praktikon

Vergelijken van resultaten van jeugdhulp

Kan dat als verschillende meetinstrumenten zijn gebruikt?

Luuk Geijsen
Ronald De Meyer
Marc Delsing
Rachel van der Rijken

Volledige referentie

Geijsen, L., de Meyer, R. E., Delsing, M. J. M. H., & van der Rijken, R. E. A. (2025). Vergelijken van resultaten van jeugdhulp: Kan dat als verschillende meetinstrumenten zijn gebruikt? *Onderzoek en Praktijk*, 64(2), 8-19. Juni 2025 (https://orthopedagogiek.eu/2025/05/26/vergelijken-van-resultaten-van-jeugdhulp/)

Inhoudsopgave

1 Samenvatting

Het systematisch inzetten van feedbacktools, zoals vragenlijsten, binnen de (jeugd)hulpverlening is een manier om de effectiviteit van de hulpverlening te evalueren en waar nodig te verbeteren. De Lerende Databank Jeugd is een databank met zorginhoudelijke gegevens van jeugdhulporganisaties waarin per behandelvorm de resultaten van de geboden hulp van de eigen organisatie worden afgezet tegen die van andere organisaties. Door resultaten onderling te vergelijken, kunnen organisaties van elkaar leren wat goed gaat en wat beter kan.
Om de vergelijking van resultaten mogelijk te maken, wordt gebruik gemaakt van statistische maten: de Effect size en de Reliable Change Index. Deze maten worden berekend uit scores op vragenlijsten die door ouders of kinderen worden ingevuld in het kader van de hulp. Veelgebruikte vragenlijsten binnen de jeugdhulp zijn de Child Behavior Checklist (CBCL) en de Strengths and Difficulties Questionnaire (SDQ). Deze vragenlijsten worden ingezet om veranderingen in probleemgedrag bij kinderen in kaart te brengen. Organisaties kiezen zelf welk instrument zij inzetten bij welke behandelvorm, maar hebben wel de behoefte om resultaten onderling te vergelijken. In dit artikel is onderzocht of dat verantwoord is als verschillende meetinstrumenten zijn gebruikt. Op basis van gegevens uit de LDJ is getoetst of behandelresultaten gemeten met de CBCL en de SDQ met elkaar te vergelijken zijn.
Uit het onderzoek blijkt dat de resultaten op beide vragenlijsten op basis van de RCI nauwelijks van elkaar verschillen. Met de CBCL wordt weliswaar een grotere Effect size behaald dan met de SDQ, maar er is geen significant interactie-effect. Dit betekent dat de verandering in probleemgedrag niet significant verschilt tussen beide meetinstrumenten. Op basis van deze bevindingen lijkt een vergelijking tussen behandelresultaten gemeten met de CBCL en de SDQ te verantwoorden, met daarbij wel enkele nuances die in dit artikel nader worden beschreven.

Kernwoorden: Benchmarken, vragenlijsten, probleemgedrag, praktijkonderzoek, behandelresultaten, Effect size, Reliable Change Index.

2 Inleiding

Het systematisch inzetten van feedbacktools, zoals vragenlijsten, binnen de (jeugd)hulpverlening is een manier om de effectiviteit van de hulpverlening te evalueren en waar nodig te verbeteren. Er zijn vanuit verschillende onderzoeken indicaties dat het systematisch verzamelen van gegevens binnen behandelingen en het bespreken van deze gegevens met de cliënt bijdragen aan een groter behandeleffect (o.a. Amble et al., 2014; Bovendeerd et al., 2021; Janse et al., 2020; Lambert et al., 2001). Hoewel jeugdhulporganisaties steeds meer gegevens verzamelen over de kwaliteit en de effectiviteit van de door hen geboden hulp, worden deze gegevens nog niet altijd optimaal benut. Professionals hebben namelijk ondersteuning nodig om de verzamelde gegevens op de juiste manier te kunnen interpreteren en effectief te kunnen inzetten tijdens de hulp (Van Yperen, 2013). Onder meer het trainen van professionals in het voeren van feedback- en verbetergesprekken, daarbij ondersteund door eenduidige resultaatoverzichten, draagt bij aan een betere benutting van verzamelde gegevens in gesprekken met cliënten en tussen professionals onderling (SEJN, 2020).

Dat nog veel winst te behalen valt in het benutten van beschikbare gegevens over de kwaliteit en effectiviteit van de jeugdhulp blijkt uit de Hervormingsagenda Jeugd 2023-2028. Eén van de opgaven in de Hervormingsagenda is namelijk: kwaliteits- en effectiviteitverbetering en blijvend leren. Het doel is beschikbare kennis over wat werkt beter te benutten en een structuur en cultuur neer te zetten waarin blijvend van elkaar wordt geleerd. Eén van de randvoorwaarden voor blijvend leren in de jeugdhulp is ‘meten om te verbeteren’ (Hervormingsagenda Jeugd 2023-2028, 2023). Het is dan ook niet verrassend dat een groeiend aantal organisaties voor (jeugd)hulpverlening in Nederland een digitaal platform voor het systematisch verzamelen en bespreken van behandelresultaten, zowel op cliëntniveau als op groepsniveau, implementeert. Daarnaast zijn er organisaties die hun gegevens samenvoegen ten behoeve van organisatie overstijgende kennisopbouw over de effectiviteit van de geboden hulp. Hierdoor ontstaat de mogelijkheid om met collega’s van andere organisaties in gesprek te gaan en behandeleffecten van gelijksoortige behandelingen met elkaar te vergelijken. Dit wordt ‘benchmarken’ genoemd. Benchmarken, mits op een methodologisch verantwoorde wijze uitgevoerd, draagt bij aan het gericht verzamelen van gegevens en het vergroten van wetenschappelijke kennis over de hulpverleningspraktijk.

In Nederland hebben verschillende initiatieven plaatsgevonden om te benchmarken met zorginhoudelijke gegevens, onder andere via Stichting Benchmark GGZ, AKWA, de Research Data Infrastructure, de Lerende Databank Jeugd en landelijke kenniscentra die de kwaliteit van een interventie bewaken, zoals MST, Triple P, Gezin Centraal en Hersenz. Niet al deze initiatieven zijn succesvol gebleken. Het samenvoegen van zorginhoudelijke gegevens is een uitdagende klus, die

vaak tijdintensief en dus kostbaar is. Daarnaast bestaat in de praktijk niet zelden weerstand tegen benchmarken, omdat deze term vaak geassocieerd wordt met een zogenaamde afrekencultuur. Dit speelt met name als het initiatief tot benchmarken niet vanuit de hulpverlenende organisaties zelf komt, maar extern georganiseerd is om de effectiviteit van de hulp en daarmee de inzet van beschikbare gelden te beoordelen (Algemene Rekenkamer, 2017; De Beurs, 2017). Wanneer een benchmark- of kwaliteitssysteem echter goed aansluit bij de behoeften van professionals en cliënten en bij de visie van de organisatie, is er meer draagvlak om ermee te werken, wat de impact op de hulpverleningspraktijk kan vergroten (Van Geffen, 2019).

De Lerende Databank Jeugd (LDJ) is een voorbeeld van een databank met zorginhoudelijke gegevens die op eigen initiatief van verschillende jeugdhulporganisaties is ontstaan. Deze organisaties hebben zich verenigd in het Samenwerkingsverband Effectieve Jeugdhulp Nederland (SEJN, zie https://www.sejn.nl/lerende-databank-jeugd/). Door de gegevens die deze 30 organisaties verzamelen samen te voegen in de LDJ ontstaat de mogelijkheid om resultaten van organisaties, teams en interventies met elkaar te vergelijken, daarvan te leren en de hulp zo nodig te verbeteren. Gegevens in de LDJ worden voor SEJN-organisaties op een overzichtelijke wijze weergegeven via LDJ-dashboards, waarbij zij de behandelresultaten van de eigen organisatie in relatie tot de resultaten van alle SEJN-organisaties kunnen zien. In de LDJ kunnen per behandelvorm de effecten van de geboden hulp (het verschil tussen een beginmeting en een eindmeting van de hulp) in de vorm van een Effect size (Cohen, 1988) van de eigen hulp vergeleken worden met de effecten van alle SEJN-leden tezamen. Daarnaast is het mogelijk om het behandeleffect uit te drukken in het percentage cliënten waarbij de problemen gelijk blijven, significant verbeteren of juist verslechteren. Hiervoor wordt per cliënt de Reliable Change Index (RCI; Jacobson & Truax, 1991) berekend, een methode om uit te rekenen of een waargenomen verandering significant is of waarschijnlijk op toeval berust. Deze RCI kan verrijkt worden met de klinische betekenis van de score op de meting aan het einde van de hulp, waardoor een praktische betekenis gegeven kan worden aan de resultaten van een behandeling: heeft de geboden hulp geleid tot een verslechtering dan wel verbetering en is de cliënt nagenoeg klachtenvrij aan het einde van de behandeling.

Een van de parameters om de zorg te evalueren is het meten van verandering in probleemgedrag bij kinderen die jeugdzorg krijgen. Probleemgedrag is in de Wet op de jeugdzorg vertaald in opgroeiproblemen, die aanleiding kunnen zijn voor het verstrekken van zorg. Om de ernst van de opgroeiproblemen in de zorg te meten, worden verschillende meetinstrumenten ingezet die emotionele en gedragsproblemen in kaart brengen. Twee van de meest gebruikte meetinstrumenten hiervoor zijn de Child Behavior Checklist (CBCL; Achenbach & Rescorla, 2001) en de Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997). Ook binnen SEJN worden deze twee meetinstrumenten gebruikt; organisaties maken hierin hun eigen keuzes. Omdat de SEJN-organisaties en teams hun resultaten toch met elkaar willen delen en bespreken, ontstond

de vraag of resultaten gemeten met de CBCL en de SDQ vergelijkbaar zijn. In dit artikel willen we dit nader onderzoeken.

De CBCL/6-18 en de SDQ-O hebben dezelfde meetpretentie, worden afgenomen bij dezelfde informanten (opvoeders) en beschikken over vergelijkbare normgroepen. Daarnaast tonen verschillende onderzoeken aan dat deze instrumenten vergelijkbaar zijn wat betreft sensitiviteit en specificiteit van de totaalscore voor gedragsproblemen (zie o.a. Goedhart et al., 2003; Hendriks et al., 2020; Klasen et al., 2000; Vogels et al., 2005; Warnick et al., 2008). Het is echter niet bekend of beide instrumenten even gevoelig zijn voor het meten van verandering. Uit een onderzoek van De Beurs et al. (2015) bleek dat de Effect sizes en RCI's even groot waren bij cliënten die een SDQ hadden ingevuld als bij cliënten die een CBCL/6-18 hadden ingevuld. Het artikel maakt echter niet duidelijk of beide groepen cliënten een zelfde type behandeling hadden gekregen. Dit bemoeilijkt de interpretatie van de vergelijkbaarheid van de effecten behaald met beide instrumenten.

Om te kunnen bepalen of jeugdhulporganisaties resultaten op de SDQ en CBCL daadwerkelijk met elkaar kunnen vergelijken, is meer onderzoek nodig. In dit artikel wordt daarom op basis van praktijkdata die zijn verkregen uit de LDJ, getoetst of de behandelresultaten gemeten met de CBCL en de SDQ bij cliënten die een vergelijkbare zorgvorm hebben ontvangen, met elkaar te vergelijken zijn. Meer kennis hierover draagt bij aan een verantwoorde uitwisseling van gegevens tussen jeugdhulporganisaties en teams van professionals met als doel betere (verbeter)gesprekken te kunnen voeren over de kwaliteit en de effectiviteit van de geboden hulp.

3 Methode

3.1 Deelnemers

Gegevens zijn verzameld bij opvoeders (hierna 'ouders') van kinderen die hulp kregen bij verschillende jeugdhulporganisaties die zijn aangesloten bij het SEJN en bij wie de hulp is beëindigd vóór 22 juni 2021. De ouders hebben een CBCL/6-18 of een SDQ-O ingevuld bij aanvang én einde van de hulp. Over 81,6 % van de kinderen vulde de moeder een vragenlijst in, over 10,6 % de vader, over 6,6 % vulden beide ouders de lijst samen in en over 1,3 % werd de lijst door een andere opvoeder ingevuld. De gemiddelde leeftijd van de kinderen bij aanvang van de hulp was 11,03 jaar (SD = 3,51), 60,31% waren jongens en de gemiddelde behandelduur was 39,13 weken (SD = 18,20).

Om de hulp die cliënten hebben gehad binnen de verschillende organisaties onderling te kunnen vergelijken, is binnen het SEJN een standaard zorgproductentabel opgesteld met de meest voorkomende hulpvormen die worden geboden binnen de organisaties (zie https://www.bergop.info/wp-content/uploads/2021/05/Zorgproducten-LDJ.pdf). Alle kinderen en hun ouders hebben hulp ontvangen in de zorgproductcategorie 'Jeugdhulp Ambulant'.

3.2 Materiaal

Child Behaviour Checklist

De Child Behavior Checklist (CBCL, Achenbach & Rescorla, 2001; Verhulst & Van der Ende, 2013) meet emotionele en gedragsproblemen bij kinderen en jeugdigen. De CBCL wordt door ouders ingevuld en kent twee versies voor verschillende leeftijden van de kinderen: de CBCL voor kinderen van 1,5 tot en met 5 jaar (CBCL/1,5-5) en de CBCL voor kinderen en jongeren van 6 tot en met 18 jaar (CBCL/6-18). In dit onderzoek is de CBCL/6-18 gebruikt. De CBCL/6-18 bestaat uit 120 vragen over emotionele en gedragsproblemen, die worden gescoord op een driepuntsschaal: (0) 'Niet', (1) 'Soms', (2) 'Vaak'. Een groot deel van deze vragen behoort tot specifieke schalen voor emotionele en gedragsproblemen: Angstig/Depressief, Teruggetrokken/Depressief, Lichamelijke Klachten, Sociale Problemen, Denkproblemen, Aandachtsproblemen, Regelovertredend Gedrag en Agressief Gedrag. Een aantal van deze probleemschalen zijn te groeperen rond twee hoofddimensies: Internaliseren (de drie eerstgenoemde schalen) en Externaliseren (de twee laatstgenoemde schalen). De som van alle items vormt de 'Totaalscore'. Voor elke afzonderlijke schaal en de Totaalscore zijn normgegevens voor leeftijd en geslacht beschikbaar. Om de gevoeligheid voor verandering van de CBCL met de SDQ te kunnen vergelijken, is in dit onderzoek de Totaalscore op de CBCL gebruikt. Omwille van de

vergelijkbaarheid met de SDQ zijn de T-scores van de CBCL omgezet in deviatiescores (formule: Tscore - 50 / 10). De betrouwbaarheid van de Totaalscore op de CBCL is goed (Cronbach's alpha = 0,90; test-hertest betrouwbaarheid = 0,94; Achenbach & Rescorla, 2001; Verhulst & Van der Ende, 2013).

Strengths and Difficulties Questionnaire

De Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997) meet de aanwezigheid van emotionele en gedragsproblemen, sociale competenties, en de gevolgen van aanwezige problemen voor het dagelijks functioneren bij kinderen en jeugdigen van 4 tot en met 18 jaar. Er zijn drie versies beschikbaar: één voor ouders, één voor leerkrachten en één voor jongeren van 11 tot en met 16 jaar. In dit onderzoek is de SDQ voor ouders (SDQ-O) gebruikt. De gedragsvragenlijst bevat 25 vragen die deels negatief, deels positief geformuleerd zijn en gescoord worden op een driepuntsschaal (niet waar – een beetje waar – zeker waar). De items behoren bij vijf schalen: Emotionele problemen, Gedragsproblemen, Hyperactiviteit, Omgang met leeftijdgenoten en Prosociaal. De 20 items met betrekking tot de eerste vier genoemde schalen worden opgeteld tot een 'Totale moeilijkhedenscore'. Op basis van de normgegevens van Van Widenfeldt et al. (2003) zijn de ruwe scores omgezet in deviatiescores. De betrouwbaarheid van de totaalscore is goed (Cronbach's alpha = 0.81).

3.3 Procedure

Om te onderzoeken of het effect van de hulp vergelijkbaar is als de resultaten gemeten worden met de CBCL/6-18 dan wel de SDQ-O, is gebruik gemaakt van in de LDJ beschikbare afnames van deze vragenlijsten. De gegevens in de LDJ zijn ontdaan van herleidbare persoonsgegevens, zoals naam, geboortedatum en identificatienummers. De leeftijd (in jaren) bij de start van de hulp is wel bekend, evenals de start- en einddatum van de hulp. De cliënten zijn geïnformeerd en hebben geen bezwaar gemaakt tegen het opnemen van hun gegevens in de LDJ.

Omdat in de hulpverleningspraktijk zelden of nooit twee verschillende vragenlijsten met eenzelfde meetpretentie worden afgenomen bij dezelfde cliënt, hebben we twee groepen cliënten met elkaar vergeleken: Groep 1 bestond uit cliënten waarvan minimaal twee ingevulde CBCL/6-18 vragenlijsten beschikbaar waren (één bij aanvang en één bij einde van de hulp). Groep 2 bestond uit cliënten waarover minimaal twee SDQ-O lijsten waren ingevuld (eveneens één bij aanvang en één bij einde van de hulp). Om eventuele verschillen in behandeleffecten tussen de groepen goed te kunnen interpreteren is matching toegepast op alle in de LDJ bekende cliëntkenmerken (o.a. leeftijd en geslacht). Na matching zijn de Effect sizes en RCI's berekend voor beide groepen.

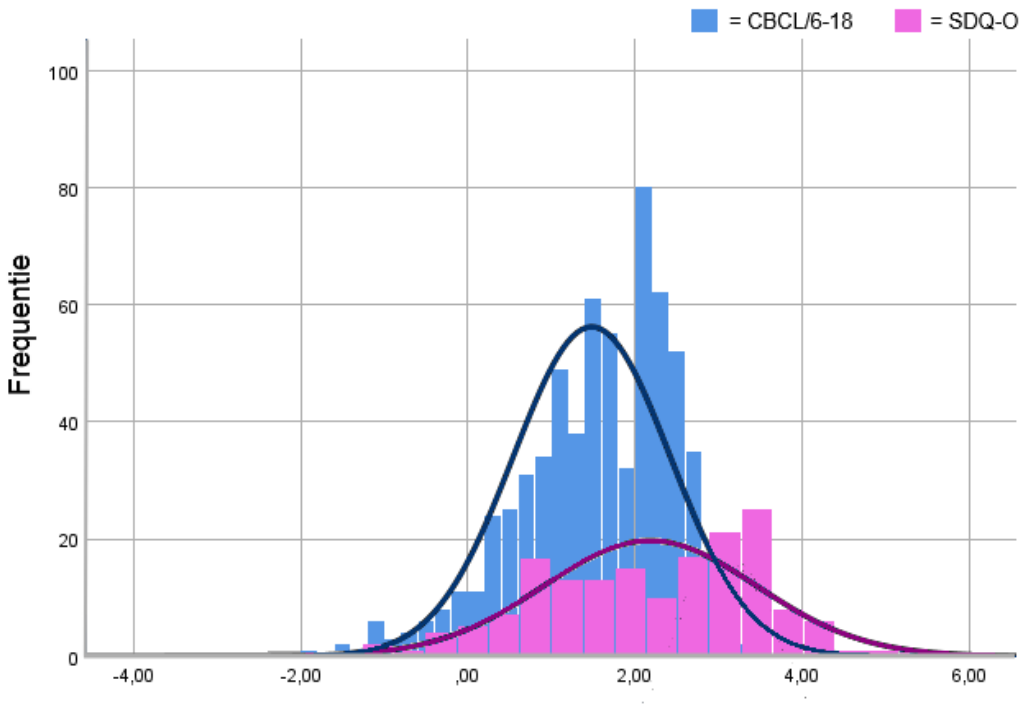
Dataselectie

Van alle op 22 juni 2021 aanwezige afnames van de CBCL/6-18 en SDQ-O, zijn de algemene gegevens (d.w.z., OnderzoeksID, InstellingsID, geslacht en leeftijd cliënt, informant, meetmoment, zorgproduct, startdatum en einddatum van de hulp), de ruwe totaalscore (SDQ-O en CBCL/6-18) en de T-score van het totaal (CBCL/6-18) vanuit de LDJ aangeleverd bij de onderzoekers. De vragenlijsten hebben het meetmoment 'aanvang' of 'einde' gekregen als deze in een periode van respectievelijk 3 maanden rondom de startdatum of 3 maanden rondom de einddatum van de behandeling waren ingevuld. Dit leverde 720 SDQ-effectmetingen en 1805 CBCL-effectmetingen op.

Bij deze 720 effectmetingen met de SDQ-O en 1805 effectmetingen met de CBCL/6-18 is nagegaan of er meerdere effectmetingen waren over dezelfde cliënt (ingevuld door verschillende informanten, bijvoorbeeld moeder en vader). Als dat het geval was, is per cliënt één effectmeting geselecteerd op basis van een 'informantenvolgorde', waarbij afnames van moedertypes als informant verkozen werden boven afnames van vadertypes, omdat moederafnames frequenter aanwezig waren in de database. Daarnaast werden alleen metingen meegenomen van kinderen die bij de eerste meting 5 jaar of ouder waren, omdat zowel de SDQ-O als de CBCL/6-18 vanaf deze leeftijd gescoord kan worden. Metingen behorende bij behandelingen die korter dan 12 weken en langer dan 100 weken duurden werden verwijderd, om het aantal registratiefouten en voortijdig afgebroken behandelingen in de dataset te minimaliseren. Voorgenoemde selecties resulteerden in 661 effectmetingen met de CBCL/6-18 en 167 met de SDQ-O.

Van de overgebleven effectmetingen is gecontroleerd of ze verschilden op de cliëntkenmerken die bekend waren in de LDJ. Voor het type hulp is niet gecorrigeerd, omdat alleen cliënten geselecteerd waren uit de zorgproductcategorie Jeugdhulp ambulant. In beide groepen was alleen sprake van specialistische hulpvormen, waarbij systeemgerichte hulp in beide groepen het meeste voorkwam. Voor de continue variabelen 'leeftijd bij aanvang van de hulp' en de 'duur van de behandeling' werd een t-toets voor onafhankelijke paren uitgevoerd op de opgeschoonde dataset. De duur van de hulp in weken verschilde significant tussen de groepen (Groep 1 (CBCL): M = 34,31, SD = 14,00; Groep 2 (SDQ): M = 37,88, SD= 19,17, $p < .01$). De leeftijd in jaren bij aanvang van de hulp verschilde niet significant (Groep 1 (CBCL): M = 11,27, SD = 3,48; Groep 2 (SDQ): M = 11,01, SD = 3,46, $p = .51$). Om mogelijke verschillen tussen de groepen in 'geslacht cliënt' en 'geslacht informanten' te toetsen, zijn chi-kwadraattoetsen uitgevoerd. Het geslacht van de cliënten (Groep 1 (CBCL): 63,3% man; Groep 2 (SDQ): 63,5% man, $p = .95$) en het geslacht van de informanten (Groep 1 (CBCL): 82,9 % moeders; Groep 2 (SDQ): 79,6 % moeders, $p = .17$) verschilden beide niet significant tussen de groepen.

Omdat het niveau van de problematiek bij aanvang invloed kan hebben op de gemeten verandering door de hulp, is ten slotte getoetst of het niveau van de problematiek bij aanvang tussen de twee groepen verschilde. Bij zowel de CBCL/6-18, als de SDQ-O worden scores boven het 90e percentiel aangemerkt als 'afwijkend' (SDQ-O, vanaf ruwe score 18, zie Goedhart et al., 2003), dan wel 'klinisch' (CBCL/6-18, vanaf T-score 63, zie Achenbach & Rescorla, 2001). Daarom is voor het vergelijken van het niveau van de problematiek bij aanvang het aantal cliënten met een score boven het 90e percentiel gehanteerd. Het was niet mogelijk om middels gemiddelde scores de problematiek bij aanvang te vergelijken. De normgegevens van de SDQ zijn namelijk gebaseerd op niet-normaal verdeelde ruwe scores, terwijl bij de CBCL T-scores worden berekend en gebruikt voor het kwalificeren van probleemgedrag. Omdat beide lijsten daarnaast behoorlijk verschillen wat betreft het aantal items (20 items (SDQ) versus 118 items (CBCL)) en specificiteit van de items (waarbij de vragen van de SDQ algemener geformuleerd zijn), is de verdeling van de scores dusdanig verschillend (zie Figuur 1) dat het vergelijken van gemiddelden geen goede manier is om te bepalen of de problematiek bij aanvang verschilt. Het niveau van de problematiek bij aanvang op basis van een score boven het 90e percentiel, verschilde significant tussen de twee groepen (Groep 1 (CBCL): 64,9%, Groep 2 (SDQ): 74,1%, p = .02).



Figuur 1. Scores bij aanvang van de hulp (omwille van de grafiek zijn de T-scores van de CBCL/6-18 omgezet in deviatiescores). Blauw = CBCL/6-18, roze = SDQ-O.

3.4 Statistische-analyses

Propensity Score Matching (PSM). Omdat de behandelduur en de problematiek bij aanvang verschilden tussen de twee groepen is matching uitgevoerd volgens de methode Propensity Score Matching (PSM; Rosenbaum & Rubin, 1983). PSM is een statistische matchingtechniek waarmee bij het analyseren van het effect van een behandeling rekening kan worden gehouden met variabelen die mogelijk samenhangen met het resultaat van de behandeling. De propensity score is de uitkomst van een multiple logistische regressieanalyse van een aantal variabelen en drukt de kans uit dat iemand op basis van deze variabelen tot bijvoorbeeld de experimentele of controlegroep behoort. In onze studie gaat het om de kans dat over een cliënt een CBCL of SDQ is ingevuld tijdens een ambulante behandeling. De cliënten die een klinische score en een niet klinische score op de CBCL hadden werden via de PSM gematcht met cliënten die respectievelijk een klinische score en een niet klinische score op de SDQ hadden, waarbij het criterium was dat de gekoppelde cliënten zo min mogelijk van elkaar mochten verschillen wat betreft propensity scores. Als cutoff point is hiervoor 0,02 gehanteerd: een kwart van een standaarddeviatie van de geschatte propensity scores (Guo et al., 2019; Rosenbaum & Rubin, 1985). Op deze manier werden twee gelijkaardige groepen geconstrueerd die evenveel kans hadden om een CBCL of SDQ te hebben ingevuld en die dus minder van elkaar verschilden wat betreft een aantal achtergrondvariabelen. De variabelen die zijn meegenomen in de regressieanalyse zijn geslacht en leeftijd van het kind, de behandelduur en het type informant (moeder, vader e.d.).

Deze matching resulteerde in twee vergelijkbare groepen van elk 160 cliënten. De groepen verschilden niet in leeftijd (in jaren) bij aanvang van de hulp (t(316) = ,16; p = .87), geslacht (χ2(2, N=320) = 1,31, p = .25), problematiek bij aanvang ((χ2(2, N=320) = 2,82, p = .24), duur van de hulp (t(310) = -,27; p = .79) en informant (χ2(2, N=320) = 0,17, p = .919). In Tabel 1 staan de kenmerken van beide onderzoeksgroepen.

Tabel 1. Gegevens onderzoeksgroepen

Instrument		Leeftijd (jaren)	Geslacht (jongens)	90e percentiel bij aanvang hulp	Duur hulp	Informant (moeder)
	N	*M (SD)*	%	%	*M (SD)*	%
CBCL/6-18	160	11,00 (3,66)	56,9	74,4	37,67 (15,35)	81,3
SDQ-O	160	11,06 (3,38)	63,8	74,4	37,16 (17,97)	81,9

Effect size

Met de gematchte dataset is vervolgens de Effect size (ES) berekend voor de twee groepen apart. Een ES is een maat om uit te drukken hoeveel eenheden van een standaarddeviatie een groep cliënten gedurende een behandeling veranderd is. De ES is in dit onderzoek berekend door de gemiddelde score op de nameting af te trekken van de gemiddelde score op de voormeting en dit verschil te delen door de gepoolde standaarddeviatie van de voor- en nameting. Een positieve ES duidt op verbetering, een ES om en nabij 0 duidt op geen verandering, een negatieve ES duidt op verslechtering. Het voordeel van de ES is dat deze niet zo gevoelig is voor de grootte van de onderzoeksgroep, dus ook bij kleine groepen bruikbaar is en over alle meetinstrumenten dezelfde betekenis heeft. Cohen (1992) hanteert de volgende betekenis voor de ES: een ES kleiner dan 0,20 is te verwaarlozen, een ES tussen 0,20 en 0,49 is een klein effect, tussen 0,50 en 0,79 is de ES middelgroot en boven de 0,80 is sprake van een grote ES (Cohen, 1992).

Reliable Change Index (RCI; Betrouwbare verandering).

Om te bepalen of er statistisch gezien bij een behandeling sprake is van een betrouwbare verandering hebben we de Reliable Change Index (Jacobson & Truax, 1991) berekend, dat is het verschil tussen de score op de voormeting en de score op de eindmeting gedeeld door de standaardfout van het verschil. De standaardfout van het verschil wordt bepaald door de standaardmeetfout alsmede de test-hertest betrouwbaarheid of interne consistentie van een vragenlijst. Statistisch gesproken wordt de RCI verondersteld normaal verdeeld te zijn met een gemiddelde van 0 en een standaarddeviatie van 1. De RCI's zijn hiermee getransformeerde z-waarden die getoetst kunnen worden. Als criterium voor "voldoende mate veranderd" hanteren we een z-waarde van 1,645. Bij een RCI groter of gelijk aan 1,645 mag er met 95% zekerheid van uitgegaan worden dat de verandering niet is toe te schrijven aan "toeval" of "ruis". De "ruis" is in dit geval de onbetrouwbaarheid van het meetinstrument. Het verschil tussen twee metingen moet in voldoende mate boven deze ruis uitkomen om te kunnen spreken van betrouwbare verandering.

4 Resultaten

Voor beide groepen zijn de ES (Cohen's d) en de RCI's berekend en de veranderingen getoetst met respectievelijk gepaarde t-toetsen en chi-kwadraat toetsen.

Effectgrootte

De t-tests laten significante verschillen zien tussen de voor- en nameting voor zowel de CBCL ($t(159) = 8{,}39$; $p < .001$) als de SDQ ($t(159) = 6{,}12$; $p < .001$). In Tabel 2 wordt de ES apart voor de CBCL en SDQ weergegeven.

Tabel 2. Veranderingen in Totale gedragsproblemen per instrument

Instrument		Aanvang		Einde				
	N	*M*	*SD*	*M*	*SD*	*p*	*Cohen's d*	*95% CI*
CBCL/6-18	160	1,52	0,74	0,93	0,95	.001	0.66	[0.51-0.86]
SDQ-O	160	2,19	1,29	1,57	1,39	.001	0.48	[0.31-0.63]

In Tabel 2 is te zien dat op beide instrumenten een significante verbetering wordt behaald, maar dat de ES behaald met de CBCL groter is dan de ES op basis van de SDQ. De ES behaald met de CBCL is middelgroot, de ES voor de SDQ is klein. Dit verschil wordt met name verklaard door de afwijkende standaarddeviaties van beide steekproeven (zie SDs Tabel 2): de spreiding bij de SDQ is op beide meetmomenten groter dan die bij de CBCL.

Om te toetsen of de veranderingen over tijd tussen de CBCL en SDQ van elkaar verschillen, is een MANOVA herhaalde metingen uitgevoerd met tijd (Aanvang en Einde) als within-subjects factor en groep (CBCL, SDQ) als between-subjects factor. We waren met name geïnteresseerd in een interactie-effect tussen tijd en groep, omdat dit weergeeft of de afname in probleemgedrag verschilt voor beide instrumenten. Dit blijkt niet het geval; het interactie-effect was niet significant ($F(1{,}318) = {,}075$, $p = .784$).

Reliable Change Index

Per cliënt is de RCI berekend en op basis van een cut-off criterium voor verandering (1,645) en de status van de cliënt bij de eindmeting (score onder of boven de subklinische grens) ingedeeld in vier effectcategorieën. In Tabel 3 staan voor de CBCL en de SDQ de aantallen met tussen haakjes de percentages behaald per effectcategorie.

Tabel 3. Verdeling effectcategorieën per instrument

Instrument/effectcategorie	-	0	+	++
CBCL/6-18	8 (5%)	93 (58%)	3 (2%)	56 (35%)
SDQ-O	11 (7%)	103 (64%)	13 (8%)	33 (21%)

Noot. - = significant verslechterd; 0 = niet significant veranderd; + = significant verbeterd, maar score in probleemgebied (90e percentiel); ++ significant verbeterd en score in normaal gebied.

De Chi-kwadraattest is significant: 2(3, N=320) = 13,178, p = .004. De gestandaardiseerde residuen wijzen uit dat de verschillen in de verdeling met name door de cellen + en ++ worden verklaard. De CBCL heeft verhoudingsgewijs meer cliënten in de categorie ++ en de SDQ verhoudingsgewijs meer cliënten in de categorie +. Wanneer we de categorieën + en ++ samennemen tot categorie + (significant verbeterd) dan is de Chi-kwadraat niet meer significant: χ^2(2, N=320) = 2,592, p = .273.

5 Conclusie en discussie

De aanleiding voor dit onderzoek was de vraag of ten behoeve van benchmarkdoeleinden de effecten die in de jeugdhulppraktijk gemeten worden met instrumenten die dezelfde meetpretentie hebben, zoals de CBCL/6-18 en de SDQ-O, vergelijkbaar zijn. Uit het onderzoek blijkt dat de resultaten op beide vragenlijsten op basis van de RCI nauwelijks van elkaar verschillen, met name wanneer drie effectcategorieën worden gehanteerd. Met de CBCL wordt weliswaar een grotere Effect size behaald dan met de SDQ, maar er is geen significant interactie-effect. Dit betekent dat de verandering in probleemgedrag niet significant verschilt tussen beide meetinstrumenten. Eerder gevonden overeenkomsten wat betreft meetpretentie, sensitiviteit en specificiteit van de CBCL en SDQ (Goedhart et al., 2003; Hendriks et al., 2020; Klasen et al., 2000; Vogels et al., 2005; Warnick et al., 2008) lijken ook te gelden voor de gevoeligheid van de instrumenten voor het meten van verandering. Het benchmarken met de CBCL en de SDQ samen is op basis van de gevonden uitkomsten dan ook te verantwoorden, maar wel met enkele nuances.

Het onderzoek geeft aanwijzingen dat de CBCL, waarschijnlijk door het grotere aantal items waarmee gedragsproblemen specifieker in kaart worden gebracht en het feit dat meer aspecten van probleemgedrag worden uitgevraagd, nauwkeuriger is in het meten van probleemgedrag. Dit vertaalt zich waarschijnlijk in een nauwkeuriger en daardoor gunstiger beeld wanneer deze vragenlijst wordt gebruikt om de behandelresultaten in kaart te brengen. Wanneer resultaten op de CBCL en SDQ met elkaar worden vergeleken, dient men zich er daarom van bewust te zijn dat een klein positief verschil in effect mogelijk toe te wijzen is aan het gebruikte instrument (CBCL), en niet aan de uitgevoerde behandeling.

In ons onderzoek is gebruik gemaakt van gegevens die zijn verzameld in de praktijk van de jeugdhulp. Daarbij zijn enkele beperkingen te benoemen. Allereerst is de groep cliënten in de dataset slechts een fractie van de cliënten die in zorg zijn. Doordat niet van iedere cliënt (voor- én na)metingen aanwezig zijn in de LDJ, kunnen de resultaten niet zonder meer gegeneraliseerd worden naar alle kinderen en jongeren die jeugdhulp ontvangen. Ten tweede is de gebruikte onderzoeksmethode (PSM) weliswaar een wetenschappelijk verantwoorde oplossing om twee groepen die in de praktijk van elkaar verschillen vergelijkbaar te maken, maar dit is methodologisch gezien niet de beste manier om de onderzoeksvraag te beantwoorden. In het ideale geval hadden we één groep van cliënten gehad die zowel de CBCL als SDQ had ingevuld bij aanvang en einde van de hulp. In dat geval was namelijk met meer zekerheid te zeggen dat de gevonden resultaten niet beïnvloed waren door cliënt- of behandelkenmerken. Een dergelijk within-subjects design is in de praktijk echter moeilijk te realiseren, omdat men ernaar streeft cliënten zo weinig mogelijk te belasten met extra vragenlijsten; Men zet bij voorkeur alleen

vragenlijsten in die ten dienste staan van de hulpverlening. Daarnaast spelen werkdruk en personeelstekorten een rol, waardoor organisaties kritisch zijn op waar hun tijd en personele inzet naartoe gaat. Met dit onderzoek hebben wij laten zien dat, ondanks dat de meest ideale onderzoeksmethode in de praktijk moeilijk haalbaar is, er toch wetenschappelijk verantwoorde manieren zijn om praktijkrelevante onderzoeksvragen te kunnen beantwoorden. Daarnaast laat het onderzoek zien dat gegevens die toch al verzameld worden in de dagelijkse praktijk van de hulpverlening ook op andere manieren benut kunnen worden om zicht te krijgen op de resultaten van jeugdhulp om daarvan te leren en de hulp zo nodig te kunnen verbeteren. Dit laat echter onverlet dat toekomstig onderzoek opnieuw zou moeten overwegen beide instrumenten bij eenzelfde groep cliënten af te nemen.

Om de beperkingen van ons onderzoek zoveel mogelijk te minimaliseren, hebben we de twee groepen (de cliënten over wie een CBCL of een SDQ is ingevuld) gematcht op alle bij ons bekende cliënt- en behandelfactoren, maar het is mogelijk dat er een andere samenhang is tussen de effectiviteit van de geboden hulp en de keuze voor een bepaalde vragenlijst. Daarbij valt bijvoorbeeld te denken aan het cognitieve niveau van de ouders. Het is denkbaar dat bij ouders met lagere cognitieve vermogens wordt gekozen voor de SDQ, omdat deze minder en eenvoudiger geformuleerde vragen bevat. Of te verwachten is dat de behandeling van cliënten met ouders met een lager cognitief functioneren ook minder effectief is, is onduidelijk. Nader onderzoek zal dit moeten uitwijzen. Een andere mogelijke beïnvloedende factor is toeval, waarbij bepaalde organisaties voor jeugdhulp die minder effectief zijn, vanwege het type behandeling of het type problematiek waarmee kinderen worden aangemeld, voor de SDQ hebben gekozen waardoor het effect dat wordt gemeten met deze lijst iets lager is. Aan de andere kant is aan de RCI’s te zien dat in beide groepen evenveel cliënten significante vooruitgang laten zien, wat erop wijst dat het waargenomen verschil is toe te wijzen aan het instrument in plaats van een mogelijk verband tussen de organisatie en de keuze van het instrument.

Dit onderzoek onderbouwt dat organisaties hun behandelresultaten, ongeacht of de CBCL/6-18 of de SDQ-O is gebruikt, met elkaar kunnen vergelijken om aanknopingspunten te vinden om de hulp te verbeteren. Daarbij valt te denken aan gesprekken over het effect van verschillende behandelvormen of werkzame elementen, over veranderingen in de behandelresultaten na bijvoorbeeld de Coronapandemie of na een stelselwijziging, over de behandeleffecten in relatie tot de problematiek bij aanvang van de behandeling, enzovoorts. De met dit onderzoek opgedane kennis draagt hierdoor bij aan een lerende beweging binnen de jeugdhulp.

5.1 Aanbevelingen

Wanneer de CBCL en de SDQ worden samengevoegd ten behoeve van benchmarking, is het van belang dat de juiste contextinformatie wordt gegeven. Het is aan te bevelen om in dashboards en rapportages aan te geven dat de behandeleffecten die gemeten worden met de CBCL waarschijnlijk een iets gunstiger beeld geven vanwege de eigenschappen van het instrument. Een belangrijke kanttekening hierbij is wel dat uit de praktijk blijkt dat de verschillen in behandeleffecten tussen organisaties doorgaans groter zijn dan de waargenomen verschillen tussen de instrumenten.

Wat betreft het meten van Totale problemen is het niet nodig dat organisaties die gezamenlijk willen benchmarken tezamen de keuze maken voor het ene dan wel het andere instrument, mits bij de interpretatie rekening wordt gehouden met bovengenoemde verschillen tussen beide instrumenten. Daarbij is een belangrijke kanttekening dat dit niet voor de verschillende schalen en subtotalen is onderzocht. Dit heeft als reden dat de betrouwbaarheid van bepaalde schalen van de SDQ niet consistent is gebleken uit eerder onderzoek (Theunissen et al., 2016). Dit heeft waarschijnlijk te maken met het lage aantal items per schaal. Voor jeugdhulporganisaties die onderscheid willen maken tussen externaliserend en internaliserend probleemgedrag, of die specifieke aspecten van probleemgedrag in kaart willen brengen, lijkt de CBCL meer geschikt dan de SDQ. Als richtlijn zouden organisaties in overweging kunnen nemen de SDQ te gebruiken voor screening en het meten van effecten van relatief lichte of kortdurende hulp, en bij langdurigere of intensievere hulp of hulp die zich richt op specifieke aspecten van gedrag alsnog de CBCL af te nemen.

Ook financiers aan wie de organisaties voor (jeugd)hulp verantwoording afleggen over de geboden hulp zouden kennis moeten nemen van eigenschappen van meetinstrumenten, zoals onderhavig onderzoek. Het is van belang voor duurzame kwaliteitsverbetering dat alle betrokken partijen inzicht hebben in de verzamelde gegevens en zich terdege bewust zijn van hetgeen wél en niet gemeten wordt om de gegevens juist te kunnen interpreteren. We hopen met dit onderzoek bij te dragen aan een onderbouwd en genuanceerd gesprek over behandelresultaten in de jeugdhulp en de jeugdhulporganisatie meer kennis en tools te geven om met gemeenten in gesprek te gaan over de resultaten van de geboden hulp.

Literatuurlijst

Achenbach, T. M. & Rescorla, L. A. (2001). *Manual for the ASEBA school-age forms & profiles.* Burlington, VT: University of Vermont, Research Center for Children, Youth & Families.

Algemene Rekenkamer (2017). *Bekostiging van de curatieve geestelijke gezondheidszorg.* Den Haag: Sdu.

Amble, I., Gude, T., Stubdal, S., Just Andersen, B. & Wampold, B. E. (2014). The effect of implementing the Outcome Questionnaire-45.2 feedback system in Norway: A multisite randomized clinical trial in a naturalistic setting. *Psychotherapy Research, 7,* 1-9.

BergOp. (z.d.). BergOp. https://www.bergop.info/_uploaded/werkveld/Zorgproducten-LDJ-mei21.pdf

Beurs, E. de (2017). *Naar aanleiding van het rapport van de Algemene Rekenkamer en STOPROM.* Bilthoven: Stichting Benchmark GGZ.

Beurs, E. de, Barendregt, M., Rogmans, B., Robbers, S., Geffen, M. van, Aggelen-Gerrits, M. van, & Houben, H. (2015). Denoting treatment outcome in child and adolescent psychiatry: a comparison of continuous and categorical outcomes. *European Journal of Child and Adolescent Psychiatry, 24(5),* 553-563.

Bovendeerd, B., Jong, K. de, Groot, E. de, Moerbeek, M. & Keijser, J. de (2021). Enhancing the effect of psychotherapy through systematic client feedback in outpatient mental healthcare: A cluster randomized trial. *Psychotherapy Research, 32(6),* 710-722.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Hillsdale, IN: Elbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155-159.

Geffen, M. van (2019). *Kwaliteitssystemen in de Nederlandse Geestelijke Gezondheidszorg: een analyse van de impact van kwaliteitssystemen op de praktijk van de zorg.* Proefschrift, Radboud Universiteit Nijmegen.

Goedhart, A., Treffers, F. & Widenfelt, B. van (2003). Vragen naar psychische problemen bij kinderen en adolescenten. De Strengths and Difficulties Questionnaire (SDQ). *Maandblad Geestelijke Volksgezondheid, 58,* 1018-1035.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire. A research note. *Journal of Child Psychology and Psychiatry, 38*, 581-586.

Guo, C., Xia, L., Mej, J., Li, C., Lin, F., Ma, L.Pu, Q. & Liu, L. (2019). A propensity score matching study of non-grasping *en bloc* mediastinal lymph node dissection versus traditional grasping mediastinal lymph node dissection for non-small cell lung cancer by video-assisted thoracic surgery. *Translational Lung Cancer Research, 8(2),* 176-186.

Hendriks, A. M., Hill, F. Ip, Nivard, M. G., Finkenauer, C., Beijsterveldt, C. E. M. van, Bartels, M. & Boomsma, D. I. (2020). Content, diagnostic, correlational, and genetic similarities between common measures of childhood aggressive behaviors and related psychiatric traits. *Journal of Child Psychology and Psychiatry, 61(12),* 1328-1338

Hervormingsagenda Jeugd 2023-2028. (2023). Verkregen op 20 december 2023 van: https://open.overheid.nl/documenten/addec5d5-279c-40de-b607-7b64e8441602/file

Jacobson, N. S. & Truax P. (1991) Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 12-19.

Janse, P. D., Jong, K. de, Veerkamp, C., Dijk, M. K. van, Hutschemaekers, G. J. M. & Verbraak, M. J. P.M. (2020). The effect of feedback-informed cognitive behavioral therapy on treatment outcome: A randomized controlled trial. *Journal of Consulting and Clinical Psychology, 88(9),* 818-828.

Klasen, H., Woerner, W., Wolke, D., Meyer, R., Overmeyer, S., Kaschnitz, W., Rothenberger, A. & Goodman, R. (2000). Comparing the German versions of the Strengths and Difficulties Questionnaire (SDQ-Deu) and the Child Behavior Checklist. *European Child & Adolescent Psychiatry, 9(4),* 271-276

Lambert, M. J., Whipple, J. L., Smart D.W., Vermeersch, D. A., Lars Nielsen, S, Hawkins, E. J. (2001). The effects of providing therapists with feedback on patient progress during psychotherapy: are outcomes enhanced? *Psychotherapy Research, 11(1),* 49-68.

Rosenbaum, P. R. & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrica, 70(1),* 41-55.

Rosenbaum, P. R. & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician, 39(1),* 33-38.

SEJN (2020). *De Lerende Databank Jeugd. Betere benutting van gegevens door gebruik van een feedbacksysteem.* Eindhoven: SEJN.

Theunissen, M.H.C., Wolff, M. de, Grieken, A. van, & Mieloo, C. (2016*). Handleiding voor het gebruik van de SDQ binnen de Jeugdgezondheidszorg. Vragenlijst voor het signaleren van psychosociale problemen bij 3-17 jarigen.* Leiden: TNO.

Verhulst, F. & Ende, J. van der (2013). *Handleiding ASEBA. Vragenlijsten voor leeftijden 6 t/m 18 jaar.* Rotterdam: ASEBA Nederland.

Vogels, A. G. C., Crone, M. R., Hoekstra, F. & Reijneveld, S. A. (2005). *Drie vragenlijsten voor het opsporen van psychosociale problemen bij kinderen van zeven tot twaalf jaar.* TNO Rapport KvL/JPB 2005.082. Leiden: TNO.

Warnick, E. M., Bracken, M. B. & Kasl, S. (2008). Screening efficiency of the Child Behavior Checklist and Strengths and Difficulties Questionnaire: A systematic review. *Child and Adolescent Mental Health, 13(3),* 140-147.

Widenfelt, B. M. van, Goedhart, A. W., Treffers, P.D.A. & Goodman, R. (2003). Dutch version of the Strengths and Difficulties Questionnaire (SDQ). *European Child and Adolescent Psychiatry, 12,* 281-289.

Yperen, T. van (2013). *Met kennis oogsten. Monitoring en doorontwikkeling van een integrale zorg voor jeugd.* Utrecht/Groningen: Nederlands Jeugdinstituut/Rijksuniversiteit Groningen.

Colofon

Praktikon B.V. is een zelfstandige, onafhankelijke organisatie voor onderzoek en ontwikkeling. We werken voor (jeugd)zorg, onderwijs en gemeenten en zetten ons in voor het verbeteren van de kwaliteit en effectiviteit in deze sectoren. Onze aanpak kenmerkt zich door persoonlijke aandacht, praktijkgerichtheid en maatwerk.

Het Praktikon-team bestaat uit gedreven professionals met voornamelijk achtergronden in de sociale wetenschappen, zoals psychologie en orthopedagogiek.

Als Praktikon vinden we bundelen van krachten en delen van kennis van groot belang. Zo werken we samen met diverse (jeugd)zorg- en onderwijsinstellingen, gemeenten, cliënten- en onderzoeksorganisaties, universiteiten, hogescholen, academische werkplaatsen en landelijke kenniscentra en samenwerkingsverbanden.

Contactgegevens

Praktikon B.V.
Postbus 6909
6503 GK Nijmegen
www.praktikon.nl

info@praktikon.nl
tel. 024-3615480
fax. 024-3611152

praktikon