2). De gepoolde spreiding is 0,62. De standaardfout wordt geschat op: $SD/\sqrt{n}$. Dit is dus respectievelijk $0{,}74/\sqrt{10} = 0{,}23$ en $0{,}48/\sqrt{10} = 0{,}15$. Indien de itemgemiddelden werkelijk 2,10 en 1,70 waren dan is de linkeroverschrijdingskans met een normaalverdeling rond 2,10 en $1{,}645S_e = 1{,}645 \times 0{,}23 = 0{,}38$. Indien derhalve de score bij M2 lager is dan $2{,}10 - 0{,}38 = 1{,}72$ dan zal deze als significant beschouwd worden (bij $\alpha = 0.05$), zoals ook blijkt uit tabel 9.1. Leggen we de verdeling van M1 ($m = 2{,}10$ met $S_e = 0{,}23$) en M 2 ($m = 1{,}70$ met $S_e = 0{,}15$) over elkaar, dan heeft 1,72 in de verdeling van M2 een linkeroverschrijdingskans van $z = (1{,}72 - 1{,}70)/S_e = 0{,}02/0{,}15 = 0{,}13$. Daarbij hoort een linkeroverschrijdingskans van 0,55. Derhalve geldt dan: $1 - \beta = 0{,}55$ (55%) en $\beta = 0{,}45$ (45%). Dat is nog steeds behoorlijk groot, maar aanmerkelijk gunstiger dan de eerder, in § 10.3 berekende $1 - \beta = 0{,}16$ en $\beta = 0{,}84$.

De resultaten

- De behandeling was vooral gericht op verbetering inzake slaapproblemen (schaal 8). Deze waren gerelateerd aan andere aspecten, waaraan in de therapie ook gewerkt werd: Angst, Depressie, Insufficiëntie van Denken & Handelen. Op deze aspecten (schalen) werd dan ook een verbetering verwacht.
- Een significante verandering treedt op bij de schalen waarop dat verwacht werd: 2, 3, 4, 8 en 10. Op schaal 8 is er een significante afname van M1 naar M3 en naar M4 (nog niet naar M2).
- In de grafiek (grafiek 3) is te zien dat de verbetering vooral plaatsvindt van M1 naar M2 behalve bij Slapen. Ook Slapen verbetert van M1 naar M2, maar is nog niet significant. De verbetering zet door bij M3 en M4 en is dan significant.
- Duidelijk is te zien hoe verbetering van Slapen (schaal 8), waarop de therapie primair gericht was, eruit springt. De toets is tamelijk sensitief: de verschillen van de andere schalen waarop vooruitgang werd verwacht (2, 3 en 4) zijn te zien in grafiek 3, maar eigenlijk beperkt. Duidelijk is te zien hoe juist de slaapproblemen, waarop de therapie vooral gericht was, sterk afnamen.
- Op schaal 3 is er een significante afname van M1 naar M2, maar niet meer van M1 naar M3 en M4, al blijft de afname nog wel beneden de p-waarde van .10 (niet-significante afname, wel trend). Tussen M2 en M3 en M4 is geen significant verschil.
  Men kan conservatief redenen en concluderen dat er een significante afname is van M1 naar M2, maar niet meer van M1 naar M3 en M4. Dit betekent dat er een vooruitgang was van M1 naar M2, maar niet meer van M1 naar M3 en M4. Men kan ook redeneren: er was een afname van M1 naar M2, en van M2 naar M3 en naar M4 was er geen verschil. De afname kan zo bezien als gehandhaafd beschouwd worden. De eerste manier van redeneren wordt aangehouden.
- Als alle verschillen éénzijdig getoetst zouden zijn, blijken er dan verschillen met tweezijdig toetsen? Tabel 9.2, hierna, is identiek aan tabel 9.1 maar nu zijn álle toetsen éénzijdig uitgevoerd. Zoals men ziet betekent dit dat Agorafobie (schaal 1) ook significante verschillen geeft. De schaal Overige (schaal 9) geeft tweemaal een trend. Dit zijn geen vreemde en onverklaarbare verschillen. Bij agorafobie speelt ook angst; als er op de angstschaal een significante vooruitgang optreedt, is het niet vreemd dat er ook vooruitgang is bij agorafobie. Als de verwachte vooruitgang optreedt bij alle verwachte schalen, dan is het – achteraf – niet vreemd dat ook de totaalscore een significante vooruitgang laat zien. Inzake de schaal Overige (9 items) blijkt er vooruitgang te zijn op 2 items: ‘alleen voelen’ en ‘zich psychisch niet in orde voelen’. Andere (gelijkblijvende) items zijn stemmen horen, teveel eten, gedachten aan seks, gedachten aan straf, schuldgevoelens.
  Kortom: als alle verschillen éénzijdig worden getoetst, blijkt hetzelfde beeld.

Tabel 9.2: Gepaarde t-toetsen over itemscores (alle verschillen éénzijdig getoetst)

	Gemiddelden				Significantie van verschillen (eenzijdig)					
SCL90-schaal	**M1**	**M2**	**M3**	**M4**	**M1&M2**	**M1&M3**	**M1&M4**	**M2&M3**	**M2&M4**	**M3&M4**
1 Agorafobie	2,00	1,71	1,57	1,57	0,086°	0,039*	0,039*	0,178	0,178	0,178
2 Angst	2,10	1,70	1,80	1,80	0,019*	0,097°	0,097°	0,172	0,172	-
3 Depressie	1,38	1,06	1,13	1,19	0,010*	0,052°	0,094°	0,167	0,082	0,290
4 Insuffiëntie van denken en handelen	2,44	1,78	1,89	1,89	0,011*	0,090°	0,026*	0,297	0,174	0,500
5 Somatiek	2,17	2,00	2,17	1,92	0,170	0,500	0,138	0,170	0,293	0,096
6 Wantrouwen en interpersoonlijke sensitiviteit	1,33	1,28	1,22	1,22	0,334	0,215	0,166	0,166	0,290	0,500
7 Hostiliteit	1,00	1,17	1,17	1,17	0,182	0,182	0,182	-	-	-
8 Slaapproblemen	4,67	4,33	3,33	2,67	0,212	0,029*	0,042*	0,042*	0,019*	0,092°
9 Overige	1,22	1,00	1,11	1,00	0,085°	0,297	0,085°	0,174	-	0,174
10 Totaal	1,78	1,53	1,54	1,49	0,000*	0,000*	0,000*	0,399	0,187	0,139

- Inzake de evaluatie van de behandeling en de keuze van vervolgstrategieën.
  Op basis van de vooruitgangstoetsingen met gepaarde t-toetsen over de itemscores, treedt de vooruitgang op waar die verwacht wordt en is een differentiële evaluatie van de behandelvoortgang mogelijk. Bij M2 blijkt dat er verbetering is op de schalen 2, 3, 4 en 10. Maar nog niet op 8, de voornaamste. De behandelstrategie was in hoofdzaak angst en spanning beter hanteerbaar maken, waarvan een indirecte vooruitgang verwacht werd op slapen.
  Ook was de cliënt de eerste maanden verhoogd gespannen, nadat hem de diagnose ASS was meegedeeld en hij zich daar zorgen over maakte.
  Direct werd slapen bewerkt door de omstandigheden van slapeloosheid uit te vragen en met adviezen te komen. De nog niet eerder gestelde diagnose autismespectrumstoornis en de vragen en het praten over het slapen zorgden ervoor dat de cliënt in de eerste periode juist méér gespannen werd over zijn slapen (zie hoofdstuk 7).
  Bij M2 blijkt vervolgens vooruitgang op de Angst (schaal 2), Depressie (schaal 3) en Insufficiëntie van Denken & Handelen (schaal 4) en het totaal (10), maar nog niet bij Slapen.
  Afgezien van een 'technische' verklaring over het slapen (de schaal telt slechts 3 items, een verschil moet daarom redelijk groot zijn om significant te worden), ligt dit in de lijn van de behandeltheorie en -strategie. De therapie wordt daarom voortgezet als voordien, het accent verschuift nu meer naar het rechtstreeks focussen op het slapen.
  Bij M3, 3 maanden later en 6 maanden na het begin, blijkt het slapen verder verbeterd en nu is er wel een significant verschil met het begin op de SCL90-schaal Slapen. Ook op schaal 10 bleef de vooruitgang significant. Maar de vooruitgang van M1 naar M3 is voor de andere schalen waarop vooruitgang verwacht werd (2, 3 en 4) nu niet meer significant beter dan bij M1 (nog wel een trend, significant bij $\alpha = 0.10$). Overigens is het verschil met M2 op deze schalen klein.
  De conclusie op dat moment is dat de behandeling de kernvariabele (Slapen, schaal 8) aan het beïnvloeden is, dat de variabelen via welke mede de kernvariabele beïnvloed werd nu iets minder aandacht kregen en iets in score zijn gestegen (achteruit gegaan, mogelijk gestabiliseerd op een bepaald niveau). De keus is nu: *a* verder inzetten op verbetering van het slapen door dat rechtstreeks te benaderen, *b* verder inzetten op de schalen Angst, Depressie en Insufficiëntie van Denken & Handelen om daarmee cliënts redzaamheid te vergroten en indirect het slapen te verbeteren, of *c* beide te doen. De voorkeur van de therapeut had *a*, maar omdat er nogal wat stress was in het leven van cliënt (het hem erg vertrouwde familielid overleed, hij moest verhuizen), werd gekoerst op *c*.

Het resultaat bij M4 laat zien dat het slapen verder daalde (nog steeds significant beter dan M1 maar iets minder significant dan bij M3 wat aan de wat grotere spreiding van de verschilscores kan liggen), dat de schalen 2, 3 en 4 op ongeveer hetzelfde niveau bleven (het verschil M1-M4 is voor schaal 4 nu weer significant). Schaal 10 bleef steeds significant lager dan bij M1.

In plaats van t-toetsen kan ook gewerkt worden met non-parametrische toetsen. Die geen normale verdeling van de scores veronderstellen Dit is gebeurd in bijlage 2. Zie hiervoor ook § 11.1.

10.6.3 Werken en toetsen met verdelingen van somwaarden

Een heel andere en nieuwe manier van werken met itemscores, is werken met somwaarden. De itemscores op schaal 8, Slapen, waren bij M1: 5, 4 en 5, weer te geven als {5,4,5}. Men kan nu redeneren dat bij de keus voor een score de invuller een zekere onzekerheid heeft. Wat geeft het meest adequaat zijn beleving en ervaring weer? Als hij 5 scoort, had hij wellicht ook 4 kunnen scoren, of 6 als die mogelijkheid er geweest zou zijn. Dit is weer te geven als (4,5,6). Als hij 4 scoort, had het wellicht ook 3 of 5 geweest kunnen zijn, weer te geven als (3,4,5). Aldus wordt elke score omgezet in een 'waarschijnlijk scorebereik' van drie scores.
Vervolgens worden alle combinaties van de scores van het waarschijnlijke scorebereik met elkaar gecombineerd. Men krijgt dan $3^3 = 27$ somscores. Die vormen een verdeling. In het geval van {5,4,5} resulteert 1 somwaardenscore van 11, 3 van 12, 6 van 13, 7 van 14, 6 van 15, 3 van 16 en 1 van 17. Deze somwaardenscores vormen dus een (symmetrische) verdeling.
Bij M2 scoort de cliënt {5,4,4}. Dus op 1 item 1 punt lager. Ook van {5,4,4} is een verdeling van somwaardenscores op te stellen.
Vervolgens worden beide verdelingen met elkaar vergeleken. Is er 50% of meer overlap, dan worden ze niet als verschillend beschouwd. Is er minder dan 50% overlap dan worden ze wel als verschillend beschouwd. Het blijkt dat de verdelingen van M1 en M2 elkaar 74% overlappen, van M1 en M3 is de overlap 19% en van M1 en M4 is deze 4%. De verschillen tussen M1 enerzijds en M3 en M4 anderzijds worden daarom als werkelijke verschillen beschouwd.
De methode wordt in iets meer detail uitgelegd in § 18.2, waar ook de overlappercentages voor de andere schalen waarop vooruitgang werd verwacht, worden gegeven (in tabel 22 en 23). De uitwerking van deze 'somwaardenstatistiek' wordt gegeven in bijlage 3.

V Afweging van mogelijkheden, bespreking werken met itemscores

11 Discussie over de methoden

11.1 Non-parametrische toetsen

Voor de toetsen op vooruitgang (tabel 9.1 en 9.2) zijn een t-toetsen gebruikt. De reden daarvoor is vooral praktisch. Dit is een bekende, makkelijk mee te werken toets. Als de voorwaarden ervoor aanwezig zijn (in principe normale verdeling van de scores; vergelijkbare spreidingen, hoewel er een toets is voor verschillende spreidingen) dan is dat om praktische redenen aantrekkelijk.
Dat de scores normaal verdeeld zijn, is bekend uit het onderzoek bij de schaalconstructie. Aangezien we evenwel vooral gericht zijn op toetsingsprocedures waarvoor geen onderzoek over groepen (en dus het normeringsonderzoek van de schaal) nodig is, is er veel voor te zeggen daar niet vanuit te gaan. Maar alleen uit te gaan van de (item)scores zoals men die vindt.

De McNemar-toets is een verdelingsvrije (non-parametrische) toets. Ook andere non-parametrische toetsen kunnen worden toegepast. Ook andere nonparametrische toetsen dan de McNemar-toets zijn mogelijk, bijvoorbeeld de Wilcoxon, t-toetsen over rangordewaarden (Sheskin, 2004). Het zijn de non-parametrische pendanten van de t-toets. Zulke toetsen zijn juist opgezet om onafhankelijk van de scoreverdeling en onderlinge spreidingen te toetsen. Non-parametrische toetsen zijn ook geschikt voor een beperkt aantal waarnemingen.
Ook is pure kansberekening mogelijk (een randomisatietoets, zie bijvoorbeeld Todman & Dugard, 2001), dat wil zeggen nagaan hoe groot de kans is dat, gegeven de aanvangsscores (M1) de vervolgscores (M2) mogelijk zijn. Ook daarbij kan men α kiezen.

In bijlage 2 zijn met de Wilcoxontoets voor gepaarde steekproeven dezelfde toetsingen gedaan als in § 10.6.2 met t-toetsen. Zoals men ziet levert dit in grote lijn dezelfde resultaten als die eerder verkregen werden met de t-toetsen.

11.2 Verschil in betekenis van werken met normgroepen, standaardmeetfout, RCI's en toetsen over itemscores

Wat is de waarde van de gebleken vooruitgang? Hoe groot is de vooruitgang? Mogen we er een waarde aan toekennen?
Op basis van vergelijken van de scores van de cliënt met de normgroepen poliklinische psychiatrische patiënten en de bevolking in z'n geheel, konden we alleen concluderen dat Slapen bij M4 verbeterd was, maar konden we niet tot andere duidelijke uitspraken komen over de vooruitgang.
De methode van werken met de standaardmeetfout en de RCI zijn in grote lijn vergelijkbaar. Op basis van de standaardmeetfout zijn er twee significante vooruitgangen, volgens de RCI's is er maar één vooruitgang significant; tussen M1 en M4 bij slapen.
Als we itemscores als afzonderlijke waarnemingen beschouwen en toetsen met een gepaarde t-toets, blijkt er (soms afwisselend op M2, M3 en M4) vooruitgang op alle verwachte schalen en vrijwel niet op de schalen waarop geen vooruitgang verwacht werd. Er is dus een duidelijk verband met de verwachtingen, wat maakt dat de betekenis van de naar voren gekomen vooruitgang niet als toeval aan de kant kan worden geschoven. Van deze methode kan gesteld worden dat die erg sensitief is. Kleine verschillen kunnen significant blijken (zoals bij de t-

toets hoort, wordt dat vooral bepaald door het verschil van gemiddelden en de grootte van de spreiding).
Een wezenlijk verschil is dat voor deze manier van toetsen geen gegevens nodig zijn van normgroepen.
Met de resultaten van de toetsing met de gepaarde t-toets over de itemscores was een differentiële taxatie van de behandelstrategie en het vervolg van de behandeling mogelijk. Werken met itemscores heeft dus veel voordelen. Mits men er gebruik van kan maken. De vraag is: is die methode toegestaan?

11.3 Toetsen over itemscores en onafhankelijkheid van waarnemingen. Is toetsen over itemscores toegestaan?

11.3.1 Onafhankelijke waarnemingen vereist

Statistische toetsen, de parametrische én non-parametrische, zijn gebaseerd op onafhankelijkheid van scores. Bij het toetsen over itemscores kan men de onafhankelijkheid van de afzonderlijke waarnemingen, de itemscores, op basis van twee overwegingen ter discussie stellen.
De eerste is dat de waarnemingen alle afkomstig zijn van één persoon en derhalve onderling in meerdere of mindere mate gerelateerd zijn of kunnen zijn.
De tweede is dat de itemscores die gebruikt worden als afzonderlijke waarnemingen, onderdeel zijn van een schaal. En bij de constructie van die schaal werd schaalhomogeniteit nagestreefd, dat wil zeggen dat afzonderlijke items hoog met de schaal correleren en laag correleren met andere schalen en met de items van andere schalen.
Beide overwegingen stellen de onderlinge onafhankelijkheid van de afzonderlijke itemscores ter discussie.

Dit is een wezenlijke kwestie want zowel parametrische als non-parametrische toetsen vereisen onafhankelijke observaties (Kerlinger & Lee, 2000, p. 145-149, 417-418). Kerlinger en Lee (2000) vinden dit een zeer belangrijke aangelegenheid aangezien men tot verkeerde conclusies over de resultaten kan komen door eventuele onderlinge afhankelijkheid van observaties over het hoofd te zien. Ook geven ze aan dat deze vereiste vaak enigszins wordt geschonden. Tegelijkertijd is dit ten dele vooral een kwestie hoe men onderlinge afhankelijkheid en onafhankelijkheid van observaties ziet.
Kerlinger en Lee (2000, p. 147) geven het voorbeeld van 5 leerlingen die een examen afleggen. Ze worden gerangschikt naar beoordeling, de hoogste krijgt nummer 1. Kerlinger en Lee (2000) vinden dat de observaties niet alle onafhankelijk zijn. Want als er al drie een score hebben gehad, bijvoorbeeld 1, 3 en 5, dan resteren nog slechts de nummers 2 en 4. Dit is een voorbeeld hoe men hierover verschillend kan denken. Als de examenopgaven onafhankelijk van elkaar worden beoordeeld met een schoolcijfer en de schoolcijfers worden vervolgens gerangschikt, waarna de nummers 1 t/m 5 aan de 5 beoordelingen worden gegeven, zijn de observaties naar ons idee volledig onafhankelijk. De beoordelingen werden immers volstrekt onafhankelijk gegeven. De rangschikking erna volgt puur uit die beoordelingen.
Zo geven Kerlinger en Lee (2000, p. 419) een voorbeeld van enkele dieren van wie agressieve gedragingen als observatie-eenheden gebruikt in plaats van agressieve dieren zelf als observatie-eenheid. Zij achten dit onjuist, want binnen één dier zullen agressieve gedragingen niet onafhankelijk zijn.
Ook dit is discutabel. Als deze dieren worden gezien als representatief voor de diersoort, de leeftijd, sekse en omstandigheden (anders gezegd als het specifieke dier zelf niet als variantie-

bron, als ‘oorzaak’ van verschillen, hoeft te worden gezien), dan is moeilijk in te zien waarom de agressieve gedragingen zelf geen eenheid van observatie kunnen zijn.

11.3.2 Waarnemingen bij éénzelfde persoon

De gepaarde t-toets wordt gebruikt voor het vergelijken van bijvoorbeeld pre- en postmetingen. Elke afzonderlijke pre- en postmeting is van dezelfde persoon. Men heeft van elke persoon dus een pre- en postmeting. De toets wordt uitgevoerd over een aantal personen.
Dat er sprake is van dezelfde persoon bij elke afzonderlijke pre- en de postmeting, maakt het mogelijk de correlatie tussen de deze metingen in de toets te verdisconteren. Dat maakt de toets erg sensitief, dat wil zeggen: de statistische power is groot (de type II fout laag).
Maar bij de gewone toepassing van de t-toets zijn de personen zijn onderling verschillend en hun scoringen zijn niet afhankelijk van elkaar. Als persoon A hoog scoort op een pre-meting zal hij vermoedelijk ook hoog scoren op de postmeting al kan die wel hoger of lager zijn. Maar als persoon A hoog scoort, wil dat nog niet zeggen dat B dat ook doet. Die kan best laag scoren, en ook in zijn geval kan het zijn dat de score op de postmeting hoger of lager is ondanks de lagere waarden die B scoort, dan de waarden van A.
Overigens geldt dit binnen een zekere range van de scores. Als een persoon de laagste waarde scoort, kan hij bij een volgende meting niet nog lager scoren; hetzelfde geldt voor de hoogste score waarboven bij een tweede meting niet kan worden gescoord: het bodem- en plafondeffect.

De onafhankelijkheid van waarnemingen bij een groep personen moet evenwel bezien worden in het licht van het contrast met andere groepen. Als men een experiment uitvoert waaraan Nederlandse GGZ-patiënten meedoen, dan kan men die personen en daarmee de waarnemingen als onafhankelijk beschouwen. Zou men echter ook Chinese of Indiase GGZ-patiënten in het onderzoek betrekken, dan zou heel goed kunnen blijken dat er een ‘persoons’- of ‘cultuurvariabele’ in het geding is. In die zin dat de Europeanen enerzijds en de Chinezen en inwoners van India anderzijds een verschillend patroon vertonen. De waarnemingen blijven in principe onafhankelijk, maar door de cultuurvariabele is er mogelijk een systematische invloed die niet onderkend werd zonder de Chinezen en inwoners van India in het onderzoek te betrekken. Deze systematische invloed geeft de personen onderling een zekere gerelateerdheid.
Er is een tweede systematische invloed werkzaam: voor een onderzoek naar het effect van een therapie tegen angst, zal men GGZ-patiënten werven die angstig zijn. Patiënten met lage angstscores worden niet in de onderzoeksgroep opgenomen. Hoewel in principe de waarnemingen onafhankelijk zijn, zullen de angstscores dat op een bepaalde manier niet zijn. Want lage angstscores zijn onwaarschijnlijk.

We beschouwen nu de persoon van wie de N=1-gegevens afkomstig zijn als ‘de populatie’. Binnen die ‘populatie’ worden de gegevens als onafhankelijk beschouwd. Zou men naast die persoon nog één of meer andere personen in één onderzoek betrekken, dan gaat de persoonsvariabele een rol spelen. Zolang de gegevens echter afkomstig zijn van die ene persoon, worden de data als onafhankelijk beschouwd.

11.3.3 Itemscores in plaats van schaalscores

Onderlinge afhankelijkheid van item(scores)

Schaalscores worden in de geldende methodologie als onafhankelijke waarnemingen beschouwd. Dat wil zeggen schaalscores van verschillende personen. Tegen itemscores zien als afzonderlijke (onafhankelijke) waarnemingen, is er het bezwaar van mogelijke of waarschijnlijke onderlinge afhankelijkheid. Tests of vragenlijsten worden immers zo geconstrueerd dat ze een zekere mate van homgeniteit hebben en een redelijke item-testcorrelatie alsmede een redelijke onderlinge itemcorrelatie.
De onderlinge gerelateerdheid van items van één schaal hoeft evenwel het beschouwen van de itemscores als onafhankelijke waarnemingen niet in de weg te staan. Waar het om gaat is of alle scoringspatronen (alle combinaties van itemscores) bij één schaal kunnen vóórkomen en een waarschijnlijkheid hebben die hoort bij de curve (bijvoorbeeld de normaalverdeling). En dat is bij het werken met itemscores het geval.
Indien de itemscores onafhankelijk zijn, is de scoreverdeling van de populatie continu en vloeiend verdeeld (gecorreleerdheid van itemscores toont zich bijvoorbeeld als een minder vloeiende curve omdat bepaalde combinaties discontinu meer vóórkomen dan andere).
Neem bijvoorbeeld de items van de SCL90-angstschaal (tabel 1.1). Het is duidelijk dat bij een angstig iemand een aantal van deze kenmerken aanwezig zal zijn. Maar dat iemand zenuwachtig is (item 2), betekent nog niet dat hij zal trillen (item 7) of vaak plotseling schrikt (item 23). Inhoudelijk betreffen de items verschillende aspecten of manifestaties van angst die in alle combinaties en met verschillende intensiteiten kunnen vóórkomen.
Tegelijk verwacht men dat scores op items als trillen, zenuwachtig, rusteloosheid (item 78) ongerelateerd zijn aan items van andere schalen, bijvoorbeeld Insufficiëntie van Denken & Handelen of Somatiek. Dit betekent niet dat als iemand hoog scoort op het item zenuwachtigheid en eveneens op trillen of rusteloosheid, hij op die laatste hoog scoort *omdat* hij op zenuwachtigheid hoog scoort. In dat geval zouden de items (eventueel ten dele) manifestaties zijn van hetzelfde. In dat geval zou een hoge scoring op één item van een schaal automatisch betekenen dat er een grote kans is van hoge scoring op een ander item van die schaal. Dan hebben alle scoringspatronen niet een evenredige kans van vóórkomen binnen de verdeling (bijvoorbeeld de normaalverdeling).
In het theoretisch extreme geval van onderlinge gerelateerdheid van items van een schaal is de correlatie tussen items van eenzelfde schaal r = 1.0. Als men één item een 3 geeft, krijgen alle items een 3. Alle mogelijke schaalscores hebben dan niet een gelijke waarschijnlijkheid. In dit theoretisch extreme geval zijn alle schaalscores veelvouden van de itemscores.
Zoals aangegeven, is dit niet het geval bij de items van de Angstschaal. Deze kunnen beschouwd worden als afzonderlijke, onderling onafhankelijke manifestaties van angst.
Men kan aannemen dat angst een psychosomatische toestand is die zich (wisselend) manifesteert in verschillende uitingen. Zo kan de angst van een persoon zich soms manifesteren als zenuwachtigheid, op andere momenten als trillen en op weer andere als rusteloosheid. Maar dan nog betekent dit niet dat deze manifestaties onderling afhankelijk zijn. Dat iemand trilt, betekent niet dat hij rusteloos is, zelfs niet als hij beide uitingen van angst afwisselend manifesteert.
Dat doorgaans de betrouwbaarheid van een test of schaal verbetert door het toevoegen van items impliceert een zekere onafhankelijkheid. Als items onderling immers sterk gerelateerd zijn, zou toevoegen niet veel uitmaken. Het toegevoegde item correleert immers hoog met al bestaande items.

De bestaande gegevens over de gerelateerdheid van items van één schaal en de interne consistentie en homogeniteit van die schaal betekent vooral dat de items gerelateerd zijn op basis van de verdeling van de item- en schaalscores in de populatie. In feite is dat een kunstmatige onderlinge gerelateerdheid.
Want ook al zijn items inhoudelijk onafhankelijk van elkaar, ze kunnen door de kenmerken van de schaal in de populatie (de 'base rate') onderling gecorreleerd zijn. Neem bijvoorbeeld de angstschaal. Een behoorlijk deel van de populatie zal bestaan uit mensen met weinig angst. Hun SCL90-angstitems zullen vooral de waarden 1 of 2 aannemen. En dus intercorreleren. Er is een aantal angstige mensen. Hun items zullen veel waarden 4 of 5 tellen. En daarmee zullen de items intercorreleren. Niet omdat de items inhoudelijk onderling afhankelijk zijn (zenuwachtigheid betekent niet per definitie trillen, het voorbeeld dat eerder genoemd werd), maar omdat ze in de populatie een zekere mate van gerelateerdheid vertonen op grond van het feit dat bij een aantal mensen alle kenmerken weinig vóórkomen en bij een beperkt maar wel substantieel deel van de populatie alle kenmerken véél vóórkomen. Het zijn vooral deze twee segmenten van de populatie die zorgen voor interne consistentie en homogeniteit van de schaal. Daar tussenin zijn dan de scores van mensen met een variërende mate van angst tussen geen of weinig en veel. Hun scores zullen naar verwachting onderling weinig correleren. De toets op onderling verband bij deze leden van de populatie is de werkelijke toets op inhoudelijke gerelateerdheid. Anders gezegd of zenuwachtigheid ook trillen betekent en beide manifestaties van elkaar of van hetzelfde zijn. Anders gezegd: er is een onderscheid tussen inhoudelijke (on)afhankelijkheid van items en itemscores en soms, niet vaak overigens, de feitelijke onderlinge correlatie op basis van de base-rate van een segment van de populatie.

De interne consistentie van de SCL90 is bekend. Arrindell en Ettema (2003, p. 31-34) geven Cronbach's α, een maat voor interne consistentie, en de itemtestcorrelatie r_{ij} voor 9 onderzoeksgroepen, waaronder poliklinische, psychiatrische patiënten en de gewone bevolking,
Zo ligt Cronbach's α voor poliklinische psychiatrische patiënten voor alle schalen boven de .80. En r_{ij} is voor deze categorie rond de .40. Voor de gewone bevolking varieert Cronbach's α van .76 (Hostiliteit, schaal 7) tot .97 voor de totaalscore (schaal 10). En voor r_{ij} geldt dat deze minimaal .29 is (schaal 10, totaal) en maximaal .44 (schaal 2, angst).
Dit zijn uitstekende waarden. Er is duidelijk sprake van interne consistentie.

Voor de cliënt van de casus (hoofdstuk 7) bleken de volgende correlaties tussen de items van de SCL90-schalen. Zie tabel 10.
De correlaties werden berekend over het koppelen van elke itemscore van een schaal met elke andere. Bij de 10 items van de SCL90-angstschaal (schaal 2) krijgt men dan 45 verschillende combinaties per meting. Over 4 metingen worden dat er 180.
Er zijn enkele significante correlaties ook al zijn ze laag. Deze significanties ontstaan vooral door het hoge aantal waarnemingen (kolom *N paren*[7]). Dit aantal paren is echter veel groter dan het aantal items en de waarnemingen over de paren gaan terug op de afzonderlijke itemscores. Toetst men over de onafhankelijke waarnemingen (N) dan is op één na geen enkele correlaties significant. Zie toelichting 6 bij tabel 10.
Alleen de correlatie van de items van schaal 8, Slapen, is tamelijk hoog. Dit weerspiegelt het plafond-effect (op de eerste meting is de schaalscore 14, dat is 1 punt minder dan het maximum). Als iemand hoog scoort op een bepaalde schaal, neigen alle items naar het maximum of bijna-maximum. Dat resulteert in een duidelijke correlatie, ook al zouden de items inhoudelijk onafhankelijk zijn. Zie tabel 11. In de volgende metingen verbetert de schaalscore op Slapen met 1 (M2), met 4 (M3 ten opzichte van M1) en met 6 punten (M4 ten opzichte van M1).

---

[7] Zijn er tien items zoals bij SCL90-schaal 2, Angst, dan levert dat 45 paren op. Er valt iets voor te zeggen om niet 45 waarnemingen, maar een lager aantal, bijvoorbeeld 10 aan te houden bij significantiebepaling.

Tabel 10: onderlinge correlaties van de items per schaal over alle itemscores

SCL90-schaal		M1 t/m M4[1]				
Nr.	Naam	N items[2]	N paren[3]	r [4]	p [4]	r^2 [5]
1	Agorafobie	7	84	-.1173	.144	
2	Angst	10	180	.1436*	.027	.0206
3	Depressie	16	480	.1399*	.001	.0196
4	Insufficiëntie Denken & Handelen	9	144	.0083	.922	
5	Somatiek	12	264	-.0295	.317	
6	Wantrouwen & Int.persoonlijke Sensitiviteit	18	612	.0872*	.016	.0076
7	Hostiliteit	6	60	-.1939	.069	
8	Slapen	3	12	.6098*	.035	.3719
9	Overig	7	84	.0174	.418	

Toelichting

* Significante waarde ($\alpha = 0.05$), éénzijdig toetsen: er is verwachting van een positieve correlatie.

[1] Alle correlaties over de 4 metingen M1 t/m M4.

[2] Aantal items van de schaal.

[3] Aantal scoreparen per meting waarover de correlatie berekend werd. Dit is 4 maal (M1 t/m M4) het aantal vergelijkingen per meting van elke itemscore met elke andere itemscore van dezelfde schaal.

[4] Correlatie r en bijbehorende p-waarde.

[5] r^2 in geval van een significante r-waarde. Geeft de door de correlatie verklaarde variantie. In percentages dit vermenigvuldigen met 100. Men ziet dat de correlaties van de schalen 2 en 3 zo'n 2% van de variantie verklaren, die van schaal 6 nog geen 1%, maar voor die van schaal 8 is het 45%.

[6] Er valt iets voor te zeggen niet te toetsen met het aantal paren, maar met het aantal items (4 metingen over 10 items bij de Angstschaal, wordt dus 4N = 40 in plaats van 180), omdat het aantal onafhankelijke waarnemingen 4N is. Doet men dat dan is r = .1436 (Angst) niet significant meer (Van den Brink & Koele, 2005, p. 102). $T = 0{,}1436 * \sqrt{([40-2]/[1 - \{0{,}1436\}^2)} = 0{,}897$.

T heeft een t-verdeling met N – 2 vrijheidsgraden (ν) indien de populatie bivariaat normaal verdeeld is. Dus T = t = 0,894. En $\nu = 40 - 2 = 38$. Bij $\nu = 35$ en $\alpha = 0.05$ moet t gelijk aan of groter zijn dan 1,690 (Van den Brink & Koele, 2005, p. 267, tabel 5). Bij $\nu = 40$ is die t-waarde 1,684. Interpoleren levert bij $\nu = 38$ een t-waarde van tenminste 1,686. En t = 0,897 (p = 0,198 op basis van interpoleren). Derhalve is r = 0,1436 niet significant.

Voor schaal 3 (Depressie), 6 (Wantrouwen en interpersoonlijke subjectiviteit) en 8 (Slapen) resulteren dan respectievelijk de volgende t-waarden: t = 1,033 ($\nu = 62$; p = 0,159; niet significant); t = 0,7324 ($\nu = 70$; p = 0,233; niet significant) en t = 2,433 ($\nu = 10$; p = 0,006; significant).

Tabel 11: SCL90-schaal 8, Slapen, resultaten van de 4 metingen

SCL90-schaal Slapen		Metingen			
Itemnr.	Inhoud item	M1	M2	M3	M4
44	Moeilijk in slaap	5	5	4	3
64	Te vroeg wakker	4	4	3	2
66	Onrustige/gestoorde slaap	5	4	3	3

Autoregressieanalyse

Als er veel observaties zijn, in de orde vanongeveer één per dag of meer, zullen observaties gerelateerd zijn. De observaties voorafgaand aan en volgend op een gegeven observatie zullen met die gegeven observatie hoger correleren dan met andere observaties. Deze afhankelijkheid vertekent het beeld en daarvoor moet gecorrigeerd worden. Dat kan ook goed met methoden van autoregressie (Van Beek, 2007; en de daar aangehaalde McCain & McCleary, 1979). De N=1-studies die we hier behandelen kenmerken zich echter niet door zoveel observaties per tijdsmoment. In de N=1-studies waarover we nu spreken, is men blij met één meetmoment per 3 à 4 maanden. Overigens als aan de voorwaarden voor normaliteit van de verdeling van residuen bij autoregressie niet voldaan wordt, werkt men met logistische regressieanalyse waar voor de genoemde onderlinge afhankelijkheid van observaties vlak na elkaar niet gecorrigeerd wordt, wat dan geen groot probleem geacht wordt (Van Beek, 2007, p. 42).

Standaardmeetfout en toetsen over itemscores

Er is nog een kwestie die in dit verband bezien moet worden. In hoofdstuk 5 ging het over de standaardmeetfout en in § 10.3 werd getoetst aan de hand van de standaardmeetfout. Heeft de t-toets of de McNemar-toets niet te maken met die standaardmeetfout? Is de standaardmeetfout van itemscores in dit verband van belang?
Het antwoord is dat de standaardmeetfout bij de t-toets niet van belang is. Een grote standaardmeetfout zorgt ervoor dat de gebleken (schaal- of itemscores) veel variëren. Dat betekent dat er een grote spreiding zal optreden rondom het gemiddelde (van een groep personen met schaalscores of van een aantal itemscores van een schaal). De t-toets toetst op basis van de spreiding. Een grote standaardmeetfout betekent een grote spreiding. En een grote spreiding betekent een lagere kans op significantie. Voor het werken met itemscores geldt in principe hetzelfde. Wel is het zo dat een grote (gemiddelde) standaardmeetfout voor de items van een schaal kan resulteren in een grote standaardmeetfout van de schaalscore.

11.3.4 Corrigeren of compenseren voor onderlinge gerelateerdheid van items

Het is goed mogelijk voor onderlinge gerelateerdheid van item(scores) te compenseren of te corrigeren. Het probleem is niet dat items eventueel onderling gerelateerd zijn, het probleem is dat eventuele onderlinge gerelateerdheid afbreuk doet aan het schatten van de relatie tussen vóór- en námeting. Als men in staat zou zijn bij het bepalen van verschillen tussen vóór- en námeting te compenseren of te corrigeren voor onderlinge afhankelijkheid van items, is deze onderlinge afhankelijkheid geen probleem.
Nu is dat goed mogelijk. Neem bijvoorbeeld M1 en M4 en de itemscores op de SCL90-schaal Slapen. De itemscores voor de items 44, 64 en 66 van de SCL90-schaal slapen bij M1 zijn 5, 4 en 5, die bij M4 zijn 3, 2 en 3. In tabel 12, hierna, is dit kolom (1). Zie ook bijlage 1. We vergelijken dan voor M1: 5 met 4, 5 met 5 en 4 met 5. Voor M4: 3 met 2, 3 met 3 en 3 met 3. De waarden waarmee wordt vergeleken staan in tabel 12 staan in kolom (3).
Men kan de correlatie berekenen tussen de scores van M1 en van M4 en vervolgens de correlatie tussen een dummyvariabele met waarden 0 (M4) en 1 (M1). De correlatie tussen de dummyvariabele en de slaapscores geeft de relatie tussen beide, de correlatie tussen de items geeft de onderlinge afhankelijkheid (als het ware bepaling van de covariantie). We krijgen de volgende tabel.

Tabel 12: Correctie voor onderlinge afhankelijkheid items

Meting	SCL90-item	Score (1)	Dummy-variabele (2)	Onderlinge itemcorrelatie, correleren met score (3)
M1	44	5	1	4
	64	4	1	5
	66	5	1	5
M4	44	3	0	2
	64	2	0	3
	66	3	0	3

- Correlatie Dummyvariabele (2) met score (1): $r_{12} = .90$ ($p = .007$, éénzijdige toets).
- Correlatie items onderling, elk item binnen M1 en M4 met elk ander item: score (1) met vergelijkingswaarden (3): $r_{13} = .73$ ($p = .05$, eenzijdige toets).
- Correlatie Dummyvariabele (2) met vergelijkingswaarden voor onderlinge correlatie items (3): $r_{23} = .90$ ($p = .007$ (éénzijdige toets, identiek aan correlatie dummyvariabele met score).
- Correlatie Dummyvariabele (2) met itemscore (1), gecorrigeerd voor de invloed van de correlatie tussen de items onderling (3): $r_{12.3} = .84$ ($p = .037$, éénzijdige toets).
- De partiële correlatie tussen de scores (1) en de dummyvariabele (2), met de invloed van de onderlinge relatie tussen de items (3) 'uitgepartialiseerd' (daarvoor gecorrigeerd), geeft dan de relatie waarin we geïnteresseerd zijn. Deze is significant. Zie ook tabel 29 in bijlage 2 over hetzelfde.

Hierbij blijft dus de correlatie tussen de dummyvariabele en de scores (die de verandering weergeeft) onderscheidbaar van de onderlinge relatie tussen (afhankelijkheid van) de items.

Een probleem hierbij is dat dit met drie items goed gaat. Er zijn dan drie onderlinge relaties tussen alle items. Maar bij meer items ontstaan er meer mogelijke itemparen. Bij 6 items (bijvoorbeeld SCL90-schaal 7, Hostiliteit, tabel 10) is dit aantal 15, bij 7 is het 21 en bij 10 wordt het 45. Het koppelen van elke itemscore aan het gemiddelde van alle andere, is geen oplossing (vaak ontstaat dan een negatieve correlatie: een hoge score wegnemen, leidt immers tot gemiddeld een lagere score voor de andere items evenals het omgekeerde). Wat men kan doen is de 1- en 0-waarden van de dummyvariabele koppelen aan alle combinaties van alle items (dat worden er dus 15 bij 6 items), maar het significantieniveau bepalen op basis van de oorspronkelijke 6 ($n = 6$).

Het is ook goed doenlijk om voor onderlinge afhankelijkheid van items te corrigeren via de verklaarde variantie, wat betekent dat de p-waarde wordt verhoogd en dat dus minder snel significantie resulteert. In tabel 10 worden de onderlinge correlaties en de kwadraten (door de correlaties verklaarde varianties) gegeven. Indien deze verklaarde variantie laag is (bijvoorbeeld één of enkele procenten) is een correctie, zelfs als men deze wenselijk zou vinden, tamelijk overbodig. Indien de correlatie hoog is, zoals bij schaal 8, Slapen, en men vindt ondanks de in § 11.3.3 gegeven argumentatie, toch dat deze correlatie de onafhankelijkheid van de waarnemingen waarop de inferentiële statistiek gebaseerd is, ondergraaft, dan kan men ook corrigeren door het aantal waarnemingen aan te passen, dat wil zeggen te verlagen waardoor een verschil minder snel significant wordt. Is de correlatie tussen de items, zoals bij Slapen, $r = .667$ wat een verklaarde variantie van 44,5% (→ 45%) betekent, dan kan men p-waarden zoeken bij 55% van n (aantal waarnemingen) en eventueel interpoleren. Dit geldt in

zijn algemeenheid. Bij de schaal Slapen met slechts 3 scores, is dat niet een werkbare procedure

De onderlinge afhankelijkheid van items kan ook bezien worden vanuit de optiek van tijdreeksen (time series). Bij time series gaat echter het doorgaans om een (lange) reeks waarden. Zoals temperatuurwaarden door de jaren of door decennia, inkomsontwikkeling, prijzen van goederen, van huizen etc. Die waarden zijn niet onderling onafhankelijk. We hebben echter gestreefd naar een in principe zo eenvoudig mogelijke procedure.

11.3.5 Regressie naar het gemiddelde

Een bezwaar dat denkbaar is, is regressie naar het gemiddelde. Als iemand, zoals de cliënt van de N=1-behandeling, hoog scoort op de items van een schaal, in dit geval Slapen, dan kan een tendens denkbaar geacht worden dat de scores bij de verviolgmeting(en) lager scoren op grond van toeval, dat wil zeggen kansberekening. En dus niet door een verandering van (het gedrag van) de persoon.
Voor regressie naar het gemiddelde is in principe te corrigeren, maar de eerste check of er wellicht sprake van is, zijn in dit voorbeeld schalen met lage itemscores. Als er sprake is van regressie naar het gemiddelde, moeten de waarden van deze items bij een vervolgmeting (enigszins) stijgen. Is dit niet het geval (en het is in onze ervaring zelden het geval) dan pleit dit tegen het aanwezig zijn van regressie naar het gemiddelde.

11.4 Kanskapitalisatie en correctie

Een belangrijke kwestie betreft het uitvoeren van toetsen over een aantal verschillen en de ‘kanskapitalisatie’ die zich daarbij voor kan doen. Dit werd al genoemd in de Box 1. Bij $\alpha = 0.05$ is er (gemiddeld) een kans van 1 op 20 dat ten onrechte wordt besloten tot een vooruitgang (dat de nulhypothese van geen vooruitgang ten onrechte verworpen wordt ten faveure van de alternatieve hypothese van vooruitgang), de type I fout (ten onrechte tot een verband besluiten).
Gaat men meer hypothesen (van vooruitgang) toetsen *over dezelfde onderzoeksgegevens* dan stijgt de kans dat één of meer ervan ten onrechte significant is/zijn. De 5%-kans op een ten onrechte significante uitslag geldt voor de gegevens van één hypothese bij één steekproef. Heeft men twee hypothesen die men wil toetsen, dan moet voor de tweede hypothese weer een nieuwe steekproef worden getrokken (nieuw onderzoek worden gedaan). Dit is praktisch erg moeilijk te realiseren, waardoor vaak (eigenlijk altijd) meer hypothesen getoetst worden over dezelfde onderzoeksgegevens. Als men verbanden/samenhangen tussen twee of meer variabelen wil nagaan, kan dat alleen maar met één onderzoeksgroep.
Dit kan verschillende omstandigheden opleveren.

- Het kan dat er maar één hypothese uit een aantal een siginficant resultaat oplevert. Het is bijvoorbeeld mogelijk dat men 20 hypothesen toetst en er één significant is. Men moet dan ernstig rekening houden met de mogelijkheid dat dit ene significante resultaat op toeval berust.
  Anders gesteld: als er in werkelijkheid geen effect zou zijn (wat men niet weet, want daarvoor doet men juist het onderzoek) dan is de kans op ten onrechte besluiten tot effect bij één hypothese 5%. Maar bij meer (onafhankelijke) hypothesen dat (zie ook): $1 - (1 - \alpha)^k$ waarbij k het aantal onafhankelijke hypothesen is.

Als men bijvoorbeeld 8 (onafhankelijke) hypothesen toetst met $\alpha = 0.05$, dan is de gemiddelde kans $1 - (1 - 0.05)^8 = 0.337$ (34%) dat één of meer hypothese(n) van deze 8 (men weet uiteraard niet welke) ten onrechte als bevestigd wordt gezien.
In zijn algemeenheid geldt dat wanneer een aantal (onafhankelijke) hypothesen getoetst wordt over dezelfde onderzoeksgegevens en er blijkt er maar één significant, men erg terughoudend moet zijn om te verklaren dat deze hypothese bevestigd is. Dit geldt temeer naarmate er meer hypothesen getoetst worden over diezelfde onderzoeksgegevens.
Het kan zijn dat er slechts enkele hypothesen significant zijn van de bijvoorbeeld 20. Dan heeft men een redelijke kans dat één ervan op toeval berust. Dat hangt mede af van hoeveel hypothesen men toetst en ook hoe ze onderling aan elkaar in een theoretisch raamwerk gerelateerd zijn. Zijn het 'losse' hypothesen, zonder een tevoren duidelijk gespecificeerd verband, dan is de kans op toevalssignificanties duidelijk aanwezig.

- Het andere uiterste is ook denkbaar: er zijn bijvoorbeeld 20 hypothesen en toetsing levert voor alle een significant resultaat op. Gemiddeld is de kans 64% dat één of meer ervan ten onrechte significant zijn.
- Gewoonlijk verkeert men in een middenpositie. Enkele hypothesen leiden tot significante resultaten, andere niet.

De correctie voor het toetsen van meer hypothesen over dezelfde onderzoeksgegevens, is de zgn. Bonferroni-correctie. Deze houdt in dat bij meer vergelijkingen de gekozen α gedeeld wordt door het aantal vergelijkingen (Nakagawa, 2004; Walsh, 2004). In de tabellen 9 zijn er 10 schalen en 30 vergelijkingen (10 vergelijkingen van M1 met M2, idem van M1 met M3 en van M1 met M4).
In geval van $\alpha = 0.05$ moet men dus bij 30 hypothesetoetsingen een gecorrigeerde α van 0,05/30 = 0,0017 aanhouden. De p-waarde die resulteert uit een verschiltoetsing, moet dus kleiner zijn dan 0,0017, wil men tot significantie besluiten. Dat is een erg lage waarde. Omwille van gemiddeld 1,5 vergelijking op 30 met een grote toevalskans op significantie wordt de statistische power van de overige 28,5 vergelijkingen aanzienlijk aangetast.
Het kritiekpunt op Bonferroni-correcties is dan ook de sterke verlaging van de statistische power (forse vergroting van de type II fout van ten onrechte niet tot een in werkelijkheid wel bestaand verband besluiten). Zoals in § 10.3 naar voren kwam is de type II fout bij een 'werkelijk' verschil van M1 en M2 van 4 punten op de SCL90-angstschaal (schaal 2) al 84%. Door te toetsen over itemscores werd dit verlaagd tot 45%. Het toepassen van de Bonferroni-correctie verhoogt de type II fout nog aanmerkelijk boven die 84%.

Het probleem van het toetsen van meer hypothesen over dezelfde steekproefdata, zou nog wel meevallen als het werkelijk zo zou zijn dat er alleen maar een kans is van 1 op de 20 getoetste onderzoekshypothesen die ten onrechte worden aangenomen. Dat valt te overzien. Als 20 onderzoekshypothesen worden getoetst, en bijvoorbeeld 4 een significant resultaat geven, dan zou het bij 1 daarvan ten onrechte kúnnen zijn (ze kunnen overigens ook alle 4 'terecht' significant zijn). Hiermee valt te leven als de onderzoekshypothesen in een goed theoretisch raamwerk zijn ingebed. En indien in de rest van de data en/of in ander onderzoek indirecte bevestigingen voor de bevestigde onderzoekshypothesen gevonden worden.
Maar een kans van 1 op 20 op ten onrechte significantie is een gemiddelde van de denkbare onderzoeken. Als men het onderzoek een aantal keren zou herhalen dan is gemiddeld 1 op de 20 hypothesen ten onrechte significant (1 op de 20 keer wordt de nulhypothese ten onrechte verworpen). Maar het kan ook gebeuren dat er 2 of 3 onterechte significanties zijn.

De situatie wordt iets gecompliceerder, maar ook gunstiger door inzake álle variabelen (schalen) verwachtingen te formuleren. Als er geen vooruitgang verwacht wordt, is het uitblijven

daarvan de verwachting. Anders gezegd: er zijn schalen waarbij men de nulhypothese wil verwerpen (als er vooruitgang verwacht wordt) en er zijn schalen waarbij men de nulhypothese niet wil verwerpen (als er geen vooruitgang verwacht wordt). Het geheel resulteert in een patroon dat als zodanig ook een uitslag is. Hoe meer het verwachte patroon blijkt, hoe sterker steun voor de gebleken vooruitgang én de onderliggende behandeltheorie. Als het verwachte patroon zich in toenemende mate manifesteert bij volgende metingen (immers als de verwachtingen valide waren zullen ze met het vorderen van de behandeling meer uitkomen), dan is dat ook een bevestiging van de vooruitgang en de onderliggende behandeltheorie.

Om de kans te verlagen tot conclusies op basis van toeval wanneer men Bonferroni-correcties achterwege laat, kan het volgende gedaan worden.

A De hypothesen goed inbedden in een theoretische raamwerk en/of behandeltheorie. De hypothesen zijn onderling ook gerelateerd. Geen 'losse' hypothesen.

B Het aantal vergelijkingen in eerste instantie beperken. Men toetst dan bijvoorbeeld niet meer tussen M2, M3, M4 en eventuele latere onderling (zoals hier wel gedaan is in de tabellen 9 van § 10.3).
Men toetst voorts in eerste instantie alleen de verwachte vooruitgang en niet over de variabelen waarop geen vooruitgang werd verwacht. In de voorbeeldcasus die gebruikt is, zou dat betekenen dat men terug gaat van 30 hypothesen naar 12 (alleen de SCL90-schalen Angst, Insufficiëntie van Denken & Handelen, Depressie, Slapen), dus 4 schalen en 3 verschillen (M1 met M2, M1 met M3, M1 met M4).

C Nagaan, zoals ook hier gedaan is, of er verschil is wanneer men de niet-verwachte vooruitgang ook éénzijdig toetst, net als bij de wel verwachte vooruitgang. Als er bij enkele schalen waarop geen vooruitgang verwacht werd dan wel significanties blijken, is dat een aanwijzing van toevalsinvloeden. Blijken de schalen waarop geen verschil verwacht werd bij éénzijdig toetsen in meerderheid ook niet significant én de vooruitgang treedt daar op waar die verwacht wordt, dan is dat een sterke aanwijzing van vooruitgang, los van toevalsinvloeden.

D Pas na *A*, *B* en *C* worden eventuele andere significanties in overweging genomen, zoals op schalen waarop geen vooruitgang werd verwacht. En dat dan uitsluitend in het kader van een zo compleet mogelijk beeld van wat er gebeurd is in de behandeling tot dat moment.
Het kan namelijk gebeuren dat er ook niet-verwachte vooruitgang is op één of meer schalen en dat die vooruitgang iets zegt over de behandeling of over het aan de behandeling te grondslag liggende behandelplan (behandeltheorie, functieanalyse etc.). Als zich echter geen verwachte vooruitgang manifesteert, maar wel niet-verwachte, dan heeft die geen waarschijnlijk geen betekenis.
De uitslagen (gebleken significanties) kunnen als patroon worden bezien. Hoe meer ze gaan in de richting van de verwachting(en) en van impliciete verwachtingen van de behandeltheorie, hoe groter de kans dat een significantie 'terecht' is. Dit wordt nog versterkt wanneer er een patroon is conform de verwachtingen dat zich steeds sterker manifesteert. Zoals in het geval van de casus dat de vooruitgang op schaal 8, Slapen, bij elke volgende meting groter wordt. Overigens geldt dit vooral wanneer er bij meting 2 nog geen grote overeenkomst is, bijvoorbeeld zo'n 40%. Als de overeenkomst eenmaal circa 80% is, is een verdere toename nog steeds goed mogelijk, maar valt eerder te verwachten dat de overeenkomst in die orde van grootte blijft.
Dit is ook in een coëfficiënt uit te drukken. Als er, zoals in de casus, op 5 schalen vooruitgang verwacht wordt en op 5 niet, wat 10 voorspellingen oplevert inzake vooruitgang en deze manifesteert zich zoals hierna in tabel 12 is aangegeven: 8 van de 10 voorspellingen komen uit, één komt half uit (er werd op schaal 10 een niet-sginficante vooruitgang

verwacht, dat wil zeggen een vooruitgang die op niveau $\alpha = 0.10$ significant zou zijn). Aldus is 8½ voorspelling uitgekomen: 85%. Dat een sterke aanwijzing voor vooruitgang als gevolg van de behandeling én daarmee ook van de onderliggende behandeltheorie.

Om een idee te krijgen van de invloed en betekenis van toeval bij verschiltoetsen, kan men volledig willekeurig mensen bijvoorbeeld een SCL90 afnemen en willekeurig enkele schalen bepalen waarop 'vooruitgang' verwacht wordt. Vervolgens toetst men of dat zo is, waarbij de schalen met verwachte 'vooruitgang' éénzijdig getoetst worden en de andere tweezijdig, en daarna ter vergelijking alle verschiltoetsen éénzijdig. Er mag dan geen verband blijken, de schalen van verwachte 'vooruitgang' mogen niet vooruitgang laten zien in vergelijking met de andere en tussen één- en tweezijdig toetsen mag geen groot verschil zijn. Is dat wel het geval dan is de methode gevoelig voor toeval en mag men uitgaan van kanskapitalisatie, ook bij toetsen waarbij werkelijk op grond van een uit te voeren behandeling op tevoren aangemerkte schalen vooruitgang verwacht mag worden.

VI Verdere bewerkingen en beslissingen inzake de behandeling, uitgaand van analyse over itemscores

12 Bewerking van de resultaten van toeten over itemscores

12.1 Overeenkomst tussen verwachtingen en uitslagen

Nu naar onze mening aangetoond is dat het werken met afzonderlijke itemscores als waarnemingen een adequate manier is om vooruitgangsverschillen te toetsen, kan bezien worden wat voor mogelijkheden deze manier van verschillen vergelijken nog meer biedt.
Om informatie te verkrijgen over het verloop van een individueel behandelproces in relatie tot de gestelde behandeltheorie en verwachtingen, kunnen de resultaten van de toetsingen over itemscores nog verder worden bewerkt.
Voortgaand op de resultaten van tabel 9.1, kan de relatie tussen de verwachting en de gebleken vooruitgang worden gespecificeerd. Daartoe wordt de uitslag per schaal vergeleken met de verwachting. Hoe meer in positieve zin gerealiseerd wordt van de verwachting, hoe beter. Daarbij telt het realiseren van een verwachte significante vooruitgang het meest, maar ook significante vooruitgang waar die niet werd verwacht, telt mee, zij het minder.
Hierbij wordt vooreerst uitgegaan van de uitslagen bij éénzijdig toetsen voor schalen waarop vooruitgang werd verwacht en tweezijdig toetsen bij schalen waar geen verschil verwacht werd. Zou men uitgaan van uitslagen bij éénzijdig toetsen bij alle schalen, dan is er sneller kans op vooruitgang bij de schalen waarop geen vooruitgang verwacht werd. En aangezien elke vooruitgang meetelt, ook die welke niet verwacht werd (al telt dat minder mee), bevoordeelt men daarbij de vooruitgangsindices.
In tabel 13 zijn de verwachtingen weergegeven van M1 naar M2. Uitgangspunt was tabel 9.1. Zoals men kan zien, zijn de uitslagen in grote lijn volgens de verwachtingen. Negen van de 10 verwachtingen komen uit bij M2: 90%. Men zou het ook kunnen kwantificeren en een uitgekomen verwachting gewicht 1 te geven, een niet-significante vooruitgang ($\alpha = 0.10$) een gewicht van ½ te geven.

Tabel 13: verschil meting 1 (M1) en meting 2 (M2)

Schaal SCL90	S +	NS +	0	NS -	S -
1 Agorafobie			●		
2 Angst	●				
3 Depressie	●				
4 Insufficiëntie van denken & handelen	●				
5 Somatische klachten			●		
6 Wantrouwen & interpersoonlijke sensitiviteit			●		
7 Hostiliteit			●		
8 Slaapproblemen			●		
9 Overige			●		
10 Totaal	●●				

Toelichting

S+	Significante vooruitgang (afname), α = 0.05, éénzijdig toetsen bij verwachte vooruitgang, anders tweezijdig
NS+	Niet-significante vooruitgang (afname), α = 0.10, idem
0	Geen verschil (gelijk blijven), α = 0.10, idem
NS-	Niet-significante achteruitgang, α = 0.10, éénzijdig toetsen (toelichting hierna)
S-	Significante achteruitgang, α = 0.05, idem
	Verwachting
•	Gebleken vooruitgang
••	Gebleken significantie op niveau α = 0.01

Van alle 10 verwachtingen zijn er bij M2 8 volledig uitgekomen en één (schaal 10) voor de helft. Dus 8,5 op 10, is 85%. Dit werd al aangegeven in § 11.4 onder *D*. Op dezelfde manier kan berekend worden dat de overeenkomst met de verwachtingen bij M3 75% is en bij M4: 80%.

De gebleken overeenstemming tussen de verwachtingen en de uitkomsten kan in een matrix worden weergegeven. Er kan dan ook een 'overeenstemmingscoëfficiënt' worden berekend: Cohen's κ (kappa; Van den Brink & Koele, 2005, p. 214-218), een overeenstemmingsmaat met maximum 1 en 0 bij afwezigheid van overeenstemming.
Men onderscheidt de uitslagen S (S is zowel S als S! dus zowel significantie op niveau α = 0.01 als α = 0.05), NS, GV, NSA (niet-significante achteruitgang) en SA (significante achteruitgang). (Niet-)significante achteruitgang bestaat eigenlijk niet. Wat bedoeld wordt is: indien tevoren achteruitgang gehypothetiseerd zou zijn, zou deze dan significant zijn bij tweezijdig toetsen? Als een matrix gemaakt wordt met als rijen de verwachting en als kolommen de uitkomst, dan liggen alle uitgekomen verwachtingen op de diagnonaal. Hoe meer een uitkomst afwijkt van de verwachting, hoe verder van de diagonaal hij af ligt.
In de tabellen 13 (13.1 en 13.2) geeft de meest linkse kolom de SCL90-schaal, de kolom 'Verwacht' geeft de verwachtingen en de kolom 'Gebleken' de uitslagen. We zien dan dat er 5 keer GV verwacht wordt (schaal 1, 5, 6, 7 en 9), daarvan is de uitslag ook 6 keer GV, 1 keer NS (schaal 10) met als uitslag S en 1 keer S (schaal 8), uitslag GV. In een matrix zien de uitslagen van tabel 9 er als volgt uit. Zie tabel 14.1.

Tabel 14.1: overeenkomst verwachtingen en uitslagen van tabel 13

Uitslagen (gebleken)

Verwacht	S(!)	NS	GV	SA	NSA
S(!) S of S!	3		1		
NS	1				
GV			5		
NSA					
SA					

S(!): Significante vooruitgang op α = 0.01 en α = 0.05.
NS: Niet-significante vooruitgang, trend (significantie bij α = 0.10)
GV: Geen verschil
NSA: Niet-significante achteruitgang
SA: Significante achteruitgang

Men ziet dus bij 10 verwachtingen 8 volledig conform de verwachting, bij 1 is er één niveau verschil (S tegen NS), bij 1 is er twee niveaus verschil (S tegen GV). In percentages uitgedrukt staat elke score op de diagnonaal (nu 3 en 5) voor het aantal volledige overeenstemmingen (dus 8). Indien een score één horizontale cel af ligt van de diagonaal is dat 50% overeenstemming. Aldus zijn er 8 overeenstemmingen en 1 halve overeenstemming van de 10. Zoals al aangegeven dus 85%.

Zou alles precies volgens de verwachtingen gegaan zijn, dus 100% overeenstemming, dan zou dat er hebben uitgezien als in tabel 14.2.

Tabel 14.2: overeenkomst indien alle uitslagen van tabel 12 volgens de verwachtingen waren geweest

Uitslagen (gebleken)

Verwacht	S(!)	NS	GV	SA	NSA
S(!) S of S!	4				
NS		1			
GV			5		
NSA				0	
SA					0

Om na te gaan of de gebleken uitslag (tabel 14.1) afwijkt van de verwachte (tabel 14.2) wordt κ berekend. Het blijkt: $\kappa = 0{,}63$. Onder de nulhypothese H_0: $\kappa = 0$ geldt dat:

$$K = \frac{n\Sigma O_{ii} - \Sigma O_{i.} \times O_{.i}}{n^2 - \Sigma_i O_{i.} \times O_{.i}}$$ (waarin O de celwaarden zijn) normaal verdeeld is met $m = 0$

en: $$S_\kappa^2 = \frac{n^2\Sigma O_{i.} \times O_{.i} + (\Sigma O_{i.} \times O_{.i})^2 - n\Sigma O_{i.} \times O_{.i} \times (O_{i.} + O_{.i})}{n\,(n^2 - \Sigma O_{i.} \times O_{.i})^2}$$

$Z = \kappa / \sqrt{(S_\kappa^2)} = 2{,}3258$. Dit resulteert in $p < 0.01$ (rechteroverschrijdingskans). De H_0 van geen verband kan dus verworpen worden ten gunste van een verband tussen de verwachting en gebleken uitslagen.

12.2 Waarderen van vooruitgang

We keren ons nu van de overeenkomst van de uitslagen met de verwachtingen, naar de vooruitgang zelf. Want niet verwachte vooruitgang is wel vooruitgang. En als het gaat om het waarderen van de behandeling, is ervoor gekozen ook niet-verwachte vooruitgang daarbij mee te laten tellen, zij het minder dan de verwachte vooruitgang.

Aan de relaties tussen verwachte en gebleken vooruitgang, gelijk blijven of achteruitgang worden waardes gegeven. Zie tabel 15.

Tabel 15: Waarden voor gebleken vooruitgang in relatie tot de verwachting

Verwachting	Gebleken	Waarde
Significante vooruitgang	Significante vooruitgang ($\alpha = 0.01$) [1]	3
Significante vooruitgang	Significante vooruitgang ($\alpha = 0.05$) [1]	2
Significante vooruitgang	Geen signif. vooruitgang, wel trend ($\alpha = 0.10$) [1]	1
Significante vooruitgang	Geen verschil[1]	0
Niet-signif. vooruitgang	Significante vooruitgang ($\alpha = 0.01$) [1]	1½
Niet-signif. vooruitgang	Significante vooruitgang ($\alpha = 0.05$) [1]	1
Niet-signif. vooruitgang	Niet-signif. vooruitgang, wel trend ($\alpha = 0.10$) [1]	½
Niet-signif. vooruitgang	Geen vooruitgang, geen verschil [1]	0
Vooruitgang (sign. of niet)	Significante achteruitgang ($\alpha = 0.01$ en $\alpha = 0.05$) [3]	-2
Vooruitgang (sign. of niet)	Niet-signif. achteruitgang, wel trend ($\alpha = 0.10$) [3]	-1
Geen verschil	Geen verschil [2]	0
Geen verschil	Significante vooruitgang ($\alpha = 0.01$) [2]	1
Geen verschil	Significante vooruitgang ($\alpha = 0.05$) [2]	½
Geen verschil	Niet-signif. vooruitgang, wel trend ($\alpha = 0.10$) [2]	¼
Geen verschil	Significante achteruitgang ($\alpha = 0.05$) [2]	-1½
Geen verschil	Niet-signif. achteruitgang, wel trend ($\alpha = 0.10$) [2]	-1
Verwachte achteruitgang	Geen verschil ($\alpha = 0.05$) [3]	0
Verwachte achteruitgang	Significante achteruitgang ($\alpha = 0.01$ en $\alpha = 0.05$) [3]	-½
Verwachte achteruitgang	Niet-signif. achteruitgang, wel trend ($\alpha = 0.10$) [3]	0
Verwachte achteruitgang	Significante vooruitgang ($\alpha = 0.05$) [3]	½
Verwachte achteruitgang	Niet-significante vooruitgang ($\alpha = 0.10$) [3]	0

Tabel 15, toelichting

[1] Eénzijdig toetsen: verwachting vooruitgang.

[2] Tweezijdig toetsen: geen verwachting. Ter vergelijking kan men nagaan of éénzijdig toetsen van verschillen bij schalen waarop geen vooruitgang verwacht werd, een ander beeld geeft (zie de eerdere opmerkingen hierover in de toelichting bij tabel 9.1 en 9.2). Dat wil zeggen of er dan meer significanties resulteren (eenzijdig toetsen geeft sneller significantie). Indien eenzijdig toetsen dan duidelijk meer significanties geeft dan tweezijdig toetsen (bij variabelen waarover geen verschil verwacht werd), is dat een aanwijzing voor 'kanskapitalisatie' (toevalsinvloeden op de resultaten).

[3] Strikt genomen kan er bij éénzijdig toetsen geen significante uitslag 'in de andere richting' zijn. De toetsingsgrootheid (t) valt in het gebied dat is aangegeven ingeval van juistheid van H_0. Er kan alleen gesteld worden dat ingeval men vooraf gesteld zou hebben dat er wél een verschil zou optreden, zonder te specificeren wat dit zou zijn (vooruitgang of achteruitgang), er bij tweezijdig toetsen een significant verschil zou zijn.[8]

Het gaat bij dit alles om *de vraag wat de vooruitgang is en hoe die gewaardeerd wordt*. Het gaat *niet om de vraag of de verwachtingen zijn uitgekomen*. Die vraag is al behandeld bij tabel 14 en bij punt *D* in § 11.4.

[8] Men toetst éénzijdig als er een specifieke hypothese is, bijvoorbeeld test B is hoger dan test A. Men toetst tweezijdig zonder specifieke hypothese, maar alleen de vraag of er überhaupt sprake is van een verschil. Bijvoorbeeld: $A \neq B$.

Optellen van de waarden die zouden resulteren als alle verwachtingen zouden uitkomen, levert een verwachte totaalwaarde op, optellen van gebleken waarden eveneens. Men kan de resulterende gebleken totaalwaarde uitdrukken in een percentage van de verwachte (gebleken waarde/verwachte waarde x 100%). Dit levert het volgende op voor het verschil van meting 1 (M1) naar meting 2 (M2).

Tabel 16: Verschil M1 en M2

Schaal	Verwacht	Gebleken	Waarde	
			Conform verwachting	Gebleken
1	GV	GV	0	0
2	S	S	2	2
3	S	S	2	2
4	S	S	2	2
5	GV	GV	0	0
6	GV	GV	0	0
7	GV	GV	0	0
8	S	GV	2	0
9	GV	GV	0	0
10	NS	S!	0,5	1,5
Totaal			8,5	7,50
Gehaald (7,50/8,5 × 100% =)			88,24% → 88%	

GV: Geen verschil
NS: Niet-significante vooruitgang, wel trend (zou significant zijn bij $\alpha = 0.10$)
S: significante vooruitgang ($\alpha = 0.05$)
S!: significante vooruitgang ($\alpha = 0.01$)

Van M1 naar M2 is er een totaalwaarde van 7,50 behaald, waar 8,5 behaald geworden zou zijn als alle uitslagen volgens de verwachting waren. Ofwel: een resultaat van 7,50/8,5 x 100 = 88%.

Dit kan ook gedaan worden voor de verschillen van M1 met M3 en van M1 met M4.

Tabel 17: Verschil M1 en M3

Schaal	Verwacht	Gebleken	Waarde	
			Conform verwachting	Gebleken
1	GV	GV	0	0
2	S	NS	2	1
3	S	NS	2	1
4	S	NS	2	1
5	GV	GV	0	0
6	GV	GV	0	0
7	GV	GV	0	0
8	S	S	2	2
9	GV	GV	0	0
10	NS	S!	0,5	1,5
Totaal			8,5	6,5
Gehaald (6,5/8,5 × 100% =)			76,47% → 76%	

Tabel 18: Verschil M1 en M4

Schaal	Verwacht	Gebleken	Waarde	
			Conform verwachting	Gebleken
1	GV	GV	0	0
2	S	NS	2	1
3	S	NS	2	1
4	S	S	2	2
5	GV	GV	0	0
6	GV	GV	0	0
7	GV	GV	0	0
8	S	S	2	2
9	GV	GV	0	0
10	NS	S!	0,5	1,5
Totaal			8,5	7,50
Gehaald (7,5/8,5 × 100% =)			88,24% → 88%	

Bij de tweede meting wordt aldus 88% van de (uiteindelijk) verwachte en nagestreefde vooruitgang gerealiseerd (ten opzichte van de eerste), bij de derde 76% (ook ten opzichte van de eerste) en bij de vierde weer 88% (eveneens ten opzichte van de eerste). Zie grafiek 4: op de Y-as de percentages, op de X-as de meetmomenten en het therapieverloop. Dit wordt in feite al weergegeven in de grafiek 3. Daarmee kan gezegd worden dat de primaire doelstelling, verbetering van slapen, is gehaald. Over het geheel gezien gaat de uitslag in de richting van de verwachting, maar komt niet op 100%. De verwachtingen waren geformuleerd als eindverwachtingen, dus het uiteindelijk te behalen resultaat.

Op basis van de tabellen 15, 16 en 17 kunnen wel inhoudelijke afwegingen gemaakt worden over het vervolg van de behandeling. Wat niet kon met de vergelijkingen op basis van normklassen en de RCI's.
We zien dat op Angst, Depressie en Insufficiëntie van Denken & Handelen (schalen 2, 3 en 4) vooruitgang is bij M2. Er is (nog) geen vooruitgang op slapen (schaal 8). De therapie lijkt iets te bewerken, maar het is nog te vroeg (3 maanden na het begin) om het uiteindelijk beoogde resultaat al te hebben gerealiseerd. Bovendien zijn er genoeg redenen die verklaren waarom de vooruitgang niet groter is (zie de casus in § 7). Cliënt piekert in deze periode juist meer over z'n slapen nu hij de diagnose autismespectrumstoornis heeft gehoord. Ook de therapeutische sessies geven hem spanning aangezien hij alles zo perfect mogelijk wil doen en de interactie met de therapeut hem vaak gespannen maakt. Er blijft dus ingezet worden op angst en spanning (de schalen 2, 3 en 4), maar er wordt meer werk gemaakt van slapen (schaal 8).
Bij M3, 6 maanden na het begin, zien we dat het slapen nu ook beter gaat, maar dat het resultaat op Angst, Depressie en Insufficiëntie van Denken & Handelen iets is teruggevallen.
De keus is nu:

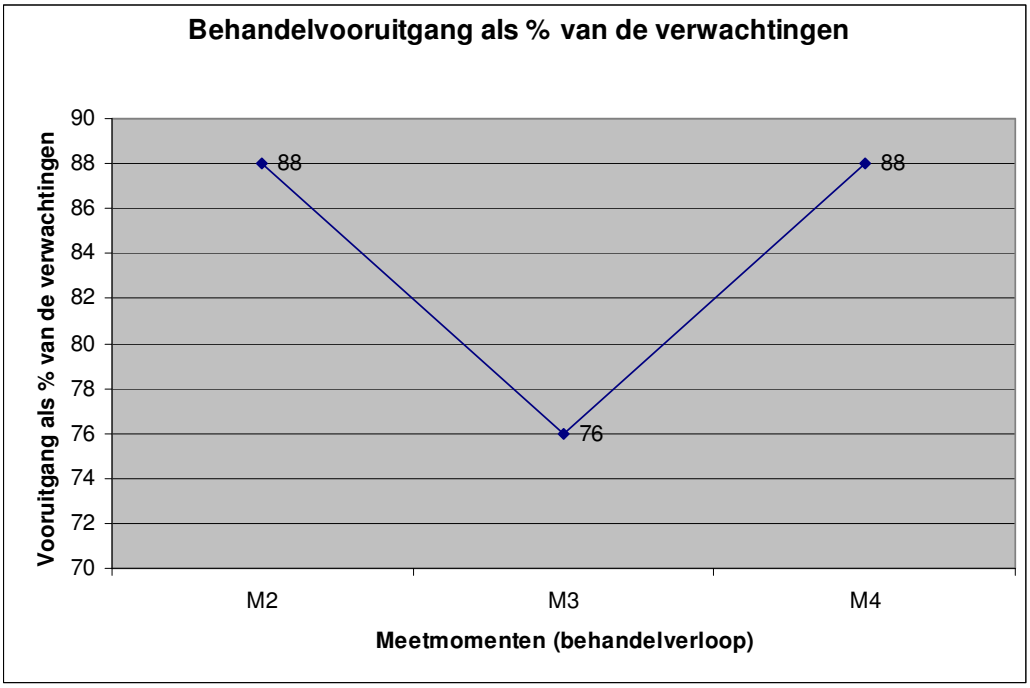
1 wordt voortgegaan met de tot dusver gevolgde strategie (slapen verbeteren via bewerken van angsten, spanningen en inadequate cognities alsmede via rechtstreekse instructies inzake slaapgedragingen),
2 nog meer accent aan het slapen geven, of:
3 alleen nu op het slapen koeren?

Zoals in de hoofdstukken 7 en 9 is aangegeven, wordt besloten tot strategie *2*. Daaruit zou bij M4 een verdere verbetering van het slapen moeten volgen, en op Angst, Depressie en Insufficiëntie van Denken & Handelen mogelijk weer verbetering of gelijkblijven aan M3.

Bij M4 blijkt het slapen verder verbeterd en Angst, Depressie en Insufficiëntie van Denken & Handelen gelijk gebleven aan M3. Schaal 10 (Totaal) was bij M2 significant verbeterd ten opzichte van M1 en dat is bij M3 en M4 zo gebleven.
Problemen met slapen waren de hoofdklacht. Waren angsten en spanningen even belangrijk geweest als slapen, dan waren er bij M3 nog de strategieën geweest om de behandeling inzake Angst, Depressie en Insufficiëntie van Denken & Handelen nu op een andere manier dan daarvóór aan te vatten en dat gelijktijdig te doen met aandacht geven aan slapen, ofwel eerst een keus te maken aan accent geven aan angst etc. dan wel vooreerst aan slapen.
Beziet men grafiek 3 dan blijkt dat de behandeling inzake verbetering van slapen zeer effectief was. Was het hoofddoel evenwel een verbetering inzake Angst, Insufficiëntie van Denken & Handelen, en Depressie (schalen 2, 3 en 4) geweest, dan zou het voor de hand liggen na M3 de functieanalyse te herzien of kritisch te bezien en op basis daarvan de therapie aan te passen.

Hoe kunnen de wisselende bevindingen (bij M2 nog geen vooruitgang inzake slapen, wel op de andere beoogde schalen; bij M3 en M4 wel vooruitgang inzake slapen, maar minder op de andere beoogde schalen) in één waarderingscijfer worden weergeveven? Dat kan op basis van de percentages gerealiseerde verwachting uit de tabellen 15, 16 en 17. Dit gebeurt in grafiek 4.

Grafiek 4



M2, M3 en M4: respectievelijk meting 2, 3 en 4 ten opzichte van M1, waar de verwachtingen werden gesteld.

Schaal 10 (SCL90-totaalscore) is hier als één van de schalen meegenomen. Dat hoeft niet. Er valt iets voor te zeggen die weg te laten, aangezien alle schalen zelf al aan bod komen. Zou voor schaal 10 een verwachting geformuleerd zijn op grond van de behandeling (hier het geval), dan is meenemen van deze schaal wenselijk. Zou men schaal 10 weglaten, dan wordt de

vooruitgang 6/8 × 100% = 75% bij tabel 9, 5/8 × 100% = 62,5% → 63% bij tabel 10 en weer 6/8 × 100% = 75% bij tabel 11. De vooruitgang (in percentages) bij het niet meetellen van schaal *10* is dus minder. Mét schaal 10 gemiddeld (M2, M3 en M4) 84%, zonder: 70,83% → 71%.

Tot dusver is steeds uitgegaan van dezelfde verwachting. Dat hoeft natuurlijk niet. Het zou realistisch geweest zijn te verwachten dat de slaapproblemen zouden verbeteren, maar niet al op de eerste verschilmeting (dus bij M2 ten opzichte van M1). Verbetering van slaapproblemen (schaal 8) zou dan verwacht worden bij M3 (ten opzichte van M1), maar dus nog niet bij M2. Aldus kan men ook per verschilmeting verwachtingen stellen en toetsen.
Zou bijvoorbeeld bij de eerste verschilmeting (dus M2 ten opzichte van M1) de verwachting gesteld zijn dat de slaapproblemen niet-significant beter zouden zijn (wel trend), dan zou tabel 16 er uitgezien hebben als tabel 19. In plaats van een vooruitgang van 88% (zoals tabel 10 laat zien), resulteert dan een vooruitgang van 107%.

Tabel 19: Verschil M1 en M2 bij andere (aangepaste, lagere) verwachtingen

Schaal	Verwacht	Gebleken	Waarde	
			Conform verwachting	Gebleken
1	GV	GV	0	0
2	S	S	2	2
3	S	S	2	2
4	S	S	2	2
5	GV	GV	0	0
6	GV	GV	0	0
7	GV	GV	0	0
8	NS	GV	0,5	0
9	GV	GV	0	0
10	NS	S!	0,5	1,5
Totaal			7	7,5
Gehaald (7,5/7 × 100% =)			107,14% → 107%	

(Dit alles zou nog sterker geweest zijn als men voor de schalen 2, 3 en 4 bij M2 ook NS verwacht had. Het totaal volgens de verwachting was dan niet 7, maar 2,5 geweest (in plaats van 3 × 2 voor de schalen 2, 3 en 4, nu 3 × 0,5). De kolom 'Gebleken' blijft hetzelfde, de vooruitgang is dan dus 7,5/2,5 × 100% = 300%!)

Bij de derde verschilmeting (M3 ten opzichte van M1), wordt dan inzake schaal 8 wel een significante vooruitgang verwacht. Tabel 17 blijft daarmee zoals die is. En ook tabel 18 blijft zo, want ook daar werd een significante vooruitgang van de slaapproblemen verwacht.

Als men tussentijds de verwachting inzake vooruitgang vergroot of verandert (dus na 3 maanden de verwachting van wel vooruitgang maar nog niet-significant en na 6 maanden verwachting van een significante vooruitgang op schaal 8), dan zijn de percentages niet meer onderling vergelijkbaar. Als immers de verwachting hoger komt te liggen, zal eenzelfde resultaat (scoring) bij een hogere vooraf-verwachting tot een lager percentage leiden. Wil men de percentages onderling vergelijkbaar houden, dan moet steeds van dezelfde verwachting uitgegaan worden. Er is wel veel voor te zeggen één verwachtingspatroon op te stellen en het uiteindelijke doel vanaf het begin als verwachting te formuleren. De verschillende meetmomenten laten dan zien 'hoe men naar dat doel toegroeit'.

13 Welke verwachtingen, de keuze van verwachtingen

Door ‘op safe’ te spelen kan men het percentage proberen op te schroeven. Zou bijvoorbeeld in de voorbeeldcasus alleen verbetering verwacht worden op slaapproblemen (schaal 8) en niet op andere, dan zou er dus inzake schaal 8 een verwachte significante vooruitgang zijn (waarde 2) en bij de andere schalen geen vooruitgang (waarden 0). Tabel 16 had er dan als volgt uitgezien (zie tabel 20).

Tabel 20: Verschil M1 en M2 (bij alleen verwachte vooruitgang op schaal 8)

Schaal	**Verwacht**	**Gebleken**	**Waarde**	
			Conform verwachting	**Gebleken**
1	GV	GV	0	0
2	GV	S	0	0,5
3	GV	S	0	0,5
4	GV	S	0	0,5
5	GV	GV	0	0
6	GV	GV	0	0
7	GV	GV	0	0
8	S	GV	2	0
9	GV	GV	0	0
10	GV	S!	0	1
Totaal			2	2,5
Gehaald (2,5/2 × 100% =)			125%	

Er is nu een vooruitgang van 125% terwijl de kernvariabele, schaal 8, op deze eerste verschilmeting (M1 naar M2) zelfs nog geen trend of significante vooruitgang heeft laten zien! De verschillen van M1 met M3 en M1 met M4 (dan is er inmiddels een significante vooruitgang op schaal 8) zouden op 4,75/2 × 100% = 237,5% (→ 238%) uitkomen (voor M1-M3) en op 5/2 × 100% = 250% (voor M1-M4)!
Door dus zo laag mogelijke verwachtingen te stellen, wordt alleen van díé verwachtingen uitgegaan. Waar dan niet-significante of significante vooruitgang optreedt (die eigenlijk ook verwacht zouden mogen worden, maar niet als verwachting is geformuleerd) resulteren dan ‘meevallers’ die meetellen bij de behaalde vooruitgang.
Dit is te voorkómen door intervisie of supervisie. De therapeut bespreekt de analyse en getaxeerde probleemsamenhang van de problemen van de cliënt/patiënt op een intervisie. Daar worden de verwachtingen dan vastgelegd. Bij het formuleren van de verwachtingen, wordt uitgegaan van de analyse op basis van het theoretisch model dat aan de therapie ten grondslag ligt. Men kiest een psychodiagnostisch instrument dat voldoende variabelen (schalen) biedt om de analyse deels in te verwoorden. De problematie en eventuele voor- én achteruitgang wordt dus mede geformuleerd in termen van de gebruikte schalen. Alleen vooruitgang bij slapen in deze casus en geen verwacht verschil op tenminste twee andere schalen, is onwaarschijnlijk gezien hoe de analyse eruit zal zien. En dus ook niet goed te verdedigen in inter- of supervisie.

14 Theoretische en statistisch-methodologische beschouwingen

- Er is sprake van *herhaalde metingen*. Men kan de herhaalde metingen overigens beperken tot één herhalingsmeting, zelfs al zijn het er meer. In het voorbeeld het vergelijken van (toetsen van verschillen tussen) M1 met M2, dan M1 met M3 en vervolgens M1 met M4. Normaal gesproken gebruikt men een gepaarde t-toets voor verschillende, afzonderlijke subjecten en voor elk subject twee (herhaalde) waarnemingen. Nu zijn alle waarnemingen van hetzelfde subject.

- Het is mogelijk dat de totaalwaarde van de gebleken uitslagen negatief is. Er treedt bijvoorbeeld geen vooruitgang op, maar wel achteruitgang. In dat geval resulteert een negatieve breuk en idem percentage: gebleken vooruitgang gedeeld door verwachte maal 100. Een negatief percentage betekent derhalve achteruitgang.

- Het is goed zich te realiseren dat het in wezen gaat om het waarderen van *vooruitgang* ten aanzien van geformuleerde verwachtingen, het realiseren van verwachtingen. Die verwachtingen worden geconcretiseerd in de schalen (variabelen) waarop wel en niet vooruitgang verwacht wordt. Het vooruitgangspercentage (in termen van percentage van de verwachting) wordt daarmee sterk beïnvloed door de verwachting, maar is er niet geheel identiek mee. Ook vooruitgang op niet-verwachte schalen telt mee bij de totale vooruitgang, zij het minder zwaar. Het percentage vooruitgang (in termen van de verwachting) loopt wel parallel met de mate waarin vooruitgang wél optreedt op de verwachte schalen en níét op de schalen waarop het niet verwacht werd, maar is niet identiek aan een onderscheid tussen. Het is geen 'discriminatiecoëfficiënt' tussen de resultaten van schalen waarop wel en niet vooruitgang verwacht werd. Als het percentage vooruitgang laag is, kan dat liggen aan het niet blijken van vooruitgang waar die wel verwacht werd, maar ook aan het juist optreden van vooruitgang waar het níét verwacht werd vice versa.

- Het gaat bij vergelijken van percentages uitsluitend om *percentages binnen dezelfde therapie*. Als men steeds van dezelfde verwachtingen uitgaat (ten opzichte van M1 of M0) zijn de percentages onderling vergelijkbaar, voor die ene cliënt/patiënt. Percentages tussen verschillende cliënten mogen niet vergeleken worden tenzij dezelfde schalen worden gebruikt en dezelfde verwachtingen worden geformuleerd, wat zelden het geval zal zijn.

- Theoretisch is er eigenlijk sprake van multivariate statistiek aangezien elke schaal (variabele) van een meetinstrument (als bijvoorbeeld de SCL90), gezien kan worden als een afzonderlijke variabele of dimensie. Het gaat dan om de variabelen waarop men verschil (vooruitgang) verwacht en eventueel, beargumenteerd, variabelen waarop (in relatie met de schalen van verwachte vooruitgang) men geen vooruitgang verwacht. Bijvoorbeeld men verwacht vooruitgang op de tendens tot agressief reageren na een agressieregulatietraining. Maar agressief reageren bij die specifieke persoon hypothetiseert men mede te ontstaan op basis van angst. Men verwacht daarom parallel aan de afname van de tendens tot agressief reageren, in het begin gelijk blijven of zelfs enige toename van de angstscore. Angst en agressie zijn daarmee inhoudelijk gerelateerd en hoewel op de ene schaal (tendens tot agressie) vooruitgang verwacht wordt, is er een duidelijke relatie met de verwachting van gelijk blijven of zelfs enige toename op de andere (angst)

- De genoemde percentages vooruitgang zijn eigenlijk een andere manier om de betekenis van vooruitgang aan te geven dan (de statische) percentages verklaarde variantie. Werken

met percentages verklaarde variantie (als indicatie van de betekenis van een gevonden significant verschil) ligt hier niet zo voor de hand. Percentages verklaarde variantie berekenen, hoort alleen te geschieden nadat eerst significantie gebleken is. Het bereken van de percentages geschiedt juist op basis van de significanties of het ontbreken. In de waardering van eventuele vooruitgang speelt in het nu gehanteerde systeem is de significantie zelf het belangrijkste criterium.

- De vooruitgang zou ook uitgedrukt kunnen worden als 'effectmaat' (effect size ES), Cohens δ (Cohen, 1988; Veerman, 2006; zie het eind van § 10.2). Dat ligt hier niet zo voor de hand. Ten eerste is het woord effectmaat feitelijk onjuist (al wordt de term in de literatuur wel ook gebruikt bij alleen vóór- en nametingen), want het gaat alleen om vooruitgang. Of de vooruitgang effect kan worden genoemd, moet in latere afwegingen aan de orde komen. Ten tweede vraagt een maat als Cohens δ een toelichting op de betekenis van de δ-waarden (tot 0.20 weinig betekenis zelfs bij significantie, 0.20-0.49 enige kleine betekenis, 0.50 tot 0.79 middelgrote betekenis, boven 0.80 grote betekenis, δ kan overigens boven de 1.0 komen). Ten derde ligt het eerder voor de hand om per schaal een δ te berekenen dan één totaalwaarde en eventueel een gemiddelde δ-waarde aan te houden al is het dan de vraag waar de δ -waarde dan precies voor staat.
  Cohens δ is overigens erg gerelateerd aan de t-waarde bij een niet-gepaarde t-toets. Een grote t-waarde impliceert een behoorlijke δ.

- Als minimaal wenselijk percentage vooruitgang kan men tenminste 50% aanhouden indien er op de belangrijkste variabelen/schalen tenminste één significante vooruitgang is. Zijn er 2 metingen, vóór en na de behandeling (M1 en M2), dan moet tenminste één van de kernvariabelen een significante vooruitgang laten zien. Zijn er meer metingen (M1, M2, M3 etc.) dan moet dit bij tenminste één verschil tussen twee metingen het geval zijn en bij de andere metingen moet er tenminste ook minimaal één trend (niet-significante vooruitgang, $\alpha = 0.10$) zijn. Uiteraard kan het vóórkomen dat men tenminste een trend of significante vooruitgang wenst op de belangrijkste kernvariabele(n).

- Als er weinig variabelen/schalen zijn, kunnen de percentages vooruitgang snel relatief hoog uitvallen als er vooruitgang is. Er zijn bijvoorbeeld twee schalen met één verwachting GV (geen verschil) en één S (significant verschil). Op beide blijkt een significante vooruitgang, dan levert dat een waarde op van $(1+2)/2 \times 100\% = 150\%$. Indien er meer schalen zijn, bijvoorbeeld 5, zal er meestal meer kans zijn dat waar geen verschillen verwacht werden, ze ook niet optreden. Dat drukt het percentage enigszins. Er valt daarom iets voor te zeggen ernaar te streven dat op tenminste 5 variabelen/schalen gemeten wordt.

- Het kan zijn dat schalen parallellie vertonen (gecorreleerd zijn). Dat kan vooraf bekend zijn of onbekend. In geval van parallellie rijst de vraag of ervoor moet worden gecorrigeerd. De uitkomst zou er immers door geflatteerd kunnen worden. Dit wordt inhoofdstuk 17 uitgewerkt.

15 Baseline-fase

Er kan met een *baseline-fase* worden gewerkt. In die fase verwacht men geen vooruitgang, dus in principe is de vooruitgang 0% (of zelfs achteruitgang) tenzij er in die baseline-fase een invloed werkzaam geacht kan worden waarvan een duidelijk effect verwacht werd op één of meer schalen. Als de meting vóór de baseline-fase meting 0 is (M0) dan is er dus de verwach-

ting van geen verschil tussen M0 en M1 (tenzij specifieke overwegingen). En wel verschil op sommige variabelen/schalen van M1 naar M2 en latere (en dus in feite ook van M0 naar M2 en latere).

Een baselinefase is designtechnisch elegant. Men heeft een fase waarin er nog geen therapie is. Maar een baselinefase vraagt ook speciale overwegingen.
De eerste speciale overweging betreft het tijdstip van de formulering van de verwachtingen. De verwachtingen worden opgesteld nadat er intakegesprekken geweest zijn en een behandelplan is opgesteld. Dat is in de baselinefase niet zonder meer het geval. Is er nog geen behandelplan dan kunnen ook nog geen verwachtingen geformuleerd worden over de schalen waarop vooruitgang wordt verwacht. In dat geval is het designtechnisch het meest correct dat de behandelaar nog geen kennis heeft van de resultaten van de nulmeting (M0). Rond of snel na M1 formuleert de behandelaar de verwachtingen en die gelden ook voor de verschilmeting van M0 naar M1. Op deze manier blijft alle vooruitgang en blijven alle verschilmetingen vergelijkbaar omdat steeds dezelfde waarden worden aangehouden.
De tweede speciale overweging is van wiskundige aard. Dit geldt vooral als men tussen M0 en M1 verschilverwachtingen van 0 definieert. Men toetst dus alleen tussen M0 en M1 met verwachting: op geen enkele schaal wordt verschil verwacht. Dat betekent dat de verwachte totaal-vooruitgang 0 is. Aangezien de verwachte vooruitgang in de noemer komt (zie de tabellen 15 t/m 19) en de gebleken vooruitgang in de teller, ontstaat er dan een onwerkbare breuk: delen door 0 is niet mogelijk of levert als uitkomst: oneindig. Alleen als ook geen enkele vooruitgang blijkt (de gebleken vooruitgang is dus ook 0).[9]
Men zou in geval van 0 als verwachte vooruitgang, kunnen werken met een kunstmatige waarde. Bijvoorbeeld 0,5 of 1. Vanuit het idee dat er altijd wel enige vooruitgang zal zijn, al was het maar op basis van toeval. Maar dat is erg arbitrair. Ook achteruitgang kan soms goed denkbaar zijn.
Beter is het daarom de verwachtingen van M1 naar M2 en latere ook al bij M0 te stellen, zowel tussen M1 en M2 en eventuele latere, als tussen M0 en M2 en eventuele latere. Men vergelijkt dus eerst M0 met M2 en vindt een vooruitgangspercentage. Men vergelijkt vervolgens M1 met M2, wat ook een vooruitgangspercentage oplevert. Vervolgens worden M0-M2 en M1-M2 vergeleken. Men vergelijkt dus eigenlijk M0 met M1 door hun vergelijkingen met M2 te vergelijken.

16 Analyses en te meten variabelen

De keus van de meetinstrumenten, dat wil zeggen de variabelen die men wil meten, dient (naast de psychometrische kwaliteit van het instrument) mede gemaakt te worden op basis van de analyse en het behandelplan die de therapeut maakt (of op basis van de verwachtingen volgens het therapeutisch protocol; in gedragstherapeutische termen: op basis van de holistische theorie, functieanalyse, betekenisanalyse, cognitiesanalyse). Dit kwam al aan de orde aan het eind van § 3. Als men weet dat agressie bij een persoon mede bepaald wordt door angst (bij angst sneller kans op impulsief-agressief reageren) dan ligt het voor de hand zowel angstmaten als agressiematen niet alleen bij de metingen te betrekken, maar ook de relatie tussen beide te formuleren. Bijvoorbeeld: aanvankelijk verwacht men na een agressieregulatietraining een afname van de tendens tot (impulsief) agressief reageren, alsmede een gelijk blijven of

---

[9] Alleen als ook geen enkele vooruitgang blijkt (de gebleken vooruitgang is dus ook 0) kan er gedeeld worden: 0/0 = 1, wat op 100% neerkomt: de vooruitgang is precies (100%) conform de verwachting, namelijk 0 Dit is natuurlijk vooral kunstmatig. Overigens valt op goede gronden te verdedigen ook 0/0 als een niet werkbare breuk op te vatten.

zelfs enige stijging van de angst. Het laatste omdat het niet uiten van de angst in de vorm van agressie tot stijging van de angst kan leiden. Pas in later instantie verwacht men óók afname van de angst.
De formulering van bijvoorbeeld de gedragstherapeutische holistische theorie, functie-, betekenis en cognitiesanalyse wordt qua inhoud deels bepaald door de concepten die men met de meetinstrumenten gaat meten. Anderzijds worden deze concepten (variabelen) en daarmee de meetinstrumenten deels gekozen op grond van de inhoud de holistische theorie, functie-, betekenis en cognitieanalyse. Het is dus een wederzijdse beïnvloeding.
Men kan de behandeltheorie schematiseren door aan te geven hoe wordt aangekeken tegen werkelijkheid van een cliënt en op welke aspecten van diens werkelijkheid de behandeling aangrijpt. En hoe vervolgens de relatie daarvan is met de voor evaluatie te meten variabelen.

Een mooi voorbeeld van een behandeltheorie die gegoten wordt in zulke schema's geven Van Yperen, Bijl en Veerman (2006). Ze maken een onderscheid tussen kernprobleem, typische hulpvraag, factoren die wel en niet beïnvloedbaar geacht worden met de interventie, tussen- en einddoelen. Van elk kan dan aangegeven worden welke aspecten ervan worden gemeten.

17 Parallellie tussen variabelen/schalen

Een speciaal punt van overweging is parallellie tussen (intercorrelaties van) variabelen/schalen. Als uit het betrouwbaarheids- en valideringsonderzoek van een instrument bekend is dat twee schalen onderling behoorlijk correleren, zal een significante vooruitgang op één schaal een behoorlijke kans geven op een trend of significante vooruitgang op de ermee correlerende schaal. Als de vooruitgang dan wordt uitgedrukt in percentages zoals eerder in tabellen aangegeven, wordt het percentage enigszins geflatteerd. Men krijgt dan op twee schalen vooruitgang terwijl er wellicht maar werkelijk vooruitgang is op één onderliggende dimensie voor beide schalen.
Dit lijkt een redelijke aanname. Maar het is niet zonder meer waar. Als er een duidelijke verwachting is dat er parallelle variabelen/schalen zijn, dan zullen daarvoor ook parallelle verwachtingen geformuleerd worden, waartegen de gebleken vooruitgang wordt afgezet. Er verandert dan niets.
Neem een (individuele) training in agressieregulatie. De agressie wordt verwacht medebepaald te worden door angst. De verwachting is dat de agressie significant zal afnemen (bij de tweede meting in vergelijking met de eerste), maar angst nog niet. Stel men meet de tendens tot directe en indirecte agressie met de BDHI (Buss-Durkee Hostility Inventory, in latere versie de Agressie Vragenlijst – AVL; Dehghani & Lange, 1993; Lange & Hoogedoorn, 1996; Lange, Pahlich, Sarucco & Smits, 1993; Meesters, Muris, Bosma, Schouten & Beuving, 1996) en beide eveneens met de Novaco agressieschalen (Novaco, 1994; Novako & Renwick, 1998). De BDHI- en Novaco-schalen worden verwacht hoog te correleren.
Stel dat vervolgens de tendens tot directe agressie significant is afgenomen, die tot indirecte agressie niet-significant is afgenomen en de angst gelijk is gebleven. Dat resulteert in de volgende tabel.

Tabel 21: Verschil M1 en M2 na agressieregulatietraining

Schaal	Verwacht	Gebleken	Waarde	
			Conform verwachting	Gebleken
BDHI: Indirecte agressie	S	NS	2	1
BDHI: Directe agressie	S	S	2	2
Novaco: Indirecte agressie	S	NS	2	1
Novaco: Directe agressie	S	S	2	2
Angst	GV	GV	0	0
Totaal			8	6
Gehaald (6/8 × 100% =)			75%	

Er blijkt een vooruitgang van 75% van de verwachting. Nu laat men de Novaco-schalen weg. Het percentage wordt dan: (1+2)/(2+2) × 100% = ¾ × 100% = 75%. Dus exact hetzelfde.

Het percentage vooruitgang kan wel beïnvloed worden door een *onbekende parallellie* tussen schalen. Schalen kunnen *empirisch gecorreleerd* zijn (tevoren niet verwacht, maar bijvoorbeeld in het valideringsonderzoek gebleken) of *inhoudelijk gecorreleerd*. Van het laatste werd hiervóór een voorbeeld gegeven inzake agressie. Een ander voorbeeld zijn de schalen Agorafobie en Angst van de SCL90 (in de tabellen de schalen 1 en 2). Er mag theoretisch enige parallellie verwacht worden tussen beide.
Als twee schalen, waarvan niet bekend is of ze intercorreleren (dus waarvan empirische correlatie niet bekend is), elk een behoorlijke vooruitgang laten zien, kan dat dus deels een gevolg zijn van hun onderlinge correlatie of van alleen onafhankelijke verbeteringen op elk afzonderlijk.
Onbekende parallellie zal vooral naar voren komen bij een vooruitgang op schalen waarop geen vooruitgang (verschil) verwacht werd. Dat is ook de parallellie die voor behandelevaluatie vooral van betekenis is. Onbekende parallellie van schalen waarop de behandeling zich niet richt (en waarop men dus geen verschil verwacht), is doorgaans geen onderwerp van betekenis.

Stel twee schalen correleren onderling. Maar dat is niet bekend. Ze worden vooralsnog beschouwd als onafhankelijk. Men verwacht bijvoorbeeld op één schaal (significante) vooruitgang (conform verwachting: waarde 2) die ook blijkt (gebleken: waarde 2, vooruitgang dus 2/2 × 100% = 100%). Op de andere schaal, met onbekende parallellie verwacht men geen vooruitgang (conform verwachting: waarde 0), maar er blijkt wel een significante vooruitgang (gebleken: waarde 0,5), dan komt de vooruitgang neer op (2 + 0,5)/2 × 100% = 125%. De 25% extra kan een gevolg zijn van de therapie, van de onderlinge correlatie van de beide schalen of van beide.
Het lijkt dan voor de hand te liggen voor zulke parallellie te corrigeren.
Aan alle correctiemogelijkheden zijn evenwel behoorlijke bezwaren verbonden. Bovendien is de overschatting niet een groot probleem. Immers de overschatting zal bij latere metingen (afgezet tegen dezelfde verwachtingen) in dezelfde mate optreden. De percentages vooruitgang bij alle meetmomenten ten opzichte van meting 1 blijven daarmee onderling vergelijkbaar al zijn ze wellicht geflatteerd.
Men wil percentages over verschillende metingen vergelijken. Dan maakt al of niet corrigeren eigenlijk niet zoveel uit. Dat zou uitmaken als men percentages van verschillende cliënten of patiënten of van verschillende therapieën zou vergelijken, maar zulke vergelijkingen kunnen

niet gemaakt worden. Alle vergelijkingen van percentages zijn individueel, gelden binnen de ene, individuele behandeling (of binnen de ene groep als men met groepsdata werkt).

Een belangrijk bezwaar tegen correctie(s) inzake parallellie is dat men de werkelijke correlaties meestal niet of maar beperkt kent voor de omstandigheden die gelden (behandelevaluatie). De werkelijke correlaties zijn alleen bekend als ze bij de constructie van een instrument en ijking op een steekproef (validering) zijn nagegaan. Soms worden schalen daarbij ook nog gecorreleerd met schalen van andere instrumenten. Men heeft daarmee echter nog niet de correlaties met de schalen die men zelf in een N=1-studie gebruikt. De intercorrelaties van de SCL90-schalen zijn uit het onderzoek van de schaalconstructie bekend, maar correlaties met andere veel gebruikte vragenlijsten voor mensen met een autismespectrumstoornis bijvoorbeeld niet. Intercorrelaties die gelden voor de hele populatie, bijvoorbeeld de correlaties tussen SCL90-schalen voor de hele bevolking, zijn niet zonder meer hetzelfde zijn voor deelpopulaties (alle mensen met autismespectrumstoornissen of persoonlijkheidsstoornissen).
In de literatuur wordt soms aanbevolen uit te gaan van een onderlinge correlatie van r = .30 (en daar dan voor te corrigeren; de methode van Goal Attainment Scaling - GAS - doet dit; Bartels, 1989; Melief, Hoekstra, Langerak, Sijben & Wevers, 1979; Sijben, Zwaan & Zwart, 1979; Veerman, 2006), maar dit is in wezen een erg arbitraire procedure. De correlaties tussen schalen van eenzelfde of van meerdere instrumenten zullen in werkelijkheid onderling erg variëren. Eenzelfde (arbitraire) waarde aanhouden is dan discutabel. Bovendien maakt het in feite eigenlijk geen verschil of men wel of niet corrigeert als op alle variabelen/schalen dezelfde correctie wordt toegepast.
Een ander belangrijk bezwaar tegen correctie is dat de betekenis van een empirische correlatie, of men die kent of niet, erg kan verschillen in verschillende behandelomstandigheden. Hierna een voorbeeld aan de hand van de relatie tussen agorafobie en angst (schaal 1 en 2 van de SCL90) bij drie verschillende cognitieve gedragstherapieën.
De eerste behandeling is tegen angsten. Het ligt voor de hand ook op de schaal voor agorafobie vooruitgang te verwacht: angst en agorafobie zijn in hun algemeenheid in zekere mate gecorreleerd.
De tweede behandeling is een cognitieve therapie uit tegen zelfdevaluatie en angst. Het ligt voor de hand dat dan schaal 2 (Angst) en 4 (Insufficiëntie van Denken & Handelen) vooruitgang zullen laten zien. Niet omdat ze (wellicht) onderling empirisch gerelateerd zijn, maar omdat ze in de behandeling als oorzakelijk gerelateerd beschouwd worden (vermindering van zelfdevaluatie zal leiden tot minder angst en vermijding). Eventueel wordt ook op depressie (schaal 3) enige vooruitgang verwacht. Maar in dit geval niet op schaal 1 (Agorafobie). Als er nu toch op schaal 1 vooruitgang is, dan kan deze vooruitgang komen door de altijd aanwezige parallellie met angst, maar ook door toepassing van de therapie op agorafobische onderwerpen. Dat laatste kan men nagaan. Anders gezegd: parallellie met agorafobie werd niet verwacht, en indien toch ook vooruitgang op agorafobie blijkt, is dat een gegeven dat van belang is voor de voortgang van de behandeling (bij de angsten speelden wellicht nog niet onderkende agorafobische aspecten mee).
De derde behandeling is specifiek gericht tegen agorafobie. Het programma begint snel met ‘exposure’ (getrapt en gesteund blootstellen aan angstoproepende situaties). De verwachting is vooruitgang op schaal 1 (Agorafobie), maar nog niet op 2 (Angst). Immers de exposure zal vermoedelijk, zeker in het begin, tot angsttoename leiden. Als schaal 1 vooruitgang laat zien, maar schaal 2 niet, kan het zijn dat er inderdaad nog geen vooruitgang op 2 was.
Concluderend: in deze drie behandelvoorbeelden over de relatie tussen angst en agorafobie, wordt in elke behandeling een verschillend verband tussen angst en agorafobie verondersteld. En bij alle drie de behandelingen is de uitkomst (of er al dan niet gelijktijdig vooruitgang is op angst en agorafobie) van veel betekenis voor de functieanalyse en vervolg van de behande-

ling. Corrigeren op basis van een of meer correlaties die men (in hun algemeenheid) bijvoorbeeld uit de onderzoeksliteratuur kent, zonder rekening te houden met de genoemde behandelinhoudelijke aspecten, doet daarom afbreuk aan de waarde van de evaluatie.

Als men toch een idee wil hebben van geflatteerde vooruitgangspercentages op grond van niet bekende gecorreleerdheid van schalen, kan de volgende procedure worden aangehouden. Het gaat dan om niet-verwachte vooruitgang op grond van nog onbekende parallellie. Indien geen vooruitgang verwacht wordt, is de verwachting GV (geen verschil), waarde conform verwachting: 0. Indien dan toch vooruitgang optreedt, bijvoorbeeld significante vooruitgang, is de gebleken waarde 0,5 (zie tabel 15).
Een rekenvoorbeeld: er zijn 4 schalen. Er is voor zover bekend geen parallellie. En er is ook in werkelijkheid geen parallellie, al weten we dat niet. Op de eerste twee schalen wordt significante vooruitgang verwacht, op de twee laatste geen verschil. De verwachtingen zijn respectievelijk 2, 2, 0 en 0 (som: 4). Stel nu dat de volgende waarden blijken: 2, 3, 0 en 0 (som 5). Het vooruitgangspercentage is $5/4 \times 100\% = 125\%$. Dit werd gehaald op de verwachte schalen. De beide schalen met verwachting GV (geen verschil) gaven ook geen verschil te zien en kregen daarmee 0 en 0. Zie tabel 22.1. Gecorreleerdheid van schalen zal hier geen rol (van betekenis) spelen.

Tabel 22.1: Rekenvoorbeeld 1

Schaal	Verwacht	Gebleken	Waarde	
			Conform verwachting	Gebleken
A	S	S	2	2
B	S	S!	2	3
C	GV	GV	0	0
D	GV	GV	0	0
Totaal			4	5
Gehaald ($5/4 \times 100\% =$)			125%	

Stel dat er in werkelijkheid wél parallellie is (al weten we dat niet) tussen de schalen A en B enerzijds en C en D anderzijds. De verwachtingen zijn hetzelfde als in tabel 22.1. Maar wat blijkt is het volgende: 2, 3, 0,5 en 0,5 (som: 6), vooruitgangspercentage: $6/4 \times 100\% = 150\%$. Zie tabel 22.2 hierna. Nu gaven de beide schalen waarop geen vooruitgang verwacht werd, wel een vooruitgangstrend (niet-significante trend) te zien. De bijdrage van deze twee schalen aan het vooruitgangspercentage van 150% is 25%. Het zou dus kunnen dat van het percentage van 150% vooruitgang 25% (mede) het gevolg is van niet-onderkende parallellie. De werkelijke vooruitgang ligt daarmee tussen 125% en 150%.
We concluderen wel tot vooruitgang. De vooruitgang van 150% is meer dan 50% (het gestelde minimum) en tenminste één schaal (A en B) laat een verwachte significante vooruitgang zien. We concluderen derhalve tot vooruitgang.

Tabel 22.2: Rekenvoorbeeld 2

Schaal	Ver-wacht	Ge-bleken	Waarde	
			Conform verwachting	Gebleken
A	S	NS	2	2
B	S	S	2	3
C	GV	S	0	0,5
D	GV	S	0	0,5
Totaal			4	6
Gehaald (6/4 × 100% =)			150%	

Van deze vooruitgang van 150% kan eventueel 1/4* × 100% = 25% een gevolg zijn van gecorreleerdheid van schaal A of B met C of D.

Conclusie: de vooruitgang op schalen waar geen verschil verwacht werd, kan veroorzaakt zijn door onbekende gecorreleerdheid van deze schalen met de schalen waarop wel vooruitgang verwacht werd en bleek. De bijdrage (als percentage van de totale verwachte vooruitgang) van de vooruitgang op schalen waarop geen verschil verwacht werd, kan een indicatie geven van de maximale (onbekende) gecorreleerdheid. Bij schalen waarop geen vooruitgang verwacht wordt, maar deze toch optreedt, moet altijd een goede uitleg gegeven worden. Zonder deze moet rekening gehouden worden met gecorreleerdheid van schalen.
Indien meer metingen gedaan worden (M2, M3 etc.) kan men ook vergelijken of dan hetzelfde patroon optreedt. Bij werkelijke gecorreleerdheid zal ook dan bij verwachting GV van zo'n gecorreleerde schaal een vooruitgang (blijven) optreden.

VII Somwaardenmethode: speciaal geconstrueerde verdeling om objectief verschillen vast te stellen

18 Objectief verschillen vaststellen tussen schaalscores of beoordelingen door meer dan één beoordelaar

18.1 Meer beoordelaars per item of meer items per schaal

Tot dusver was er bij elke meting één observatie, elk item werd als een observatie beschouwd: de itemscore die door de onderzochte persoon werd aangegeven. In de praktijk van behandelevaluatie komt het ook veel voor dat men observatie- of beoordelingsinstrumenten over de persoon (patiënt of cliënt) laat invullen door meerdere beoordelaars, bijvoorbeeld sociotherapeuten of groepsleiders.
Puur statistisch redenerend is dat eigenlijk niet nodig als van een instrument normgegevens als spreiding en betrouwbaarheidscoëfficiënten bekend zijn. Men kan dan verschillen tussen metingen benaderen op de manieren als aangegeven in de hoofdstukken 4 en 5 alsmede § 10.2 t/m § 10.4: normgroepen en RCI. Maar daaraan waren nadelen verbonden van een lage statistische power, waarvoor werken met itemscores een alternatief bood.
Een andere manier is het werken met gemiddelde scores van alle beoordelaren. Men kan dan middelen en de gemiddeldes als waarnemingen beschouwen. Of men kan consensusoverleg tussen de beoordelaren arrangeren, zodat een consensusscore ontstaat.
Aan het middelen van beoordelingen kleven evenwel vooral inhoudelijke bezwaren. Als beoordelingen gemiddeld worden, wordt in feite uitgegaan van verschillen tussen beoordelaren die zijn toe te schrijven aan beoordelingsverschillen. Het is evenwel mogelijk dat verschillen terug te voeren zijn op hetzij andere ontmoetingskaders (bijvoorbeeld een verschil tussen een psychomotore therapeut en een sociotherapeut) of op kwalitatieve verschillen in gedrag van sociotherapeuten (elke sociotherapeut lokt door zijn eigen, persoonsspecifieke optreden ander gedrag uit van de patiënt en beoordeelt dat dus iets anders). Gaat men van zo'n visie uit dan doet het juist afbreuk aan de informatie die men verzamelt door de waarnemingen van alle beoordelaars te middelen, en is het wenselijk voor het specifieke van de verschillende beoordelaars oog te hebben.
Ook kan het zijn dat bij vóórmetingen beoordelaars sterker verschillen dan bij nametingen en follow-up metingen. Dat afgenomen beoordelaarverschil kan een indicatie zijn van gedrag dat de cliënt of patiënt geleerd heeft.
Vergelijkbaar met het nemen van de gemiddelden van de beoordelaars is een procedure waarbij men de meest afwijkende score laat vervallen en werkt met het gemiddelde van de beoordelaars die het dichtst bij elkaar liggen. Zijn scoremogelijkheden 1 t/m 5 en drie beoordelaars scoren 1, 2 en 4, dan wordt waarde 4 weggelaten en resulteert gemiddelde 1,5. Aangezien dit gebeurt zowel bij de vóór- als nameting, is het toegestaan.
Dit wordt evenwel een probleem als bijvoorbeeld de drie beoordelaars 1, 2 en 3 scoren. Welke 'meest afwijkende' score moet dan vervallen: 1 of 3? Al kan men er in zo'n geval voor kiezen steeds de hoogste of laagste waarde te laten vervallen. Overigens ook een discutabele oplossing.
Indien men verschillen tussen beoordelaars wil verkleinen door consensusbesprekingen, heeft men te maken met groepsprocessen als bijvoorbeeld de formele of informele status of visieverschillen tussen beoordelaars over een bepaalde cliënt of patiënt.
Vanuit de klassieke manier van dataverwerking valt er dan iets voor te zeggen om beoordelaars als variantiebron op te nemen, als een invloed op zichzelf. Men toetst dan over alle data of beoordelaars onderling consistent verschillen. Zo ja, dan kan dit beoordelingsverschil van

behandelaars in mindering gebracht worden op de statistische error waartegen effecten worden getoetst.
Wanneer bij de verschillende metingen verschillende beoordelaars betrokken zijn, bijvoorbeeld de beoordelaars A, B en C bij M1, B, C en D bij M2, C, D en E bij M3, dan heeft het berekenen van specifieke beoordelaarvariantie minder zin. Want men kan het niet aan dezelfde persoon (beoordelaar) koppelen.
Het is mogelijk, op de manier zoals in de hoofdstukken 6 en 9 t/m 12 uitgelegd, te werken met itemscores van een schaal als waarnemingen en daarbij per item de gemiddelden van de beoordelaars te nemen. Wil men de daaraan genoemde bezwaren vermijden, dan wordt hiervoor de volgende procedure gepresenteerd.

18.2 Somwaardenverdelingen als toetsingsgrootheid

In § 10.6.3 werd heel kort een derde manier beschreven (naast de non-parametrische McNemar-toets en de gepaarde t-toets over itemscores) om met itemscores te werken. Deze zal hier iets meer in detail worden toegelicht.[10]

Vergelijkbaar met het onafhankelijk van elkaar scoren van een gebeurtenis door bijvoorbeeld drie verschillende beoordelaars, zijn scores op 3 items van een vragenlijst. Bijvoorbeeld de SCL90-schaal 8, Slapen. Zie tabel 23.1. M1 t/m M4 zijn de vier metingen.

Tabel 23.1: Scores schaal 8, Slapen

Nr. item	Item	M1	M2	M3	M4
44	Moeilijk in slaap	5	5	4	3
64	Te vroeg wakker	4	4	3	2
66	Onrustige/gestoorde slaap	5	4	3	3

Gezocht: een methode om op objectieve gestandaardiseerde wijze verschillen tussen meetmomenten zijn vast te stellen.

De procedure is als volgt. Op item 64, 'Te vroeg wakker' scoort de cliënt bij M1 4. We gaan er vanuit dat hij voor 4 koos, maar dat dit eventueel ook wel 3 of 5 had kunnen zijn, echter geen 1 of 2. Zijn 'waarschijnlijke scorebereik' was daarmee (3,4,5).
Evenzeer had hij in plaats van de 5 bij M1 op item 44 ook 4 kunnen scoren of 6. Het is zo dat 6 niet als scoremogelijkheid bestaat, maar hij had 6 kunnen kiezen als die had bestaan ('erger dan 5'). Aldus levert de waarde 5 als waarschijnlijk scorebereik (4,5,6).
Samenvattend: bij M1 is de scoring {5,4,5} die bestaat uit de elementen (4,5,6), (3,4,5) en (4,5,6).

Vervolgens worden alle combinaties van de 3 waarschijnlijke scorebereiken nagegaan. Elke combinatie levert een somwaarde. Zie tabel 23.2.

[10] De methode is in essentie afkomstig van Marinus Spreen.

Tabel 23.2: Somwaarden van de waarschijnlijke scorebereiken, schaal 8, Slapen

SCL90-item			Som	SCL90-item			Som	SCL90-item			Som
44	**64**	**66**		**44**	**64**	**66**		**44**	**64**	**66**	
4	3	4	**11**	5	3	4	**12**	6	3	4	**13**
4	3	5	**12**	5	3	5	**13**	6	3	5	**14**
4	3	6	**13**	5	3	6	**14**	6	3	6	**15**
4	4	4	**12**	5	4	4	**13**	6	4	4	**14**
4	4	5	**13**	5	4	5	**14**	6	4	5	**15**
4	4	6	**14**	5	4	6	**15**	6	4	6	**16**
4	5	4	**13**	5	5	4	**14**	6	5	4	**15**
4	5	5	**14**	5	5	5	**15**	6	5	5	**16**
4	5	6	**15**	5	5	6	**16**	6	5	6	**17**

Er resulteert, zoals men kan zien: 1 x 11, 3 x 12, 6 x 13, 7 x 14, 6 x 15, 3 x 16 en 1 x 17. Dat zijn 27 somwaarden in 7 frequenties (de frequenties 1, 3, 6, 7, 6, 3 en 1).

Deze waarden kunnen in een grafiek gezet worden. Zie grafiek 5.1.a.

Grafiek 5.1.a: Verdeling van de somwaarden van {5,4,5}

11	X
12	X X X
13	X X X X X X
14	X X X X X X X
15	X X X X X X
16	X X X
17	X

Of op de volgende manier.

Grafiek 5.1.b: Verdeling van de somwaarden van {5,4,5}

Ditzelfde kan ook gedaan worden voor de scores bij M2. De scores zijn {5,4,4}, wat de volgende elementen oplevert: (4,5,6), (3,4,5) en (3,4,5). Er resulteert dan de volgende grafiek

Grafiek 5.2.a: Verdeling van de somwaarden van {5,4,4}

10	X
11	X X X
12	X X X X X X
13	X X X X X X X
14	X X X X X X
15	X X X
16	X

Of op de volgende manier.

Grafiek 5.2.b: Verdeling van de somwaarden van {5,4,4}

7							
6							
5							
4							
3							
2							
1							
	10	**11**	**12**	**13**	**14**	**15**	**16**

We zien dezelfde verdeling, evenwel één somwaarde verschoven. Vervolgens worden beide grafieken tegen elkaar afgezet (vergeleken). Zie grafiek 5.3.

Grafiek 5.3.a: Overlap in somwaarden M1 en M2 voor SCL90-schaal 8, Slapen

10	*X*						
11	**X**	*X*	*X*				
12	**X**	**X**	**X**	*X*	*X*	*X*	
13	**X**	**X**	**X**	**X**	**X**	**X**	*X*
14	**X**	**X**	**X**	**X**	**X**	**X**	X
15	**X**	**X**	**X**	X	X	X	
16	**X**	X	X				
17	X						

X: Waarde M1
X: Waarde M2
X: Waarde van M1 en M2 gemeenschappelijk

Of op de volgende manier. Rood is de verdeling van M1, geel die van M2, oranje de overlap.

Grafiek 5.3.b: Verdeling van de somwaarden van {5,4,5} en {5,4,4}

7								
6								
5								
4								
3								
2								
1								
	10	**11**	**12**	**13**	**14**	**15**	**16**	**17**

We zien dan dat van de 27 cellen van elke verdeling, er bij 20 overlap is. Dat is 20/27 = 74%. We kiezen als criterium dat beide verdelingen verschillen indien de overlap minder is dan 50%. Dat is hier niet het geval. We besluiten dat er geen verschil is.

Aldus kunnen we ook M1 met M3 vergelijken: {5,4,5} met {4,3,3}. We krijgen dan grafiek 5.4.

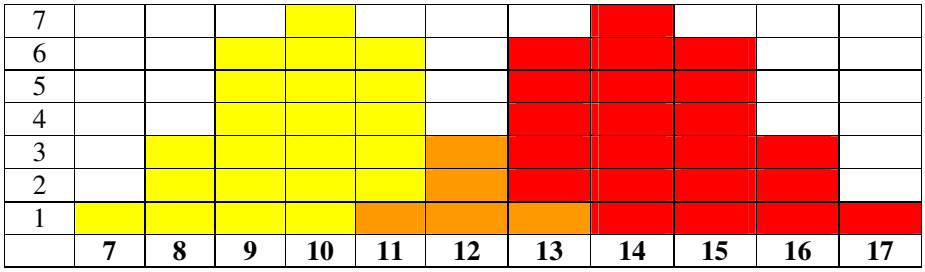
Grafiek 5.4.a: Overlap in somwaarden M1 en M3 voor SCL90-schaal 8, Slapen

7	*X*						
8	*X*	*X*	*X*				
9	*X*	*X*	*X*	*X*	*X*	*X*	
10	*X*	*X*	*X*	*X*	*X*	*X*	*X*
11	**X**	*X*	*X*	*X*	*X*	*X*	
12	**X**	**X**	**X**				
13	**X**	X	X	X	X	X	
14	X	X	X	X	X	X	X
15	X	X	X	X	X	X	
16	X	X	X				
17	X						

X: Waarde M1
X: Waarde M3
X: Waarde van M1 en M3 gemeenschappelijk

Of op de volgende manier. Rood is de verdeling van M1, geel nu die van M3, oranje de overlap.

Grafiek 5.4.b: Overlap in somwaarden M1 en M3 voor SCL90-schaal 8, Slapen



We zien hier een overlap van 5 cellen, ofwel $5/27^{e}$ = 18,5% → 19%. Dit is duidelijk minder dan 50% en ook hier besluiten we tot een structureel verschil van M1 en M3 op schaal 8, Slapen.

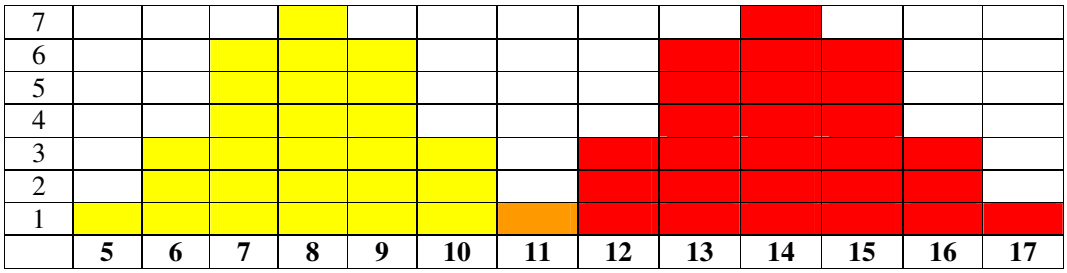
Vergelijken we tot slot M1 met M4: {5,4,5} met {3,2,3}, dan krijgen we de volgende grafiek 15.5. Er is dan slechts bij één cel overlap: 1/27 = 0,037. Er is dus een overlap van 3,4%. Dat is veel minder dan het criterium van $\leq$ 50%. We besluiten derhalve dat er tussen M1 en M4 een structureel verschil is.

Grafiek 5.5.a: Overlap in somwaarden M1 en M4 voor SCL90-schaal 8, Slapen

5	*X*						
6	*X*	*X*	*X*				
7	*X*	*X*	*X*	*X*	*X*	*X*	
8	*X*	*X*	*X*	*X*	*X*	*X*	*X*
9	*X*	*X*	*X*	*X*	*X*	*X*	
10	*X*	*X*	*X*				
11	**X**						
12	X	X	X				
13	X	X	X	X	X	X	
14	X	X	X	X	X	X	X
15	X	X	X	X	X	X	
16	X	X	X				
17	X						

X: Waarde M1
X: Waarde M4
X: Waarde van M1 en M4 gemeenschappelijk

Grafiek 5.5.b: Overlap in somwaarden M1 en M4 voor SCL90-schaal 8, Slapen



Wederom is de rode verdeling die van M1 en de gele nu die van M4, oranje is de overlap.

Aldus biedt het berekenen van somwaarden, die resulteren uit 'waarschijnlijke scorebereiken' een criterium om aan te geven wat als 'werkelijke', structurele vooruitgang kan worden beschouwd.

We hebben hier met 3 items van één schaal gewerkt. Maar het hadden, zoals in het begin van deze paragraaf gesteld, ook 3 beoordelaren A, B en C kunnen zijn die elk een oordeel uitspraken op bijvoorbeeld een 5-puntsschaal.

Tot slot de SCL90-angstschaal, schaal 2 in onze tabellen. M1 en M2 werden gegeven in tabel 1.1. Bij M1 is de totaalscore 21, bij M2: 17, bij M3 en M4: 18 en 18. De verschillen van respectievelijk M2, M3 en M4 met M1 zijn: respectievelijk 4, 3 en 3, de percentages overlap als volgt zijn: 45%, 57% en (uiteraard nogmaals) 57%.
In tabel 24 volgen de scores voor de 4 meetmomenten van de variabelen waarop verschil werd verwacht.

Tabel 24: Overlap in verdelingen van somwaarden SCL90-schaal 8, Slapen

SCL90		Metingen									
		M1	M2			M3			M4		
Schaal	N	Score	Score	V	Overlap	Score	V	Overlap	Score	V	Overlap
2	10	21	17	4	45% *	18	3	57%	18	3	57%
3	16	22	17	5	34% *	18	4	45% *	19	3	57%
4	9	22	16	6	23% *	17	5	31% *	17	5	31% *
8	3	14	13	1	74%	10	4	19% *	8	6	4% *
10	90	160	138	22	9% *	139	21	9% *	134	26	9% *

Toelichting
Onder M1-score, M2-score, M3-score en M4-score worden de schaalscores gegeven van de genoemde schalen (schaalnummers). Onder M2-V, M3-V en M4-V de verschillen van de schaalscores tussen respectievelijk M2, M3 en M4 met M1.
De percentages met * (minder dan 50% overlap) worden beschouwd als te wijzen op een structurele en objectieve verschillen. Dit komt overeen met wat in de toetsende statistiek als significant wordt beschouwd.

Vergelijkt men deze tabel met tabel 9.1, hierna nu weergegeven als tabel 25, de gepaarde t-toets over itemscores, *dan ontstaat een vergelijkbaar beeld.* Waar in tabel 9.1 een significante waarde ($\alpha = 0.05$) werd bereikt, is in tabel 24 de overlap minder dan 50%, wat een objectief en

structureel verschil inhoudt. Waar in tabel 24 een of trend ($\alpha = 0.10$) bleek, is de overlap net geen 50% of minder dan 60%.

Tabel 25 = tabel 9.1: Resultaten gepaarde t-toets (verschillen één- en tweezijdig getoetst)

	Gemiddelden				Significantie van verschillen (één- en tweezijdig)					
SCL90-schaal	**M1**	**M2**	**M3**	**M4**	**M1&M2**	**M1&M3**	**M1&M4**	**M2&M3**	**M2&M4**	**M3&M4**
1 Agorafobie	2,00	1,71	1,57	1,57	0,172	0,078	0,078	0,356	0,356	0,356
2 Angst#	2,10	1,70	1,80	1,80	0,019*	0,097°	0,097°	0,172	0,172	-
3 Depressie#	1,38	1,06	1,13	1,19	0,010*	0,052°	0,094°	0,167	0,082	0,290
4 Insuffiëntie van denken en handelen#	2,44	1,78	1,89	1,89	0,011*	0,090°	0,026*	0,297	0,174	0,500
5 Somatiek	2,17	2,00	2,17	1,92	0,337	1,000	0,275	0,339	0,586	0,191
6 Wantrouwen en interpersoonlijke sensitiviteit	1,33	1,28	1,22	1,22	0,668	0,430	0,331	0,331	0,579	1,000
7 Hostiliteit	1,00	1,17	1,17	1,17	0,363	0,363	0,363	-	-	-
8 Slaapproblemen#	4,67	4,33	3,33	2,67	0,212	0,029*	0,042*	0,042*	0,019*	0,092°
9 Overige	1,22	1,00	1,11	1,00	0,169	0,594	0,169	0,347	-	0,347
10 Totaal#	1,78	1,53	1,54	1,49	0,000*	0,002*	0,000*	0,798	0,374	0,278

Toelichting tabel 25

- M1, M2, M3 en M4: meting 1, 2, 3 en 4. Tussen elke meting lag circa 3 maanden.
- Met # aangegeven in de meest linkse kolom de schalen waarop vooruitgang werd verwacht.
- In de 4 kolommen onder gemiddelden de gemiddelde itemwaarde per schaal bij de 4 metingen.
- Als eerder aangegeven zijn gepaarde t-toetsen gebruikt.
- Gegeven zijn p-waarden. Significante verschillen op niveau $\alpha = 0.05$ zijn donker gearceerd en met * vermeld, significante verschillen op niveau $\alpha = 0.10$ zijn licht gearceerd en met ° vermeld (significante p-waarden tussen 0.05 en 0.10 worden als trend beschouwd).
- De p-waarde voor Slapen (schaal 8) bij M4 is hoger dan die bij M3 terwijl de gemiddelde itemscore is afgenomen (verbeterd van 3,33 naar 2,67). De reden is dat het verschil M1-M4 getoetst werd met een nonparametrische toets (Wilcoxon). De spreiding van de verschillen van M1 naar M4 is namelijk 0 (scores bij M1: 5, 4, 5; bij M4: 3, 2, 3), waardoor de toetsingsgrootheid t niet te berekenen is. Hetzelfde geldt voor het verschil van M2 met M3.
- Verschillen tussen M1 enerzijds en M2, M3 en M4 anderzijds zijn éénzijdig getoetst bij de schalen waarop vooruitgang verwacht werd: er werden verschillen verwacht, verbetering van M2 ten opzichte van M1, daarna voortgaande verbetering of gelijkblijven bij M3 en M4.Verschillen tussen M2, M3 en M4 onderling van deze schalen zijn ook éénzijdig getoetst bij de schalen waarop vooruitgang verwacht werd: áls er verschil zou zijn, werd dat geacht een verbetering te zijn of gelijk blijven. Bij schalen waarop geen vooruitgang verwacht werd, is tweezijdig getoetst.
- Over enkele verschilmetingen waren geen toetsen mogelijk, op grond van te weinig of te ongelijke spreiding in de scores. Deze verschillen zijn aangegeven met een liggend streepje (-).

Dit kan ook nog vergeleken worden met de resultaten van de McNemar-toets van tabel 6 in § 10.6.1.

Tabel 26 (= tabel 7): Verschillen tussen M1 en M2, M3 en M4 over de schalen waarop verbetering werd verwacht

SCL90-schaal	Verschillen ten opzichte van M1								
	M1 – M2			M1 – M3			M1 – M4		
	Vooruit	Achteruit	χ^2	Vooruit	Achteruit	χ^2	Vooruit	Achteruit	χ^2
2	4	0	4,0*	4	1	1,8	4	1	1,8
3	5	0	5,0*	5	1	4,0*	4	1	1,8
4	5	0	5.0*	5	2	1,3	4	0	4,0*
8	1	0	1,0#	3	0	3,0#	3	0	3,0#*
10	26	8	9,5*	29	9	10,5*	26	7	10,9*

Toelichting

* Significant op 0.05-niveau,

\# In plaats van McNemar-toets pure kansberekening omdat er slechts 3 items van de schaal zijn

De tabellen 24, 25 en 26 samenvattend vergelijkend, resulteren bij de drie manieren van toetsen de volgende resultaten. Zie tabel 27.

Tabel 27: McNemar-toets, t-toets en somwaardenmethode vergeleken

SCL90-schaal	Verschillen ten opzichte van M1								
	M1 – M2			M1 – M3			M1 – M4		
	McN	t-toets	Somw.	McN	t-toets	Somw.	McN	t-toets	Somw.
2	*	*	*		°	°		°	°
3	*	*	*	°	°	*		°	°
4	*	*	*		°	*	*	*	*
8				°	*	*	*	*	*
10	*	*	*	*	*	*	*	*	*

Toelichting tabel 27

- SCL90-schalen:
  2 = Angst, 3 = Depressie, 4 = Insufficiëntie van Denken & Handelen, 8 = Slapen, 10 = Totaal
  Dit zijn de schalen waarop vooruitgang verwacht werd op grond van het behandelplan.
- McN: McNemar-toets (bij schaal 8, bestaand uit 3 items, pure kansberekening), t-toets: gepaarde t-toets; somw.: somwaardenmethode
- M1: meting 1, M2: meting 2, M3: meting 3, M4: meting 4
- Alle verschillen éénzijdig getoetst
- * significant op 0.05-niveau (somwaardenoverlap < 50%)
- ° significant op 0.10-niveau (somwaardenoverlap < 60%)

VIII Algemeen kader, sommeren N=1-studies

19 Verwachtingshypotheses deponeren

Het verdient sterke aanbeveling dat de de analyse (de behandeltheorie) met collegae te bespreken en er een verslag van te maken. Dit verslag en de gestelde verwachtingen ergens worden 'gedeponeerd'. Het wordt door een collega medeondertekend en vervolgens bijvoorbeeld bij het afdelingssecretariaat of bij een commissie van enkele collegae gearchiveerd, liefst bij een centraal secretariaat van de instelling. Hiervan wordt een aparte administratie bijgehouden waartoe de behandelaar geen toegang heeft. Dit voorkomt dat men achteraf verwachtingen gaat bijstellen.
Aldus disciplineert men zichzelf en collegae in intervisie, waardoor men zichzelf ook dwingt goed na te denken over welke verwachtingen men zal formuleren en de rationales ervan.

20 Sommeren van N=1-studies

Als men gegevens heeft over verschillende behandelingen die zijn opgezet als N=1-studie, kan bezien worden of deze kunnen worden gesommeerd. Dit is een vrij complexe aangelegenheid die een eigen bespreking verdient. Maar het is erg belangrijk. In de hoofdstukken 1 en 2 werd al opgemerkt dat de American psychological Association (Task force, 1995; Van Yperen & Bijl, 2006) een cumulatie van tenminste 8 N=1-studies een volwaardig alternatief achtte voor een random clinical trial (RCT). En een RCT is de gouden standaard voor effectonderzoek.
Hierna enkele opmerkingen.
Wij zien N=1-studies, ook 8 of meer bij hetzelfde type probleem en cliënt/patiënt, niet als volkomen volwaardig alternatief voor een RCT, maar wel als een gedegen voorbereiding erop.
Vooruitgangspercentages als berekend werden in § 12.1, de tabellen 16, 17 en 18, kunnen gesommeerd worden. Al zijn ze niet onderling qua grootte vergelijkbaar. Bij één behandeling wordt bijvoorbeeld op twee SCL90-schalen vooruitgang verwacht, bij een andere behandeling op vier andere SCL90- of andere schalen.
Dit is vergelijkbaar met Goal Attainment Scaling waarbij doelrealisatie voor elke cliënt of patiënt kan verschillen bij dezelfde waarden (Bartels, 1989; Melief, Hoekstra, Langerak, Sijben & Wevers, 1979; Sijben, Zwaan & Zwart, 1979; Veerman, 2006). Wat men voor één persoon een grote vooruitgang noemt, kan voor een ander een kleine vooruitgang zijn.
Indien bij alle personen in de sommatiestudie dezelfde meet- en evaluatie-instrumenten gebruikt worden, kan men zonder problemen werken met de schaaltotalen of schaalgemiddeldes. In de voorbeeldcasus ging het slapen van M1 naar M4 vooruit van gemiddeld 4,67 naar 4,33, vervolgens 3,33 en uiteindelijk 2,67 (tabel 9.1). Als men over 10 personen zulke scores heeft, zijn er dus 10 waarnemingen per meting, waarover verschillen kunnen worden berekend en getoetst. Daarover kunnen toetsen worden uitgevoerd. Ook kan men de M1-waarden (schaaltotalen of -gemiddeldes) vergelijken met alle M2-, M3- en M4-waarden.
De vraag is vervolgens of men van elke persoon de afzonderlijke itemscores als waarnemingen kan opvoeren. Dit levert een forse toename van het aantal waarnemingen op en daarmee een behoorlijk grotere kans op significanties. Hier speelt de vraag naar onafhankelijkheid van die waarnemingen een rol, die ook al aan de orde kwam in § 11.3.
Deze manier van sommeren (van afzonderlijke itemscores over alle subjecten) is te verdedigen, maar correcter lijkt het de subjecten elk afzonderlijk als een 'conditie' op te voeren. Men heeft dus binnen conditie 1 (= subject 1) vier metingen met bij elke meting 90 waarnemingen met de SCL90. Of anders: men heeft 4 metingen en binnen elke meting 10 condities (= subjecten).

21 Referenties

Arrindell, W.A. & Ettema, J.H.M. (2003), *SCL-90 (Symptom Checklist 90). Handleiding bij een multidimensionele psychopathologie-indicator*. Lisse: Swets & Zeitlinger.

Arindell, W.A., Boosma, A., Ettema, J.H.M. & Stewart, R. (2004), Verdere steun voor het multi-dimensionale karakter van de SCL-90-R. *De Psycholoog,* april, 39 (4), 195-201.

Bartels, A.A.J. (1993). Gedragstherapie ten behoeve van extreem gedragsgestoorde geestelijk gehandicapten: een N=1-studie. *Gedragstherapie*, juni, 26 (2), 99-117.

Bartels, A.A.J. (1989), De evaluatiescore: een eenvoudige en flexibele variant van Goal Attainment Scaling. *Gedragstherapie,* sept., 22 (3), 191-204.

Beek, J. van (2007), *Cognitieve gedragstherapie via Interpay bij paniekklachten: Een studie op N=1-niveau*. Amsterdam: Universiteit van Amsterdam (UvA), vakgroepen Klinische Psychologie en Psychologische methodenleer, Faculteit Maatschappij en Gedragswetenschappen, Afdeling Psychologie.

Beurs, E. de & Barendregt, M. (2008), *Mogelijkheden voor therapie-effectonderzoek in de tbs-sector: komen tot een evidence base onder zorgprogramma's*. Den Haag: Ministerie van Justitie, Wetenschappelijk Onderzoeks- en Documentatiecentrum (WODC), Utrecht: Nederlands Instituut voor Forensische Psychiatrie en Psychologie.

Bishop, Y.M.M., Fienberg, S.E., Holland, P.W., Light, R.J. & Mosteller, F. (1995). *Discrete Multivariate Analysis. Theory and Practice*. The MIT Press, Cambridge (12e dr.).

Brink, W.P. van den & Koele, P. (2005), *Statistiek*. *Toepassingen* (deel 3). Amsterdam: Boom.

Campbell, D. & Stanley, J. (1963), *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences.* Londen/New Jersey: Lawrence Erlbaum.

Cronbach, L.J. & Furby, L. (1970), How should we measure change – or should we? *Psychological Bulletin,* 74, 68-80.

Dehghani, B. & Lange A. (1993). Het meten van agressiviteit: Validatie van de Nederlandse versie van de Buss-Durkee Hostility Inventory. *Gedrag en Gezondheid, 21,* 298-304.

Drenth, P.J.D. & Sijtsma, K. (2006), *Testtheorie*. Houten: Bohn Stafleu Van Loghum (4e herz. Dr.; 1e dr.: 1965).

Ellis, J. (2007), *Statistiek voor de Psychologie. Toetsen voor twee gemiddelden en toetsingstheorie* (deel 2). Amsterdam: Boom.

Gageldonk, A. van & Bartels, A.A.J. (1990), *Evaluatieonderzoek in de jeugdhulpverlening. Resultaten van een overzichtsstudie*. Deel I: Resultaten. Deel II: Instrumenten. Leiden: DSWO-press.

Gageldonk, A. van & Bartels, A.A.J. (1991), Evaluatieonderzoek in de jeugdhulpverlening *Kind en Adolescent,* febr., 12 (1), 1-18.

Hafkenscheid, A., Kuipers, A. & Marinkelle, A. (1998). De vragenlijst als effectmaat bij 'N=1': hoe bruikbaar zijn statistische definities van 'klinische significantie' en betrouwbare verandering? *Gedragstherapie,* 31, 221-240.

Hays, W.L. (1970), *Statistics*. Londen/New York: Holt, Rinehart & Winston.

Hermans, D., Eelen, P. & Orlemans, J.W.G. (2007), *Inleiding tot de gedragstherapie*. Houten: Bohn Stafleu Van Loghum, 6e herziene druk.

Hersen, M. & Barlow, D.H. (1976), *Single-case experimental designs: Strategies for understanding behavior.* New York: Pergamon.

Kazdin, A.E. (1981), Drawing valid inferences from case studies. *Journal of Consulting and Clinical Psychology,* 49, 183-192.

Kerlinger, F.N. & Lee, H.B. (2000), *Foundations of Behavioral Research*. London: Wadsworth, Thomson Learning.

Lange, A. & Hoogendoorn, M. (1996). De Buss-Durkee Hostilty Inventory-Dutch (BDHI-D). *Gedragstherapie,* 29 (1), 55-60.

Lange, A., Pahlich, A., Sarucco, M. & Smits, G. (1993). Psychometrische kenmerken en validiteit van de Nederlandse Buss-Durkee Hostility Inventory (BDHI-N). *Nederlands Tijdschrift voor de Psychologie,* 48, 234-236.

Lehman, E.I. & D'Ambrera, H.J.M. (1975), *Nonparametrics. Statistical Methods Based on Ranks*. San Fransisco: Holden-Day & New York: McGraw-Hill.

Lindgren, B. (1993), *Statistical Theory*. New York: Chapman & Hall.

Maassen, G.H. (2003), Principes voor het meten van reliable change (2): reliable change indices and practice effects. *Nederlands tijdschrift voor de Psychologie,* 58, 69-79.

McCain, L.J. & McCleary, R. (1979), The Statistical Analysis of the Simple Interrupted Time-series Quasi-experiment. In: T.D. Cook & D.T. Campbell (eds.), *Qausi-experimentation: Design and analysis issues for filed settings*. Chicago: Rand McNally. Pp. 233-293. (Aangehaald bij Van Beek, 2007.)

Melief, W., Hoekstra, M., Langerak, E., Sijben, N. & Wevers, I. (1979), *Doelen stellen en evalueren. Een handleiding tot het gebruik van Goal attainment Scaling* (GAS). Alphen a/d Rijn: Samsom (samson sociale en Culturele Reeks).

Meerling (Subfaculteit psychologie, vakgroep Methoden & Technieken, Rijksuniversiteit Leiden)(1989), *Methoden en technieken van psychologisch onderzoek. Model, observatie, beslissing* (deel 1). Meppel: Boom. Pp. 109-129 (vooral 111).

Meesters, C., Muris, P., Bosma, H., Schouten, E. & Beuving, S. (1996). Psychometric evaluation of the Dutch version of the Aggression Questionnaire. *Behaviour Research and Therapy,* 34 (10), 839-843.

Nakagawa, S. (2004), A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behavioral Ecology,* June, 15 (6), 1044-1045.

Novaco, R.W. (1994). Anger as a risk factor for violence among the mentally disordered. In J. Monahan & H.J. Steadman (eds.), *Violence and Mental Disorder*. Chicago: The University of Chicago Press. Pp. 21-59.

Novaco, R.W. & Renwick, S.J. (1998). Anger predictors of the assaultiveness of forensic hospital patients. In E. Sanavio (ed.), *Essays in honor of H.J. Eysenck: Behavior and Cognitive Therapy today.* London: Pergamon Press.

Rossi, P.H., Lipsey, M.W. & Freeman, H.E. (2004). *Evaluation. A systematic approach* (7th ed.). Thousand Oaks: Sage.

Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston/New York: Houghton Mifflin Co.

Sheskin, D. (2004), *Handbook of parametric and nonparametric statistical procedures*. London/New York: Chapman & Hall/CDR. Pp. 633-664 en 725.

Spreen, M., Stam, G. & A.A.J. Bartels (2003), N=1-statistiek in de Dr. S. van Mesdagkliniek. Een praktisch voorbeeld van de SCL-90 Klachtenlijst als effectindicator van therapie. *GGz Wetenschappelijk,* 7 (2), 12-20.

Strien, P.J. van (1986), *Praktijk als wetenschap. Methodologie van sociaal-wetenschappelijk handelen.* Assen: Van Gorcum.

Sijben, N., Zwaan, E.J. & Zwart, F.M. (1979).Gedragstherapeutische middelen: beschrijving van een aantal meetinstrumenten. In: J.W.G. Orlemans, P. Eelen, W.P. Haaijman (red.), *Handboek voor Gedragstherapie,* hfdst. D3, p.2-7 (afl. 5). Deventer: Van Loghum Slaterus.

Swanborn, P.G. (1999), *Evalueren. Het ontwerpen, begeleiden en evalueren van interventies.* Amsterdam: Boom.

Swanborn, P.G. (2003), *Case-study's. Wat, wanneer en hoe?* Amsterdam/Meppel: Boom (3e dr., 1e: 1996).

Task Force on Promotion and Dissemination of Psychological Procedures, Division of Clinical Psychology, American Psychological association (1995), Training in dissemination of empirically-validated psychological treatments: Report and recommendations. *The Clinical Psychologist,* 48, 3-23.

Teeken, V. (2006), *Cognitieve gedragstherapie bij een 34-jarige man met een Autisme Spectrum Stoornis en slaap- en spanningklachten. N=1-studie in het kader van de VGCt-opleidingsroute*. Opvraagbaar voor leden VGCt: info@vasthi.nl.

Todman, J.B. & Dugard, P. (2001*), Single-Case and Small-n Experimental Designs. A Practical Guide to Randomization Tests.* London/New Jersey: Lawrence Erlbaum.

Veerman, J.W. (2006), Methoden voor het kwantificeren en toepassen van effecten. In: Projectgroep zicht op effectiviteit (T. van Yperen & J.W. Veerman, red.), *Zicht op effectiviteit. Bronnenboek voor praktijkgestuurd onderzoek in de jeugdzorg*. Deel 3. NIZW/Praktikon/VWS (www.jeuginterventies.nl), voorlopige uitgave (nulversie). Pp. 69-88 (hfdst. 15).

Veerman, J.W., Damen, H. & Brink, L.T. ten (2000). *Een werkmodel voor evaluatieonderzoek in de jeugdzorg*. Nijmegen: Academisch Centrum / Praktikon.

Walsh, B. (2004), Multiple Comparisons: Bonferroni and False Discovery Rate. *Lecture Notes for EEB581,* May, 1-17.

Williams, R.H. & Zimmerman, D.W. (1996), Are Simple Gain Scores Obsolete? *Applied Psychological Measurement,* March, 20 (1), 59-69.

Yperen, T. van & Bijl, B. (2006), De opzet van praktijkgestuurd onderzoek. In: Projectgroep zicht op effectiviteit (red.: T. van Yperen & J.W. Veerman), *Zicht op effectiviteit. Bronnenboek voor praktijkgestuurd effectonderzoek in de jeugdzorg.* Deel 3. Pp. 7-30 (hfdst. 12).

Yperen, T. van, Bijl, B. & Veerman, J.W. (2006), Op weg naar veelbelovend. In: Projectgroep zicht op effectiviteit (red.: T. van Yperen & J.W. Veerman), *Zicht op effectiviteit. Bronnenboek voor praktijkgestuurd onderzoek in de jeugdzorg*. Deel 1. NIZW/Praktikon/VWS, voorlopige uitgave (nulversie). Pp. 21-37.

Bijlage 1: Itemscores op de SCL90-schalen van de cliënt uit hoofdstuk 7 (tabellen 28)

Schaal 1, Items Agorafobie (7) – Geen vooruitgang verwacht	M1	M2	M3	M4
13 Angst openbare ruimten	2	2	2	1
25 Bang alleen uit huis te gaan	1	1	1	1
47 Reizen in Openbaar Vervoer	2	2	1	2
50 Bepaalde dingen vermijden	3	2	2	2
70 Niet op gemak bij anderen.	4	3	3	3
75 Zenuwachtig alleen gelaten te worden	1	1	1	1
82 Bang flauw te vallen	1	1	1	1

Schaal 2, Items Angst (10) – Vooruitgang verwacht	M1	M2	M3	M4
2 Zenuwachtig	3	2	2	2
17 Trillen	2	2	2	2
23 Plots schrikken	2	2	2	2
33 Bang voelen	2	2	2	2
39 Hartklop	2	2	2	2
57 Gespannen	3	2	2	2
72 Angst/paniek	3	2	2	2
78 Rusteloos	1	1	2	2
80 Gevoel dat nare gebeurtenis zal plaatsvinden	1	1	1	1
86 Angst/gedachten over angst	2	1	1	1

Schaal 3, Items depressie (16) – Vooruitgang verwacht	M1	M2	M3	M4
3 Nare gedachten/idee niet kwijt	2	1	1	1
5 Geen seks interesse	1	1	1	1
14 Weinig energie	1	1	2	1
15 Denken aan eind	1	1	1	1
19 Weinig eetlust	1	1	1	1
20 Gauw huilen	1	1	1	1
22 Verstrikt	2	1	1	2
26 Jezelf schuld geven	2	1	1	1
29 Eenzaam voelen	1	1	1	1
30 In de put	2	1	1	1
31 Piekeren	2	2	2	3
32 Nergens belangstelling	1	1	1	1
51 Gevoel leegte	2	1	1	1
54 Wanhopig toekomst	1	1	1	1
59 Denken dood	1	1	1	1
79 Niets waard	1	1	1	1

Schaal 4, Items insufficiëntie denken en handelen (9) – Vooruitgang verwacht	M1	M2	M3	M4
9 Moeite onthouden	2	2	2	2
10 Piekeren	3	2	2	3
28 Belemmering uitvoer	1	1	1	1
38 Langzaam wegens steeds controleren	4	3	2	3
45 Steeds controleren	4	2	2	2
46 Moeite beslissingen nemen	3	2	2	2
55 Moeite concentreren	1	1	2	1
65 Zelfde moeten doen	3	2	2	2
71 Gevoel alles moeite	1	1	2	1

Schaal 5, Items Somatiek (12) – Geen vooruitgang verwacht	M1	M2	M3	M4
1 Hoofdpijn	2	2	2	3
4 Duizelig	2	2	3	2
12 Pijn op borst	2	2	2	2
27 Rug	1	2	2	2
40 Misselijk	2	2	2	2
42 Spierpijn	3	2	3	2
48 Adem	2	2	1	1
49 Warm/koud	2	2	2	2
52 Verdoofd gevoel	2	2	2	2
53 Brok in keel	3	3	3	2
56 Slap voelen	2	1	2	1
58 Zware ledematen	3	2	2	2

Schaal 6, Items Wantrouwen & Interpersoonlijke Sensitiviteit (18) – Geen vooruitgang verwacht	M1	M2	M3	M4
6 Kritisch staan tegenover anderen	1	2	2	1
7 Idee ander beheerst jouw leven	1	1	1	1
8 Gevoel ander schuld aan jouw problemen	1	1	1	1
18 Gevoel anderen niet te vertrouwen	2	1	1	1
21 Verlegen/niet op gemak	1	1	1	1
34 Snel gekwetst voelen	1	1	1	1
35 Idee dat anderen jouw geheime gedachten kennen	1	1	1	1
36 Gevoel dat anderen je niet begrijpen	1	1	1	2
37 Gevoel dat anderen onaardig zijn	1	1	1	1
41 Tegenover anderen mindere voelen	1	1	1	1
43 Gevoel dat anderen je in gaten houden	1	1	1	1
61 Niet op gemak als anderen kijken	2	1	1	1
68 Gedachte dat anderen je niet begrijpen	2	1	1	1
69 Pijnlijk bewust eigen aanwezigheid	1	1	1	1
73 Niet op gemak eten in openbare ruimtes	2	3	3	2
76 Gevoel dat anderen je niet op juiste waarde schatten	1	1	1	1
83 Gevoel dat anderen misbruik van je maken	2	2	1	2
88 Nooit nauw verbonden	2	2	2	2

Schaal 7, Items hostiliteit (6) – Geen vooruitgang verwacht	M1	M2	M3	M4
11 Gemak geïrriteerd	1	2	2	2
24 Woede-uitbarstingen	1	1	1	1
63 Aandrang slaan	1	1	1	1
67 Aandrang vloeken	1	1	1	1
74 Vaak in ruzies verzeild raken	1	1	1	1
81 Schreeuwen/smijten	1	1	1	1

Schaal 8, Items slaapproblemen (3) – Vooruitgang verwacht	M1	M2	M3	M4
44 Moeilijk in slaap	5	5	4	3
64 Te vroeg wakker	4	4	3	2
66 Onrustig/gestoorde slaap	5	4	3	3

Schaal 9, Items Overige (7) – Geen vooruitgang verwacht	M1	M2	M3	M4
16 Stemmen horen	1	1	1	1
60 Te veel eten	1	1	1	1
62 Gedachten niet jezelf	1	1	1	1
77 Alleen voelen	2	1	1	1
84 Gedachten seks	1	1	1	1
85 Gedachten straf	1	1	1	1
87 Idee iets verkeerd met lichamelijke klachten	1	1	2	1
89 Schuldgevoelens	1	1	1	1
90 Gedachten dat psychisch niet in orde	2	1	1	1

Schaal 10, Totaal (90): vooruitgang verwacht bij $\alpha = 0.10$.

Bijlage 2: Wilcoxon gepaarde steekproef
Non-parametrische toets (Wilcoxon) in plaats van de t-toetsen uit § 10.6.2

In plaats van t-toetsen (tabellen 9.1 en 9.2) is hier de Wilcoxontoets voor gepaarde steekproeven uitgevoerd, de non-parametrische pendant van de t-toets

Tabel 29: t-toetsen en Wilconxon vergeleken

	Gemiddelden				t-toetsen			Wilcoxon		
SCL90-schaal	**M1**	**M2**	**M3**	**M4**	**M1&M2**	**M1&M3**	**M1&M4**	**M1&M2**	**M1&M3**	**M1&M3**
1 Agorafobie	2,00	1,71	1,57	1,57	0,172	0,078°	0,078°	0,180	0,109	0,109
2 Angst#	2,10	1,70	1,80	1,80	0,019*	0,097°	0,097°	0,034*	0,113	0,113
3 Depressie#	1,38	1,06	1,13	1,19	0,010*	0,052°	0,094°	0,022*	0,071°	0,022*
4 Insuffiëntie van denken en handelen#	2,44	1,78	1,89	1,89	0,011*	0,090°	0,026*	0,022*	0,089°	0,034*
5 Somatiek	2,17	2,00	2,17	1,92	0,337	1,000	0,275	0,361	1.000	0,311
6 Wantrouwen en interpersoonlijke sensitiviteit	1,33	1,28	1,22	1,22	0,668	0,430	0,331	0,686	0,463	0,361
7 Hostiliteit	1,00	1,17	1,17	1,17	0,363	0,363	0,363	0,317	0,317	0,317
8 Slaapproblemen#	4,67	4,33	3,33	2,67	0,212	0,029*	0,042*	0,159	0,054°	0,042*
9 Overige	1,22	1,00	1,11	1,00	0,169	0,594	0,169	0,090	0,297	0,090

Toelichting en bespreking van de toetsen
- M1, M2, M3 en M4: meting 1, 2, 3 en 4. Tussen elke meting lag circa 3 maanden.
- Cursief en met # in de meest linkse kolom de schalen waarop vooruitgang werd verwacht.
- In de 4 kolommen onder gemiddelden de gemiddelde itemwaarde per schaal bij de 4 metingen. Hier zijn dus gemiddelden per schaal gegeven, als in de tabel 2, waar in de tabellen 3, 4.1 en 4.2 per schaal 4 totalen werden vermeld.
- Als eerder aangegeven Wilcoxon-toets voor gepaarde steekproeven, correctie voor ties (zelfde scores op verschillende items van de schaal).
- Gegeven zijn p-waarden. Significante verschillen op niveau $\alpha = 0.05$ zijn donker gearceerd en met * aangegeven, significante verschillen op niveau $\alpha = 0.10$ zijn licht gearceerd en met ° aangegeven (significante p-waarden tussen 0.05 en 0.10 worden als trend beschouwd).
- Verschil Slapen van M1 naar M4 was ook eerder al met Wilcoxon getoetst omdat de standaardmeetfout van het verschil 0 was waardoor geen t-waarde berekend kon worden.
- Verschillen tussen M1 enerzijds en M2, M3 en M4 anderzijds zijn éénzijdig getoetst bij de schalen waarop vooruitgang verwacht werd: er werden verschillen verwacht, verbetering van M2 ten opzichte van M1, daarna voortgaande verbetering of gelijkblijven bij M3 en M4. Verschillen tussen M2, M3 en M4 onderling van deze schalen zijn ook éénzijdig getoetst bij de schalen waarop vooruitgang verwacht werd: áls er verschil zou zijn, werd dat geacht een verbetering te zijn of gelijk blijven.
  Bij schalen waarop geen vooruitgang verwacht werd, is tweezijdig getoetst.

Ter vergelijking is hierna multiple regressie uitgevoerd met een dummy-variabele. Deze heeft waarden 1 (steeds voor meting 1) en 0 (steeds voor de metingen 2, 3 en 4 als die met meting 1 vergeleken worden). Gegeven zijn de p-waarden in het vak 'Correlaties dummyvariabele'. Ook nu is éénzijdig getoetst waar verschil verwacht werd en tweezijdig als dat niet verwacht werd.

Tabel 30: Multiple regressie met dummy-variabele

	Gemiddelden				Correlaties dummyvariabele		
SCL90-schaal	**M1**	**M2**	**M3**	**M4**	**M1&M2**	**M1&M3**	**M1&M4**
1 Agorafobie	2,00	1,71	1,57	1,57	-	0,433	0,433
2 *Angst*[#]	2,10	1,70	1,80	1,80	0,085°	0,085°	0,140
3 *Depressie*[#]	1,38	1,06	1,13	1,19	0,017*	0,055°	0,017*
4 *Insuffiëntie van denken en handelen*[#]	2,44	1,78	1,89	1,89	0,027*	0,029*	0,048*
5 Somatiek	2,17	2,00	2,17	1,92	0,430	1,000	0,275
6 Wantrouwen en interpersoonlijke sensitiviteit	1,33	1,28	1,22	1,22	0,756	0,524	0,471
7 Hostiliteit	1,00	1,17	1,17	1,17	0,341	0,341	0,341
8 *Slaapproblemen*[#]	4,67	4,33	3,33	2,67	0,391	0,092°	0,037*
9 Overige	1,22	1,00	1,11	1,00	0,075°	0,278	0,075°

Toelichting en bespreking van de toetsen

- M1, M2, M3 en M4: meting 1, 2, 3 en 4. Tussen elke meting lag circa 3 maanden.
- Cursief en met [#] in de meest linkse kolom de schalen waarop vooruitgang werd verwacht.
- In de 4 kolommen onder gemiddelden de gemiddelde itemwaarde per schaal bij de 4 metingen.
- Gegeven zijn p-waarden. Significante verschillen op niveau $\alpha = 0.05$ zijn donker gearceerd en met * aangegeven, significante verschillen op niveau $\alpha = 0.10$ zijn licht gearceerd en met ° aangegeven (significante p-waarden tussen 0.05 en 0.10 worden als trend beschouwd).
- Verschillen tussen M1 enerzijds en M2, M3 en M4 anderzijds zijn éénzijdig getoetst bij de schalen waarop vooruitgang verwacht werd: er werden verschillen verwacht, verbetering van M2 ten opzichte van M1, daarna voortgaande verbetering of gelijkblijven bij M3 en M4.
- Bij schalen waarop geen vooruitgang verwacht werd, is tweezijdig getoetst.
- Zoals in § 11.3.4 uitgelegd is de onderlinge correlatie tussen de items alleen bij Slapen hoog (over M1 t/m M4: $r = .61$, $p = .035$). In § 11.3.4 werd daarvoor gecorrigeerd door te werken met partiële correlatie: op de correlatie van de scores met de dummyvariabele wordt voor de onderlinge correlatie tussen de items gecorrigeerd. Doet men (éénzijdig toetsen) dit dan is de partiële correlatie van de itemscores met de dummyvariabele voor M1-M2: $r = .50$, $p = .391$; voor M1-M3: $r = .82$, $p = .092$; en voor M1-M4: $r = .84$, $p = .037$.

Bijlage 3: Somwaardenstatistiek

Er zijn bijvoorbeeld 3 beoordelaars, A, B en C. Of er zijn 3 items van een schaal: de items 1, 2 en 3. We spreken hierna alleen over beoordelaars, maar alles wat daarvoor geldt, is ook van toepassing op items van een schaal.
Stel dat er sprake is van een 5-puntsschaal wat betreft beoordelingen.
De scores van de drie beoordelaars A, B en C zijn bijvoorbeeld respectievelijk 1, 2 en 3. Dit wordt als volgt weergegeven: {1,2,3}. Deze {1,2,3} heet de *uitgangsverdeling* U. En bestaat uit 3 scores: 1, 2 en 3. Het aantal scores van de uitgangsverdeling {1,2,3} is 3. Dit is n. Dus: n = 3 in dit geval. De som van de scores van de uitgangsverdeling is 1 + 2 + 3 = 6. $S_n = S_3 = 6$. Gedetailleerder: $S_{U1(3)} = 6$. Dit betekent: de som van de scores van uitgangsverdeling U_1 met 3 items is 6.
Elke score heeft een *waarschijnlijk scorebereik*. Dat wil zeggen: A scoort 1, maar had wellicht ook 0 kunnen scoren (als er een mogelijkheid tot het scoren van 0 geweest zou zijn) of 2. Zo scoort B nu 2, maar hij had wellicht ook 1 of 3 kunnen scoren. Dit wordt als volgt weergegeven: A scoort (0,1,2), B scoort (1,2,3) en C scoort (2,3,4). (0,1,2) of (2,3,4) heet een *element*. Het is een element van de uitgangsverdeling.
De uitgangsverdeling bestaat dus uit scores, elke score kan worden weergegeven als een element. Ofwel: {1,2,3} = (0,1,2) + (1,2,3) + (2,3,4). De scores 0, 1 en 2; 1, 2 en 3; en 2, 3 en 4 zijn de *elementscores*. Elk element heeft (hier) 3 scores.
Alle combinaties van elementscores worden nu berekend. Dus de 0 van A met 1 van B en 2 van C t/m de 2 van A + de 3 van B en de 4 van C. Dit levert de volgende tabel.

A	B	C	Som	A	B	C	Som	A	B	C	Som
0	1	2	**3**	1	1	2	**4**	2	1	2	**5**
0	1	3	**4**	1	1	3	**5**	2	1	3	**6**
0	1	4	**5**	1	1	4	**6**	2	1	4	**7**
0	2	2	**4**	1	2	2	**5**	2	2	2	**6**
0	2	3	**5**	1	2	3	**6**	2	2	3	**7**
0	2	4	**6**	1	2	4	**7**	2	2	4	**8**
0	3	2	**5**	1	3	2	**6**	2	3	2	**7**
0	3	3	**6**	1	3	3	**7**	2	3	3	**8**
0	3	4	**7**	1	3	4	**8**	2	3	4	**9**

Dus 1 maal *somwaarde* of *somscore* 3, idem 1 maal 9 etc. In totaal: 1 × 3, 1 × 9, 3 × 4, 3 × 8, 6 × 5, 6 × 7, en 7 × 6.

Deze somwaarden zijn in een verdeling te zetten en in een grafiek uit te drukken. Zie hierna.

Grafiek B.1: Somwaarden U_1 = {1,2,3}

7							
6							
5							
4							
3							
2							
1							
	3	**4**	**5**	**6**	**7**	**8**	**9**

De uitgangsverdeling was {1,2,3}. Stel dat dit de aanvangsmeting was. Er volgt een follow-up meting en die levert op {3,3,3}. De uitgangsverdeling bestaat nu uit drie dezelfde scores en derhalve drie dezelfde elementen: {3,3,3} = 3 × (2,3,4).

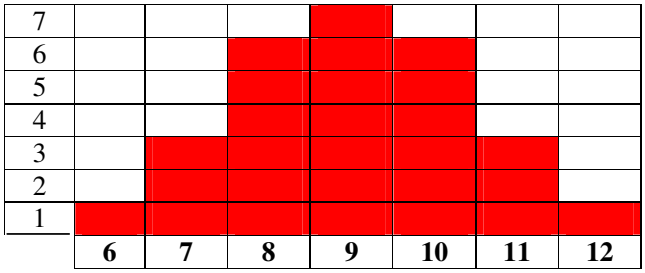
We krijgen de volgende tabel. De somscores van de drie beoordelaars zijn dan:

A	B	C	Som	A	B	C	Som	A	B	C	Som
2	2	2	**6**	3	2	2	**7**	4	2	2	**8**
2	2	3	**7**	3	2	3	**8**	4	2	3	**9**
2	2	4	**8**	3	2	4	**9**	4	2	4	**10**
2	3	2	**7**	3	3	2	**8**	4	3	2	**9**
2	3	3	**8**	3	3	3	**9**	4	3	3	**10**
2	3	4	**9**	3	3	4	**10**	4	3	4	**11**
2	4	2	**8**	3	4	2	**9**	4	4	2	**10**
2	4	3	**9**	3	4	3	**10**	4	4	3	**11**
2	4	4	**10**	3	4	4	**11**	4	4	4	**12**

Ofwel: 1 × 6, 3 × 7, 6 × 8, 7 × 9, 6 × 10, 3 × 11 en 1 × 12.

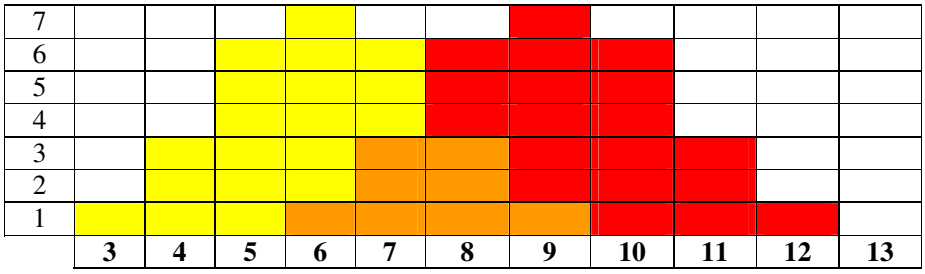
Ook dit kan in een grafiek worden uitgedrukt. Zie hierna.

Grafiek B.2: Somwaarden U_1 = {3,3,3}



Men kan de beide verdelingen samenvoegen en zien hoeveel procent overlap er is.
De gele cellen geven weer {1,2,3}, de rode geven weer{3,3,3}. De oranje cellen geven de overlap.

Grafiek B.3: Somwaarden van U_1 {1,2,3} en U_2 {3,3,3} met overlap/overeenkomst



Toelichting

Bij tellen blijkt elke verdeling 27 cellen te hebben. Er is bij 8 overlap. Dat is 8/27 x 100% = 29,6%. We beschouwen beide uitgangsverdelingen als verschillend als hun overlappercentage van hun grafiek gelijk aan of minder is dan 50%. In dit geval is dat zo (29,6%). Derhalve beschouwen we de uitgangsverdelingen {1,2,3} en {3,3,3} als structureel en objectief verschillend.

Tot nu is alles uitgeschreven en konden we tellen. Maar met bijvoorbeeld 5 of 8 beoordelaars (dan wel 5 of 8 items), om nog maar niet te praten over 20 of 30 of meer items, wordt dat ondoenlijk, en moet men naar algemene formules.

Eerst een definiëring van termen. Daarbij wordt aangesloten bij het tot dusver gehanteerde voorbeeld {1,2,3}.

De uitgangsverdeling U ={a,b,c} heeft 3 scores (scores van de uitgangsverdeling): a, b en c.
De uitgangsverdeling bestaat uit 3 elementen: {a,b,c} = (a-1,a,a+1) + (b-1,b,b+1) + (c-1,c,c+1).
n is het aantal scores van de uitgangsverdeling (en het aantal elementen) en is nu 3.
De som van de scores van de uitgangsverdeling S_n is: $S_n = a + b + c$. $S_3 = a + b + c$. Wil men uitgeven bij welke U de som hoort, dan kan dat als volgt: $S_{U1(3)}$ betekent: de som van uitgangsverdeling 1 (U_1) met n = 3.
De uitgangsverdeling heeft een aantal somscoresfrequenties of somwaardenfrequenties. In het voorbeeld {1,2,3} was dat: 1×3, 3×4, 6×5, 7×6, 6×7, 3×8 en 1×9. In dit geval zijn er 7 somwaardenfrequenties (1, 3, 6, 7, 6, 3, 1 ofwel: 2×1, 2×3, 2×6, 1×7).

n_f: het aantal somwaardenfrequenties. $n_f = 2n + 1$.
De hoogste vóórkomende frequentie, is de middelste (mediaan) van de rij: 7 in de rij 1, 3, 6, 7, 6, 3, 1.
Het symbool is: i_{mf}. Het nummer van deze frequentie is n + 1. Bij n = 4 is het dus de vijfde op rij.

S_f is het totaal aan somswaarden of somscores bij een gegeven n. Dit is gelijk aan 3^n. In het voorbeeld: voor {1,2,3} geldt: $S_f = 3^3 = 27$.

Daarbij is 3 het 'waarschijnlijke scorebereik'. Dat wil zeggen als men, uitgaande van {1,2,3} komt tot (0,1,2) + (1,2,3) + (2,3,4) is het waarschijnlijke scorebereik (waarin de 'werkelijke score' valt) gelijk aan 3. Elk element, (0,1,2), (1,2,3) en (2,3,4) heeft 3 (element)scores.
Maar men had natuurlijk ook 5 kunnen kiezen). {1,2,3} leidt dan tot: (-1,0,1,2,3) + (0,1,2,3,4) + (1,2,3,4,5). Het waarschijnlijk scorebereik wordt gegeven met q. Het aantal cellen van een verdeling van somwaarden is derhalve q^n.
We werken overigens altijd met een waarschijnlijk scorebereik van 3. S_f is de som van alles frequentiewaarden, het totaal van alle frequenties.
Dus bij n = 3 is dat: $S_f = 1 + 3 + 6 + 7 + 6 + 3 + 1 = 2 \times 1 + 2 \times 3 + 2 \times 6 + 7 = 3^3 = 27$.

De frequentie met de hoogste frequentiewaarde, dus bij n = 3 is dat 7 (de frequenties zijn dan immers: 1, 3, 6, 7, 6, 3 en 1) heeft als somwaarde of somscore S_n. Bij n = 3, is 7 de hoogst vóórkomende frequentie en dat is bij somwaarde 6 (7×6). Want voor {1,2,3} geldt $S_n = 6$.
Hieruit volgt dat alle uitgangsverdelingen met eenzelfde S_n gelijk zijn qua vorm en alleen verschillen in hun plaatsing op de X-as.

Bovendien zijn alle uitgangsverdelingen symmetrisch. De hoogst somscore of somwaarde met de vóórkomende frequentie is daarom tevens het gemiddelde van de verdeling. De verdeling wordt derhalve qua vorm volledig door S_n en n gedefinieerd.

$f_{max(n)}$ is de hoogst vóórkomende frequentie(waarde) bij een gegeven n. In geval n = 3, is deze 7, bij n = 4 geldt: $f_{max(4)} = 19$.

De frequenties bij verschillende grootten van n zijn uit elkaar af te leiden.

N	**n_f**	**S_f**	**Frequenties**								
1	3					1	1	1			
2	5	$3^2 = 9$			1	2	3	2	1		
3	7	$3^3 = 27$		1	3	6	7	6	3	1	
4	9	$3^4 = 81$	1	4	10	16	19	16	10	4	1

Bij n = 1, dus bijvoorbeeld {2} zijn/is het element: (1,2,3). Elke score van het element komt éénmaal voor. Zie de tabel. Bij n = 2, dus bijvoorbeeld {2,3} = (1,2,3) + (2,3,4) zijn de frequenties 1, 2, 3, 2 en 1. Ofwel 1×3, 2×4, 3×5, 2×6 en 1×7. Het aantal frequenties is $2n + 1 = 2 \times 2 + 1 = 5$. Het aantal somwaarden (somscores) en cellen in een grafiek bij een uitgangsverdeling met n = 2, is: $3^2 = 9 = 1 + 2 + 3 + 2 + 1$.

Elke cel in de tabel hierboven is gelijk aan de som van de drie cellen erboven: de cel linksboven, recht erboven en rechtsboven. Zo is voor de hoogste frequentie bij n = 3: 2 + 3 + 2 = 7. En voor de derde frequentie op rij bij n = 4 (die 10 bedraagt) geldt: 10 = 1 + 3 + 6. Het symbool hiervoor is: $f_{max(n)}$. Bij n = 4 geldt dus: $f_{max(4)} = 19$.
Aldus is de tabel met behulp van een spreadsheet programma als Excel naar believen uit te breiden.

Het berekenen van de percentages overlap gaat als volgt.

1 Men neemt het aantal cellen van de verdeling van somwaarden. Dit is 3^n.
In het geval van n = 4, is dit dus: $3^4 = 81$.

2 Men bepaalt de sommen van beide uitgangsverdelingen, S_n. Neem bijvoorbeeld U_1 = {2,3,4,5} en U_2 = {3,3,4,5}. S_4 bij U_1 is: 2 + 3 + 4 + 5 = 14. S_4 bij U_2 is: 3 + 3 + 4 + 5 = 15. De gemiddeldes van de beide somwaardenverdelingen zijn derhalve 14 en 15. Dat is een verschil van 1. Dit noemen we k.

3 De meest vóórkomende frequentie is 19: $f_{max(4)} = 19$. Dit is te zien in de tabel.

4 De overlap wordt gegeven met de formule $3^n - \Sigma_{f(k)}$. Daarbij staat $\Sigma_{f(k)}$ voor de som van evenveel frequenties, te beginnen met $f_{max(n)}$, als er nodig zijn om twee sommen S_n te overbruggen. In geval van n = 4 en $S_{n1} = 14$ en $S_{n2} = 15$, wordt dat als volgt.
$3^n - \Sigma_{f(k)}$ wordt nu: $[3^4 - (19 + 16 + 16)] / 3^4 = (81 - 51)/81 = 30/81 = 0{,}370$.
Uitgedrukt als percentage: 37%. De waarden 19, 16 en 16 die werden afgetrokken van $3^4 = 81$ komen uit de tabel. Genomen wordt de hoogste frequentiewaarde plus alle in grootte daarop volgenden tot men k heeft bereikt; k is nu 3 (het verschil van S_{n1} met S_{n2}). dus er worden 3 waarden, te beginnen met $f_{max(n)}$ afgetrokken.

Over deze methode kan in algemene zin nog het volgende worden aangegeven.

De somwaarde met de hoogste frequentie van zo'n verdeling is gelijk aan de som van de n elementen van {a,b,c}, derhalve dus a + b + c. De somwaarden met frequentie 1 zijn a + b + c – n en a + b + c + 1. De somwaarden met de op één na laagste en op één na hoogste frequentie zijn gelijk aan het aantal elementen van {a,b,c} derhalve 3.

Uit het voorbeeld in tabel 16.1 en grafiek 5.1 over {5,4,5}: de somwaarde met de hoogste frequentie is 5 + 4 + 5 = 14, wat onder tabel 16.1 is aangegeven en in grafiek 5.1 te zien is. De somwaarden met frequentie 1 zijn 14 – 3 = 11 en 14 + 3 = 17.
Het aantal frequenties is 2n + 1. Dus bij {1,2} met 2 elementen is het aantal frequenties 5, namelijk: 1, 2, 3, 2 en 1; bij {1,2,3} is het aantal frequenties 7, namelijk: 1, 3, 6, 7, 6, 3 en 1.
Indien men werkt met 3 elementen, dus met {a,b,c} is het totaal aantal somwaarden 3^n. In het geval van 3 items dus 27. zoals ook uit de grafieken 5.1 t/m 5.5 naar voren komt.
Verdelingen als {a,b,c} zijn volledig bepaald door het gemiddelde van a, b en c, of anders gesteld: door de som van a + b + c. Indien de waarden van de elementen a, b of c veranderen, maar de som blijft hetzelfde, resulteert dezelfde verdeling.

Voor maximaal 10 items of beoordelaars, worden de percentages niet-overlap gegeven in de volgende tabel.

Percentages *niet-overlap* van twee somwaardenverdelingen bij een gegeven aantal beoordelaars, gegeven verschillen van somwaarden

	Aantal beoordelaars (of aantal items)									
Verschil tussen somscores	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	33	33	26	23	21	19	18	17	16	15
2	67	56	48	43	40	37	34	32	31	29
3	100	78	70	65	58	54	51	48	45	43
4	100	89	81	75	70	66	63	60	57	55
5		100	93	88	83	79	75	71	69	66
6		100	96	93	89	85	82	79	77	74
7		100	100	98	95	92	90	87	85	82
8		100	100	99	97	95	93	91	89	87
9			100	100	99	98	97	95	94	92
10			100	100	100	99	98	97	96	95
11			100	100	100	100	99	99	98	97
12			100	100	100	100	100	99	99	98
13				100	100	100	100	100	99	99
14				100	100	100	100	100	100	99
15				100	100	100	100	100	100	100

Bij grote n-waarden (> 30) is de berekening via Excel goed te doen, maar de bepaling van het percentage overlap blijft toch een heel rekenwerk. In principe geldt dat men dan zou kunnen werken met benaderingen vanuit de statistische normaalverdeling. Mits evenwel de somwaardenverdelingen door de standaardnormaalverdeling benaderd worden. Dit blijkt evenwel niet het geval.

Uitgaande van het gegeven dat de spreiding $s = \sqrt{\{\Sigma d^2/(N-1)\}} = \sqrt{\{\Sigma(X-\overline{X})^2/(N-1)\}}$ waarbij N het aantal scores van de verdeling is, is de volgende snelle spreidingsberekening mogelijk. Uitgegaan wordt van de n-waarde in de somwaardenstatistiek. Indien $U_1 = \{1,2,3\}$ dan is S_n en daarmee het gemiddelde 6 ($S_{U1(3)} = 6$). Alle somwaarden van de maximale frequentie (de waarde 6 komt 7 maal voor) kunnen worden vergeten want die leveren als bijdrage aan de spreidingsberekening waarde 0. Immers: $X-\overline{X} = 6 - 6 = 0$. Men heeft dus alleen te maken met de andere waarden. De andere waarden kleiner dan 6 en groter dan 6 komen overeen, de verdeling is volstrekt symmetrisch. De andere waarden zijn: 6×5, 6×7, 3×4 en 3×8, 1×3 en 1×9.
De 6 waarden van 5 verschillen elk 1 van het gemiddelde 6. Dat is dus $6 \times 1 = 6$. De 3 waarden van 4 verschillen elk 2 van het gemiddelde 6, levert op: $3 \times 2^2 = 12$. De ene waarde van 3 verschilt 3 van het gemiddelde 6, levert op: $3^2 = 9$. De som van de kwadraten van de verschillen met het gemiddelde voor de 'linkervleugel' van de verdeling, is daarmee: $6 + 12 + 9 = 27$. Voor de rechtervleugel geldt hetzelfde zodat $\Sigma(X-\overline{X})^2 = (2 \times 27) = 54$. En dus: $s = \sqrt{\{54/(27-1)\}} = \sqrt{2,08} = 1,44$.

Somwaardenverdeling en statistische normaalverdeling
De somwaardenverdeling is symmetrisch en de vraag kan opkomen of er vergelijkbaarheid is met de statistische normaalverdeling. Dat zou berekeningen vereenvoudigen. De somwaardenverdeling is evenwel niet met de statistische normaalverdeling vergelijkbaar, het verschil is te groot.
Aangezien de somwaardenverdelingen volledig symmetrisch zijn en alleen verschillen inzake hun plaats op de X-as, is de spreiding voor uitgangsverdeling U_2 ook: $s = 1,44$. Deze s wordt 1 in de standaardnormale verdeling (als de somwaardenverdeling normaal verdeeld zou zijn). U_1 had als gemiddelde 6 (de som van de elementen van de uitgangsverdeling), voor $U_2 = \{3,3,3\}$ is het gemiddelde 9. We zagen hiervóór dat U_1 en U_2 een overlap hebben van 8 van de 27 cellen is 30%. Gaan we dit na met de spreidingen, dan blijken de gemiddelden van beide verdelingen 3 te verschillen ($m_{U1} = 6$, $m_{U2} = 9$). Dat betekent dat de gemiddelden $3/1,44 = 2,08$ spreidingen uit elkaar te liggen.
Indien de somwaardenverdeling als een normale verdeling beschouwd wordt, is de rechteroverschrijdingskans van U1 = {1,2,3}bij $\sigma = 2,08$: 0,0188. De verdelingen zijn beide identiek en symmetrisch, de linkeroverschrijdingskans van U2 = {3,3,3} bij $\sigma = -2,08$ is daaraan gelijk. De overlap is daarmee $2 \times 0,0188 = 0,0376$. In procenten: 3,8%.
Dit is een heel andere en veel kleinere overlap dan de feitelijke overlap die we eerder vonden bij het vergelijken van de twee somwaardenverdelingen van U1 en U2, namelijk 29,6% → 30%. We moeten besluiten dat de somwaardenverdelingen wel symmetrisch zijn, maar niet als normaalverdeling beschouwd kunnen worden. Althans niet zonder transformaties of normalisering (zoals stanines). Somwaardenverdelingen benaderen vanuit de standaardnormaalverdeling is geeft meer problemen dan het oplost.

Spreiding somwaardenverdelingen
In algemene bewoordingen geformuleerd, wordt de s van een somwaardenverdeling berekend als de som van een kwadratenreeks. De kwadraten vinden hun oorsprong in de oorspronkelijke verschillen met het gemiddelde. Deze zijn 1, 2, 3 etc. De kwadratenreeks wordt dan 1, 4, 9, 16 etc. Het aantal keren dat elk kwadraat telt (het aantal kwadraten in de reeks) hangt af van de frequentiewaarde van de betreffende somwaarde. Bij n = 3 (3 beoordelaren), waaruit een verdeling met 27 somwaarden resulteert, geldt dus dat N = 27. Bij n = 4 (4 beoordelaren) heeft de verdeling $3^4 = 81$ somwaarden en zijn de frequenties: 1, 4, 10, 16, 19, 16, 10, 4, 1. De som van de kwadratenreeks, $\Sigma(X-\overline{X})^2$ wordt bij n = 4 dus als volgt bepaale: $2 (16 \times 1 + 10 \times 4 + 4 \times 9 + 1 \times 16) = 2 \times 108 = 216$. Derhalve $s = \sqrt{\{216/(81 - 1)\}} = \sqrt{2,7} = 1,64$. Berekeningen zijn eenvoudig met Excel uit te voeren.