Wetenschappelijk Onderzoek- en
Documentatiecentrum
*Ministerie van Veiligheid en Justitie*

# Verslag expertmeeting "Designs van effectstudies in justitiële contexten"

Datum 7 maart 2013

# Colofon


| | |
|---|---|
| Afzendgegevens | **Afdeling Extern Wetenschappelijke Betrekkingen (EWB)** |
| | Turfmarkt 147<br>2511 DP Den Haag<br>Postbus 20301<br>2500 EH Den Haag<br>www.wodc.nl |
| Contactpersoon | Dr. J. Mulder<br>*Onderzoekscoördinator* |
| | T 070 370 65 61<br>F 070 370 79 48 |
| Projectnaam | Expertmeeting effectstudies |
| Ons kenmerk | 2299 |
| Auteurs | J. Mulder<br>A. Daalder<br>F.L. Leeuw |

# Inhoud

# Inleiding

Op 19 oktober 2012 heeft het Wetenschappelijk Onderzoek en Documentatie Centrum (WODC) van het ministerie van Veiligheid en Justitie een expertmeeting georganiseerd met als thema "designs van effectstudies in justitiële contexten". Het doel van de bijeenkomst was een discussie te voeren over designs op grond waarvan in meer of mindere mate uitspraken over effectiviteit te doen zijn. De uitgenodigde experts waren afkomstig uit diverse wetenschappelijke disciplines (zie hieronder, alleen de voornaamste affiliaties worden genoemd).

| | *Discipline* | *Verbonden aan* |
|---|---|---|
| prof. dr. G.H. de Bock | Oncologische Epidemiologie | RuG |
| dr. J. Boom | Ontwikkelingspsychologie | UU |
| dr. B.H. Bulten | Hoofd diagnostiek, onderzoek en opleidingen Forensiche Zorg | Pompestichting, RU |
| prof. dr. M. Deković | Orthopedagogiek | UU |
| prof. dr. E. (Eelko) Hak | Klinische Farmacoepidemiologie | RuG |
| dr. M.W.J. Koeter | Methodologie Psychiatrie en Verslaving | AIAR, UvA |
| prof. dr. Niels C.L. Mulder | Openbare Geestelijke Gezondheidszorg, Acute Psychiatrie, Bemoeizorg | EUR |
| prof. dr. M.M. Rovers | Evidence-Based Chirurgie | UMC St Radboud |
| Prof. dr. G.J.J.M. Stams | Forensische Orthopedagogiek | UvA |
| Prof. dr. J.W. Veerman | Speciale Kinder- en Jeugdzorg | Praktikon, RU |
| Prof. dr. R.R.J.M. Vermeiren | Forensische Jeugd Psychiatrie | LUMC, VU |
| dr. D. Webbink | Beleidsevaluaties Econometrie | EUR |
| Prof. dr. T. van Yperen | Monitoring en Innovatie Zorg voor Jeugd | RuG |
| Drs. A. Daalder | Beleidsevaluaties | WODC |
| Prof. dr. Frans Leeuw | Recht, Openbaar Bestuur, Beleidsevaluaties | WODC, MU |
| Dr. J. Mulder | Evaluaties van gedragsinterventies | WODC |
| Dr. B. Wartna | Recidive Monitor | WODC |

De aanleiding voor het organiseren van deze expertmeeting was de veelheid aan omstandigheden waaronder effectevaluaties in justitiële contexten plaatsvinden, waardoor niet steeds dezelfde eisen aan het effectstudie-design gesteld kunnen worden. Zo wordt dezelfde training soms zowel intra- als extramuraal gegeven. Het minimaal aantal deelnemers voor een interventie is vaak lastig te bereiken, of een training is voor zoveel deelnemers geschikt en wordt ook door zoveel deelnemers gevolgd dat het vormen van een controlegroep lastig is. Sommige

interventies daarentegen zijn alleen geschikt voor een zeer klein aantal justitiabelen, wat het werken met controlegroepen dan weer bemoeilijkt.

Om tot uitspraken over effecten van een gedragsinterventie te komen, moet onderzoek aan bepaalde eisen voldoen. Deze eisen dienen het eenduidig kunnen toeschrijven van de gemeten veranderingen aan de ingezette gedragsinterventie. Er zijn meerdere designs voorhanden, afhankelijk van de intensiteit van de interventie en het aantal deelnemers. Het meest geroemde design is het *randomised controlled trial* design, ofwel het RCT-design. Sommige onderzoekers nemen geen genoegen met een ander design om van een effectstudie te mogen spreken, anderen maken gebruik van alternatieven. Het quasi-experimentele design wordt in brede kring als werkbaar alternatief beschouwd. Variaties op dit design en andere alternatieven voor het RCT-design zullen in dit stuk worden besproken. Ook wordt ingegaan op factoren die van invloed zijn op mogelijkheden in onderzoek naar de werking en effectiviteit van gedragsinterventies. Geëindigd wordt met een discussie over *what works* in effectonderzoek en wat reëel lijkt in het huidige en toekomstige evaluatieonderzoek.

## Hiërarchie van designs

Binnen Criminologie en gerelateerde disciplines wordt een hiërarchie gehanteerd waarin studies worden gerangschikt naar de mate waarin gesproken mag worden van effecten, ofwel sterkte van het design. Deze hiërarchie heet de *Maryland Scale of Scientific Methods* (MSSM; Sherman et al., 1998). Daarin worden de volgende niveaus onderscheiden:

Niveau 1. Correlatie tussen een interventie ter preventie van criminaliteit en een maat van criminaliteit of criminogene risicofactoren op een enkel moment.
Niveau 2. Temporele volgorde tussen het programma en de criminaliteit of risico uitkomsten duidelijk aangetoond, of de aanwezigheid van een vergelijkingsgroep zonder aangetoonde vergelijkbaarheid met de behandelde groep.
Niveau 3. Een vergelijking tussen experimentele en controlegroep.
Niveau 4. Vergelijking tussen controle en experimentele groep, waarbij gecontroleerd wordt voor andere factoren dan het programma zelf, of gebruikmakend van deelnemers aan beide groepen die onderling slechts kleine verschillen laten zien.
Niveau 5. Random toewijzing en analyse van vergelijkbare eenheden van experimentele en controlegroep.

In de medische wetenschappen worden meerdere onderscheidingen gemaakt, waaronder één die zeven niveaus hanteert:

1. case-series. In geval van zeldzame en chronische ziektes. Een meta-analyse van *case studies* kan dan tot bredere uitspraken leiden dan de individuele cases op zich.
2. patiënt of *case control study*. De case control study, ofwel n=1 studie is een variatie op de pragmatische *trial* (zie niveau 6), die soms verklarend is en waarbij naar de werking van het medicament wordt gekeken. Die variant is alleen te gebruiken bij chronische patiënten, omdat er een min of meer stabiele situatie binnen de patiënt moet zijn om te kunnen wisselen tussen verschillende interventies of controles. Dat geldt ook voor *cross-over trials*.
3. retrospectieve studies op basis van eerder geregistreerde data.
4. prospectieve cohort-studies. In een cohortonderzoek worden personen die al dan niet blootgesteld zijn aan een risicofactor (zoals een schadelijke stof of een leefstijlfactor) gedurende lange tijd (meestal jaren) opgevolgd. De onderzochte populatie moet bij aanvang vrij zijn van de te onderzoeken uitkomst, zodat op deze wijze de incidentie van de uitkomst in de groep met blootstelling en de groep zonder blootstelling kan worden berekend.
5. geclusterde *randomised controlled trial* (randomiseren binnen huisartsenpraktijken, apotheken, regio's, arrondissementen, et cetera).
6. pragmatische trial, waarbij de selectie minder streng is dan bij een RCT en waarbij niet een placebo, maar een gangbare behandeling wordt gesteld tegenover de experimentele conditie. Ook is hierbij meer sprake van openheid over wie welke behandeling krijgt (in tegenstelling tot de double-blind randvoorwaarde).

7. de verklarende trial waarbij sprake is van double-blind werken, placebo gecontroleerde laboratoriumomstandigheden en random toewijzing aan beide condities.

## Alternatieve designs

De onderscheiden niveaus in de hiërarchieën hierboven zijn in principe helder, maar in de praktijk zijn er vele variaties op de beschreven designs mogelijk. In bijlage I staan voorbeelden van designs beschreven die tot nu toe in WODC onderzoek gebruikt zijn om te komen tot uitspraken over de effectiviteit van justitiële interventies. Kort beschreven zijn de volgende designs gebruikt in deze voorbeelden (geen uitputtende lijst):

- *Double-blind, placebo-controlled field experiment* (één studie);
- Quasi-experimenteel design met één of meer controlegroepen, al dan niet met behulp van *propensity score matching* samengesteld (vier studies);
- Geaccelereerd cohort model (één studie);
- Gestapeld n=1 design (één studie).

Tijdens de expertmeeting zijn nog meer alternatieven voor het RCT-design besproken. Deze worden hieronder kort samengevat. Dit overzicht pretendeert niet volledig te zijn.

*Natuurlijke experimenten*
Er doen zich soms situaties voor, waardoor zonder vooropgezet onderzoeksplan zich omstandigheden voordoen waarbij de invloed van twee condities getoetst kan worden. Een voorbeeld is de relatie tussen financiële ondersteuning van studenten en hun presteren. Het is lastig uitleggen aan studenten dat een deel van hen wel financiële ondersteuning krijgt en een deel niet, en dat ze random worden toegewezen aan wel of geen ondersteuning, ten behoeve van het onderzoeksdesign. Er zal dan eerder worden gekozen voor het geven van ondersteuning aan alle studenten. Wanneer die ondersteuning wegvalt als gevolg van bezuinigingen, kan toch getoetst worden wat het effect was van die ondersteuning. Dit is een vorm van een natuurlijke experiment.

*Zelen design*
In het Zelen design (Zelen, 1979; Homer, 2002) worden deelnemers *random* toegewezen aan ofwel een experimentele of een controle conditie. Doorgaans worden deelnemers voorafgaand aan de random toewijzing op de hoogte gesteld van de studie en de twee condities. In het Zelen design wordt pas na randomisering uitleg gegeven aan deelnemers. Dit design omzeilt het probleem dat op kan treden bij het geven van *informed consent* voorafgaand aan het randomiseren, waarbij deelnemers wellicht het risico niet willen nemen om in de controleconditie (in geval van een veelbelovende nieuwe interventie) of juist de experimentele conditie (in geval van een goed bekend staande controleconditie) terecht te komen. Deelnemers zouden bijvoorbeeld een controleconditie prefereren, wanneer de bestaande behandelvorm in het verleden reeds goede resultaten heeft geboekt, terwijl de experimentele, nieuwe behandelvorm nog niet bewezen effectief is.

In geval van bijvoorbeeld zeldzame, levensbedreigende ziektes kan het ethischer zijn om ouders niet voor de randomisatie te vertellen dat er een kans is dat hun kind behandeld zal worden met een nieuwe methode, terwijl dat misschien niet door kan gaan. Een ander voorbeeld waarin het

Zelen design gebruikt is, betreft het onderzoek naar het belang van een nieuwe vorm van *screening* in de detectie van een bepaalde vorm van darmkanker (Hardcastle et al., 1996). Het onderzoek werd gedaan onder alle 45-75 jarigen in een regio in Engeland. Het was belangrijk dat er geen vorm van selectie zou optreden in het willen meewerken aan de *screening*. Leden van de controlegroep werd daarom niet verteld van de studie en zij kregen geen interventie via deze weg. De nieuwe vorm van screening werd gedurende tien jaar elke twee jaar herhaald. Leden van de controlegroep werden elke twee jaar opgeroepen voor de standaard vorm van screening, zodat zij niet zouden merken dat zij in de controlegroep zaten. In totaal werkten meer dan 150.000 personen mee aan de studie. De auteurs spreken van een *unselected population-based randomised controlled trial*. De auteurs concludeerden dat de screening bijdroeg aan de reductie van deze vorm van darmkanker. Hadden mensen kunnen kiezen voor het meedoen aan de deze vorm van *screening*, dan hadden wellicht maskerende selectie-effecten kunnen optreden.

Ook het Zelen design is gevoelig voor de weigering van een behandeling. Daarnaast kan het onuitvoerbaar zijn wanneer er sprake is van erg opvallende vormen van dataverzameling of behandeling. Het kan zijn dat er dan geen gebruik gemaakt kan worden van sterk definiërende in- en exclusiecriteria. Beide factoren leiden ertoe dat hoge aantallen nodig zijn voor de groepen controle- en experimentele conditie (Torgerson & Roland, 1998). In de medische wereld wordt vaak gewerkt met een *2-op-1 randomisatie*. De nieuwe interventie wordt dan aan tweemaal zoveel mensen aangeboden als de groep die *care as usual* krijgt.

*Stepped wedge design*
*Stepped wedge trials* zijn gerandomiseerde trials waarin elke deelnemer eerst de controle conditie krijgt en later de experimentele interventie. Alleen het moment waarop de te toetsen interventie wordt gegeven, wordt gerandomiseerd. Randomisatie kan op individueel of clusterniveau. Dit design is vooral nuttig wanneer het niet haalbaar is om de interventie aan iedereen tegelijkertijd te geven, en voor de evaluatie van interventies waarvan de effectiviteit in gelimiteerde, onderzoeksgestuurde omstandigheden al was aangetoond, en nu moet worden aangetoond op grotere schaal. Het design is ook nuttig voor het evalueren van temporele veranderingen in het interventie effect (Hughes, 2008).

Een voorbeeld van het gebruik van dit design is een onderzoek naar de toepassing van een supervisie programma voor ex-gedetineerden in hun proeftijd (voorwaardelijke invrijheidsstelling) (Pearson et al., 2010). In dit onderzoek werden drie regio's met elkaar vergeleken. Regio A waarin het programma al was geïmplementeerd voor er over een experimentele studie werd gesproken. Regio B waar men voor een experimentele opzet koos, maar waar sommige locaties beter geschikt waren voor implementatie dan andere. Daarom werd daar besloten tot een gefaseerde invoering. Regio C zag geen brood in een gefaseerde implementatie en besloot om het programma te implementeren op alle locaties tegelijkertijd. Gedurende de implementatie in regio C bleek dat het niet haalbaar was om het voltallige personeel op tijd te trainen en werd er alsnog besloten tot een gefaseerde invoering. In regio A was veel geïnvesteerd in de invoering van het programma en het trainen van het personeel, waardoor de uiteindelijke implementatie als erg goed werd beoordeeld. In regio B werden locaties

zorgvuldig één voor één getraind, opdat er geen contaminatie kon ontstaan tussen locaties van de experimentele en controlelocaties. In deze regio verliep zowel het onderzoek als de implementatie het best. De uitkomsten van het onderzoek waren nog niet bekend ten tijde van de publicatie. Volgens de auteurs heeft het stepped wedge design het voordeel dat de implementatie van een grootschalig programma beter uitvoerbaar is, zonder verlies van statistische integriteit.

Problemen die bij het *stepped wedge design* kunnen spelen, zijn bijvoorbeeld verschillen tussen deelnemende instellingen en maatschappelijke veranderingen zoals bezuinigingen of ambulantisering waardoor er sprake kan zijn van veranderende populaties. Een ander probleem bij een design waarbij mensen op de wachtlijst staan voor de interventie, is dat er geen lange follow-up periodes mogelijk zijn. Mensen die toe zijn aan een behandeling kun je niet een jaar lang op de wachtlijst laten staan. Dan is er dus maar een korte follow-up periode mogelijk. Een antwoord op de vraag of effecten beklijven, wordt dan lastig. Hetzelfde probleem kan zich echter ook voordoen bij RCT's. Bij deelnemers aan RCT's kan ook sprake zijn van tussentijdse interventies in de follow-up periode. Gesteld wordt dat het probleem econometrisch oplosbaar is door de instrumentele variabelen aanpak (IV).

*Instrumental variables (IV-)design*
Het *instrumental variables* (IV-) design is een methode afkomstig uit de econometrie, waarin de effecten van verborgen bias in observationele studies worden opgespoord (Earle et al., 2001; Stukel et al., 2007). Een instrumentele variabele is sterk verbonden met de waarschijnlijkheid dat iemand een behandeling krijgt, en niet direct verbonden met de uitkomst van de behandeling. De assumptie is dat een valide IV alleen via de interventie effect heeft op de uitkomstmaat en niet correleert met andere ongemeten *confounders* (Angrist & Pischke, 2009; Bushway & Apel, 2010). Deze 'ongecorreleerdheid' kan alleen op theoretische gronden aannemelijk worden gemaakt en is statistisch niet toetsbaar (Posner et al., 2002). De IV-benadering houdt rekening met ongemeten selectie-effecten en reduceert '*residual confounding'* en is daarom een aantrekkelijke methode om effecten van interventies te evalueren (Posner et al., 2002). In de praktijk is het echter een zeer lastige klus om een valide IV te identificeren en te meten die aan de genoemde criteria voldoet. Daarnaast kan deze benadering, in tegenstelling tot een RCT, alternatieve verklaringen voor de gevonden effecten niet uitsluiten (Sherman, 2010).

De IV-benadering is afkomstig uit de econo(metr)ische literatuur (zie (Angrist & Pischke, 2009) voor een overzicht) en is midden jaren negentig door Levitt in de criminologie geïntroduceerd om de effecten van gevangenisstraf op (geregistreerde) criminaliteit te beschrijven (Levitt, 1996). In Nederland heeft Vollaard (Vollaard, 2010) een IV-benadering gebruikt om de effecten van de SOV/ISD-maatregel op aangiften van woninginbraak en autodiefstal te onderzoeken. Als instrument gebruikte hij de bestuurlijke perikelen rondom de invoering van de SOV/ISD-maatregel. Hierdoor ontstonden door natuurlijke variatie gemeenten die enkel van elkaar verschilden doordat ze de SOV/ISD-maatregel voor veelplegers op een ander moment in de tijd invoerden. Het verschil in aangifte van woninginbraak en autodiefstal tussen de gemeenten, schreef Vollaard toe aan het insluitingseffect van veelplegers in de SOV/ISD-maatregel. In

hoeverre de SOV/ISD-maatregel effect heeft op de recidive van de deelnemers is niet onderzocht.

*Regression discontinuity (RD-)design*
Hierbij worden deelnemers aan groepen toegewezen op basis van een cutoff-score op een bepaalde variabele bij een voormeting (e.g., Van Loon, Van der Meulen & Minnaert, 2011). Zo kunnen bijvoorbeeld de mensen die het meest behoefte hebben aan een interventie in de experimentele groep geplaatst worden. Voor het RD-design zijn een criteriumwaarde en een heldere aanwijzingsregel nodig. Dat is wenselijk, omdat de mensen die als eerste instromen wel eens heel anders kunnen zijn dan de mensen in de buurt van de cutoff. Het verschil in groepen door tussentijdse ontwikkelingen doet zich veelvuldig in de jeugdzorg voor. Als de aanwijzingsregel bekend is (bijvoorbeeld 'met 200 leerlingen zit de wijkschool vol'), is het ook mogelijk voor de instroomregel te corrigeren. Men kan dan inzoomen op de laatste 50 van de eerste 200 en de eerste 50 van de volgende 200.

*Clusterrandomisatie*
Een ander alternatief voor klassieke randomisering is het randomiseren op het niveau van clusters. Deze methode heeft als nadeel dat er veel respondenten nodig zijn. Afhankelijk van de grootte van het cluster zijn zes tot zevenmaal zoveel respondenten nodig. Een voorbeeld van cluster randomisatie is het random toewijzen van deelname aan een programma ter preventie van geweld op scholen dat ondermeer inzet op het veranderen van de omgangsvormen tussen leerlingen onderling, leerlingen en docenten, en docenten onderling. Elke school geldt dan als een cluster. Bij cluster-randomisatie moet goed gelet worden op *recruitment bias*.

*Vergelijkende studies*
In de bemoeizorg heeft men ervaring met het werken met twee verschillende methoden die allebei erkend zijn. Zo kan een instelling vanwege nieuw beleid methode B invoeren en methode A handhaven. Te verdedigen is dat wordt gekozen voor beide methoden, omdat onbekend is welke van de twee beter is. Na prerandomisering worden respondenten via ROM gevolgd. Dan wordt op een gerandomiseerde wijze informatie over de effecten verkregen. Dit is te verdedigen op basis van een goede argumentatie richting de medisch-ethische commissie. Immers, sommige zaken zijn niet te onderzoeken in een (acute) omgeving waarbij mensen van tevoren om *informed consent* moet worden gevraagd.

## Kenmerken van de justitiële setting van invloed op design

Zoals uit bovenstaande beschrijving van mogelijke onderzoeksdesigns al blijkt, zijn er verschillende scenario's mogelijk voor effectstudies. Tijdens de expertmeeting werd duidelijk dat onderzoekers bij effectonderzoek in de verschillende vertegenwoordigde disciplines tegen dezelfde problemen aanlopen. Hieronder worden zij kort samengevat.

*Random toewijzing*
Tot nu toe is er binnen het justitiële werkveld van uitvoerders, rechters en beleidsmakers flinke weerstand geweest tegen randomiseren. Soms gaat het om ethische problemen, soms spelen juridische en praktische bezwaren een rol. Bij medische interventies, onder andere chirurgische ingrepen, is het randomiseren ook niet altijd mogelijk of wenselijk (Barkun, Aronson, Feldman, et al., 2009). De laatste tijd verschijnen er Nederlandse publicaties waarin het randomiseren in enkele gevallen wel is gelukt in het justitiële veld. Volgens Bruinsma en Weisburd (2007) is er sinds de eeuwwisseling sprake van een toename in experimenteel onderzoek binnen criminologie. James, Asscher, Dekovic, Van der Laan en Stams (2012) beschrijven bijvoorbeeld hoe randomiseren zeer moeizaam is gelukt in de context van nazorg voor jeugdige gedetineerden. Ook een WODC-onderzoek naar de effecten van voedingssupplementen op gedrag van gedetineerden was vormgegeven als een *randomized field experiment* (Zaalberg et al., 2009). Dit onderzoek wordt in de bijlage beschreven. Een complicerende factor in ander onderzoek is dat de meeste interventies als maatregel door de rechter worden opgelegd. In het buitenland (bijvoorbeeld de VS, zie e.g., Farrington & Welsh, 2005) werken rechters vaker mee aan randomisatieprocedures, maar in Nederland is dit zelden het geval. Het kan zijn dat het 'slechts' een kwestie is van voorlichten, samenwerken en tijd om dit ook in Nederland gedaan te krijgen.

*Aantal deelnemers*
Een ander veel voorkomend probleem is het bereiken van het benodigd aantal deelnemers aan zowel de conditie experimentele als controlegroep, zodat eventuele werkelijke verschillen ook (statistisch) aangetoond kunnen worden. In geval van sommige interventies kunnen wel voldoende aantallen gehaald worden voor de experimentele groep, maar dan is het vinden van deelnemers aan de controlegroep een probleem. Er is vaak sprake van landelijke uitrol van gedragsinterventies, waardoor iedereen die voor de interventie in aanmerking komt, deze ook krijgt. In andere gevallen is een interventie voor een zeer specifieke groep bedoeld, waardoor de populatie om een controlegroep uit te halen ook zeer klein is. Een voorbeeld is een interventie voor gedetineerde jongeren met een zedenproblematiek. Het halen van voldoende aantallen is op de lange termijn wellicht wel mogelijk met een verlengde onderzoeksperiode. Vaak past dit niet binnen het beschikbare onderzoeksbudget, of binnen de termijnen die gehanteerd worden in het kader van erkenning verleend door de Erkenningscommissie.

*Follow-up*
Hieruit voortvloeiend is het doen van follow-up onderzoek bij deelnemers complex. Vaak is er al sprake van het verlopen van de periode waarbinnen een justitiële maatregel geldt, waardoor deelnemers moeilijk terug te

vinden zijn. Er kan wel toestemming gevraagd worden aan deelnemers om mee te doen aan follow-up onderzoek, maar dat resulteert in sterke selectie-effecten in de groep die daaraan mee wil doen. Alleen indien de deelnemers nog in contact staan met justitie, reclassering of een andere ketenpartner is het terugvinden van eerdere deelnemers mogelijk, waardoor een herhalingsmeting voldoende non-biased, uitvoerbaar, kostenverantwoord en daarmee realistisch kan zijn. Dan nog is het lastig om de effecten van een gedragsinterventie te verbinden met recidive-uitkomsten een paar jaar later (meestal minstens twee jaar). Hoewel zeer wenselijk voor beleid, is het voor onderzoek vrijwel onmogelijk om rekening te houden met wat er in de tussentijd gebeurt. Het is daarom zeer belangrijk om (naast recidive) goede uitkomstmaten in de nabije toekomst van de interventie te kiezen, zodat de interventies een 'fair trial' krijgen.

*Geen testfase*
Een verschil tussen de vertegenwoordigde disciplines tijdens de expertmeeting betrof de mate waarin van een testfase gebruik wordt gemaakt. In bijvoorbeeld farmacologische studies is de testfase van nieuwe medicijnen niet weg te denken. In geval van (justitiële) gedragsinterventies wordt zelden de tijd genomen voor de testfase. Het onderzoek naar gedragsinterventies is sinds het programma Terugdringen van Recidive (2002) hand in hand gegaan met het beleid om te gaan werken met *evidence-based* interventies. Vanaf die tijd zijn interventies beter uitgewerkt op papier. Soms betrof het interventies die voortbouwen op eerdere, bestaande interventies, soms nieuwe interventies die gebaseerd zijn op interventies uit de VS (en die daar bewezen effectief zijn). Na de planevaluatie door de Erkenningscommissie was het aan het WODC of de instellingen zelf om procesevaluaties uit te voeren, gevolgd door effectevaluaties. Zoals in de onderzoeken in Bijlage I te zien is, werd ingezet op het evalueren van interventies *alsof zij al waren uitontwikkeld*. Achteraf gezien is er geen tijd genomen voor de exploratieve testfase, waarin uitgezocht kan worden voor wie de interventie bedoeld is, wat essentiële onderdelen in de uitvoering zijn, welke ruimte er is voor afstemming op de specifieke behoeften van de groep of het individu (responsiviteit). Ook de ontwikkeling en validering van instrumenten die de programmadoelen beogen te meten (sensitiviteit, betrouwbaarheid, validiteit) hebben relatief weinig tijd gekregen. Het overslaan van een dergelijke testfase heeft het onderzoek gaandeweg parten gespeeld (ondermeer onduidelijkheid over waaraan gevonden uitkomsten toe te schrijven zijn).

Als voorbeelden (zie verder bijlage I) kan geconstateerd worden dat Halt en Stay-in Love bij implementatie nog niet goed uitontwikkeld of getest waren op de werkzame mechanismen. Gedurende het onderzoek bleken deze interventies nog in ontwikkeling te zijn. Toch werden zij al onderworpen aan experimenteel onderzoek. Er kwamen geen sterke resultaten uit naar voren, wat kan komen door de premature inzet van het experimentele onderzoek. In geval van Halt bleek dat de interventie bij sommige subgroepen wel werkte. Dit zijn belangrijke uitkomsten die eigenlijk in een testfase naar voren moeten komen. In principe is er zeker een aantal jaren nodig voor de ontwikkel- en testfase van een interventie.

## Beperkingen van het RCT-design

In het stuk hiervoor is beschreven waarom een RCT-design pas nagestreefd kan worden na een testfase waarin het programma en het instrumentarium goed uitontwikkeld worden. Alhoewel de interne validiteit van een RCT zeer sterk is, heeft het design ook een aantal nadelen. Deze worden hieronder beschreven.

*Selectie-effecten*
Officieel moeten deelnemers aan een studie toestemmen met de randomisatieprocedure. Degenen die weigeren mee te doen, vallen buiten de onderzoeksgroep en daarmee ontstaat er wellicht een *bias* in de groep die wel mee wil doen. Dit wordt lang niet altijd gerapporteerd, waardoor de generaliseerbaarheid van de uitkomsten eigenlijk minder groot is dan het lijkt. De selectie van de deelnemers aan de experimentele conditie, en vereiste homogeniteit in deze groep, maakt de generaliseerbaarheid van bevindingen naar andere populaties moeilijk.

*Externe validiteit minder sterk*
Tijdens een RCT wordt een laboratoriumachtige opzet gecreëerd waarin de experimentele groep alleen die interventie krijgt die wordt onderzocht, waarbij de interventie zo strak mogelijk volgens plan van aanpak wordt uitgevoerd. In de praktijk echter, kunnen mensen ook andere interventies krijgen dan alleen de gedragsinterventie, al dan niet overlappend in tijd. Daarnaast kan de uitvoering in de praktijk minder strak zijn dan tijdens een dergelijke studie. Bijvoorbeeld om rekening te houden met responsiviteit.

*Geen zicht op "wat er waarom werkt"*
Hoewel de attributie van veranderingen in principe duidelijk is doordat het enige verschil tussen experimentele en controlegroep eruit bestaat dat de ene groep wel en de andere niet een interventie kreeg, blijft het onduidelijk hoe de veranderingen tot stand zijn gekomen. Het zicht op de werkzame mechanismen wordt zelden meegenomen in RCT's. Sommige auteurs stellen dat het meenemen van moderatoren dit tekort wegneemt (e.g. Farrington, Gottfredson, Sherman, & Welsh, 2002). Anderen stellen dat dit onvoldoende is (e.g. Pawson & Tilley, 1997; Pawson, 2002).

*Kosten van het werken met experimentele designs*
De ethische, juridische en praktische bezwaren die uitvoerders hebben zijn wellicht niet onoverkomelijk wanneer er veel tijd en moeite wordt gestoken in het overtuigen van de verschillende partijen van het belang van randomiseren. Dit speelt zeker een rol in de voorbereidingsfase, maar ook nog tijdens het onderzoek. Als men medewerking verleent aan randomiseren, is het zaak goed toezicht te houden op dit proces[1]. Doordat de voorbereiding en uitvoering van het randomiseren veel tijd kost, zijn er financieel consequenties voor het onderzoek. Dit speelt een remmende rol in het inzetten van een randomisatieprocedure.

[1] Zie artikel James et al. (2012).

## Behoefte evaluatieonderzoek anno 2013

Het belangrijkste doel van effectevaluaties van interventies is niet alleen te weten te komen of de interventie werkt, maar ook bij wie deze werkt, onder welke omstandigheden, en waardoor. In het geval van gedragsinterventies die zijn beoordeeld en erkend door de Erkenningscommissie Gedragsinterventies van het ministerie van Veiligheid en Justitie is er tot nu toe voornamelijk sprake van een beoordeling op papier (planevaluatie). Erkenningstermijnen zijn vijf jaren geldig. Na vijf jaren dient de doeltreffendheid van de interventie aangetoond te worden. Doeltreffendheid betreft de mate waarin de interventie de beoogde doelgroep bereikt, wordt uitgevoerd zoals bedoeld en de beoogde veranderingen op de programmadoelen haalt. Daarbij dient er een plan te zijn voor een effectstudie, die dan binnen drie jaren dient te worden afgerond, inclusief uitspraken in termen van recidive. Van een aantal interventies zijn al procesevaluaties uitgevoerd door het WODC en van een enkele interventie is al een effectstudie afgerond (MST door Asscher et al., *in press*). Bij een aantal interventies is gebleken dat het opzetten van een RCT niet mogelijk zal zijn, vanwege landelijke uitrol van de ene interventie, en zeer lage aantallen deelnemers bij de andere. Bovendien is gebleken, nu een aantal interventies aan doeltreffendheidstudies worden onderworpen, dat een aantal aspecten nog niet goed is uitgewerkt.

Resumerend is er de komende tijd behoefte aan de volgende ontwikkelingen.

*Tijd nemen voor de testfase*
Duidelijk is geworden dat er één aspect van het onderzoek eerst besproken moet worden; de testfase waarin de interventie zich bevindt. Een interventie die goed uitgewerkt is, is op papier niet meteen klaar voor effectonderzoek. Eerst dient uitgezocht te worden wat de werkzame bestanddelen zijn van de interventie, hoe die werkzaam zijn en tot welke resultaten deze kunnen leiden. Daarbij hoort ook de ontwikkeling van sensitieve instrumenten die de veranderingen in kaart kunnen brengen. Wanneer dat is uitgewerkt, kan gedacht worden aan effectonderzoek. Afhankelijk van de grootte van de doelgroep voor wie de interventie bedoeld is en de wijze van uitrol (bij voorkeur nooit meteen landelijk) kan ook eerst cohort onderzoek gedaan worden. Een RCT is pas na dit voorwerk mogelijk en wenselijk.

Het is de vraag of het noodzakelijk is om de effectiviteit van een interventie die al in het buitenland is aangetoond, nogmaals te onderzoeken wanneer dezelfde interventie in Nederland wordt aangeboden. Het kan zijn dat er verschillen in de uitvoering zullen zijn, bijvoorbeeld als gevolg van contextuele factoren zoals verschillen in juridische kaders. Ook kunnen er verschillen zijn in de *care as usual* waarmee de interventie wordt vergeleken. Volgens Nederlanders is de *care as usual* in Nederland beter dan in andere landen zoals de VS.

In ieder geval dient de testfase lang genoeg te duren om te weten te komen wat de essentiële onderdelen en werkzame mechanismen van de interventie zijn. Eerder heeft een effectevaluatie nog geen zin. Uitkomsten

zijn dan niet eenduidig te verklaren. Een mogelijkheid is om meerdere RCT's te doen, waarbij gevarieerd wordt met elementen van de interventie, zodat zicht komt op de werkzame mechanismen, of tenminste de voorwaarden voor werkzaamheid. McCulloch, Altman, Campbell, et al. (2009) hebben een model uitgewerkt voor de ontwikkeling van chirurgische innovaties, dat goed past bij onderzoek naar gedragsinterventies. Zij onderscheiden 4 fasen: *Idea*, *Development & Exploration*, *Assessment* en *Long-term Study*. Voor elke fase zijn bepaalde onderzoeksmethoden geschikt, en daarmee zijn van elke fase bepaalde typen uitkomsten te verwachten. Tijdens fase 1 zijn *structured case reports* geschikt. Deze kunnen het idee achter de interventie ondersteunen, en dienen om dramatische successsen en falen aan te tonen. In fase 2 tijdens *development* zijn *prospective development studies* mogelijk, tijdens *exploration* database onderzoek of haalbaarheids RCT. In fase 3 zijn RCT's geroepen, eventueel met modificaties of alternatieve designs. Ook in het chirurgisch werkveld zijn RCT's niet altijd mogelijk. In fase 4 is het routinematig bijhouden van gegevens wenselijk (*routine outcome monitoring*; ROM), en *rare case reports* zijn noodzakelijk.

*Doorontwikkelen van uitkomstmaten*
Er is behoefte aan het doorontwikkelen van instrumenten en sensitieve, betrouwbare en valide uitkomstmaten. Op dit moment zijn de instrumenten die de ontwikkelingen op de uitkomstmaten van de interventies meten nog niet volledig uitgetest en gevalideerd. Ondertussen worden interventies al wel afgerekend op basis van de uitkomsten. Dit is een premature en onwenselijke gang van zaken. In feite behoort dit te gebeuren tijdens de testfase van een interventie.

*Alternatieve designs nodig*
De RCT is een sterk design voor het aantonen van de effectiviteit van een interventie bij een bepaalde groep, onder bepaalde omstandigheden. Het design is niet altijd mogelijk bij effectonderzoek in justitiële setting (en daarbuiten). Bij zedendelinquenten is het bijvoorbeeld erg lastig om een goede controleconditie te gebruiken, zolang er nog geen goede *care as usual* en een veelbelovende nieuwe interventie zijn. De verdere ontwikkeling van alternatieve designs voor effectonderzoek is dan ook wenselijk en nodig. De behoefte aan alternatieve designs betreft niet alleen het effectonderzoek. Ook het onderzoeksdesign ten tijde van de testfase behoeft nog doorontwikkeling, bijvoorbeeld het n=1 design.

*Realiseren randvoorwaarden effectonderzoek*
Naast ruimere tijd voor een testfase van programma's, de doorontwikkeling van instrumenten en designs om vooruitgang te meten, is er de komende tijd behoefte aan het beter realiseren van randvoorwaarden voor effectonderzoek. Het is belangrijk om politie en rechters te betrekken bij dit snijvlak van onderzoek en beleid, ondermeer om te verkennen op welke wijze enige vorm van randomisering mogelijk zou kunnen worden.

Op dit moment is de relatie met recidive voor de Erkenningscommissie en beleidsmakers de meest wenselijke uitkomstmaat, maar die kan pas twee jaren na afloop van een interventie onderzocht. Dat kan onderzocht worden, maar in die twee jaren tijd zijn er vele ontwikkelingen en invloeden denkbaar die van invloed zijn op de relatie tussen de effecten van de gedragsinterventie en de recidive twee jaren later. Zicht op wat er in die

tussentijd gebeurt bij de ex-gedetineerde zou zeer wenselijk zijn. Dit is bijna niet haalbaar, tenzij de ex-deelnemer in aanraking blijft met Justitie in kader van nazorg, of als gevolg van recidive. De enige manier waarop recidive zinvol aan een gedragsinterventie verbonden kan worden, is wanneer er sprake is van een zeer zuivere vergelijkbaarheid tussen controle- en experimentele groep. Mede daarom is het van belang om interventies niet meteen landelijk uit te rollen, zolang de effectiviteit ervan nog niet is aangetoond.

Een positieve ontwikkeling betreft het beleid om voortaan het registreren bij uitvoerende organisaties beter te laten lopen, door het inzetten van *Routine Outcome Monitoring* (ROM). Wanneer dit eenmaal loopt, zal het doen van onderzoek sneller en makkelijker worden, aangevuld met bijvoorbeeld *theory-based evaluations* (Leeuw, 2012). Ook kan er in de toekomst wellicht meer gebruik gemaakt gaan worden van het samenstellen van een controlegroep op basis van *propensity score matching* (zie Tollenaar & Van der Laan, 2012). Wat de ROM betreft is het wel belangrijk goed te beseffen dat de ontwikkeling ervan ook nog in de testfase zit. Het is erg belangrijk om nu meteen goed uit te zoeken wat gemeten wordt, hoe en waarom en daar voldoende tijd voor te nemen.

# Referenties

Angrist, J. D. & Pischke, J.S. (2009). Mostly Harmless Econometrics: An Empiricists Companion. Princeton: Princeton University Press.

Apel, R. J. & Sweeten, G. (2010). Propensity score matching in criminology and criminal justice (Chapter 26, Part III-B: Estimation of impacts and outcomes of crime and justice: Innovation in quasi-experimental design). In: *Handbook of Quantitative Criminology*. A.R. Piquero & D. Weisburd (eds). New York: Springer.

Asscher, J.J., Dekovic, M., Manders, W.A., Van der Laan, P. & Prins, P.J.M. (*in press*). A randomized controlled trial of the effectiveness of multisystemic therapy in the Netherlands: Post-treatment changes and moderator effects. *Journal of Experimental Criminology,* DOI:10.1007/s11292-012-9165-9.

Barkun, J.S., Aronson, J.K. Feldman, L.S., Maddern, G.J. & Strasberg, S.M. (2009). Surgical Innovation and Evaluation I: Evaluation and stages of surgical innovations. *The Lancet, 374,* 1089-1096.

Bottomley, A. (1997). To randomise or not to randomise: methodological pitfalls of the RCT design in psychosocial intervention studies. *European Journal of Cancer Care, 6*, 222-230.

Bruinsma, G.J,N. & Weisburd, D. (2007). Experimental and quasi-experimental criminological research in the Netherlands. *Journal of Experimental Criminology, 3,* 83-88.

Bushway, S.D. & Apel, R.J. (2010). Instrumental Variables in Criminology and Criminal Justice. In: *Handbook of Quantitative Criminology (Eds A.R. Piquero & D. Weisburd).* 595-612.

Buysse, W. & Loef, L. (*in press).* Eerst denken, dan doen. Doeltreffendheid van de cognitieve vaardigheidstraining (CoVa) voor justitiabelen. Den Haag: WODC.

Cook, T.D., Shadish, W.R., & Wong, V.C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New
findings from within-study comparisons. *Journal of Policy Analysis and Managemen*t,
27, 724-750.

Duncan, T.E., Duncan, S.C., Strycker, L.A., Li, F., Alpert, A., 1999*. An Introduction to Latent Variable Growth Curve Modeling: Concepts, Issues, and Applications*. Mahwah NJ: Lawrence Erlbaum Associates.

Earle, C.E., Tsai, J.S., Gelber, R.D., Weinstein, M.C., Neumann, P.J. & Weeks, J.C. (2001). Effectiveness of chemotherapy for advanced lung

cancer in the elderly: Instrumental variable and propensity analysis. *Journal of Clinical Oncology (19,4)*, 1064-1070.

Ergina, P.L., Cook, J.A., Blazeby, J.M., Boutron, I., Clavien, P.A., Reeves, B.C. & Seiler C.M. (2009). Surgical Innovation and Evaluation II: Challenges in evaluating surgical innovations. *The Lancet, 374,* 1097-1104.

Farrington, D.P., Gottfredson, L.W., Sherman, B.C., & Welsh, B.C. (2002). The Maryland Scientific Methods Scale. In L.W. Sherman, D.P. Farrington, B.C. Welsh & D.L. MacKenzie (red.), *Evidence-based crime prevention.* Londen: Routledge.

Farrington, D.P. & Welsh, B.C. (2005). Randomized experiments in criminology: What have we learned in the last two decades? *Journal of Experimental Criminology (1,1),* 9-38.

Ferwerda, H.B., Van Leiden, I.M.G.G., Arts, N.A.M. & Hauber, A.R. (2006). Halt: Het alternatief? De effecten van Halt beschreven. *O&B 244*. Den Haag: WODC.

Glazerman, S., Levy Dan.M., & Myers, D. (2002). *Nonexperimental replications of*
*social experiments: A systematic review*. Westport, CT: Smith Richardson Foundation.

Hardcastle, J.D., Chamberlain, J.O., Robinson, M.H., Moss, S.M., Amar, S.S., Balfour, T.W., James, P.D. & Mangham, C.M. (1996). Randomised controlled trial of faecal-occult-blood screening for colorectal cancer. *The Lancet, 348 (9040)*: 1472-7.

Homer, C.S.E. (2002). Using the Zelen design in randomized controlled trials: debates and controversies. *Journal of Advanced Nursing, 38(2),* 200-207.

Hughes, J.P. (2008). Stepped wedge design. In: *Wiley Encyclopedia of Clinical Trials*. DOI: 10.1002/9780471462422.eoct449.

James, C., Asscher, J.J., Dekovic, M., Van der Laan, P.H., & Stams, G.J.J.M. (2012). Endeavors in an experimental study on the effectiveness of an aftercare program in the Netherlands: Research note. *Criminal Justice Policy Review*,
http://cjp.sagepub.com/content/early/2012/04/18/0887403412442891.

Kempes, M.M., Pelt, L. van, Beerthuizen, M.G.C.J., Boom, J., Brugman, D. (2010). Programma-integriteit en effecten van Stay in Love+. Een preventieprogramma voor 12-15 jarige VMBO scholieren dat partnergeweld beoogt te voorkomen. Den Haag: WODC.

Koeter, M.W.J., Bakker, M. (2007). Effectevaluatie van de Strafrechtelijke Opvang Verslaafden (SOV). Den Haag: WODC.

Leeuw, F.L. (2012). Linking theory-based evaluation and contribution analysis: Three problems and a few solutions. *Evaluation, 18(3),* 348-363.

Levitt, S.D. (1996). The effect of prison population size on crime rates: Evidence from prison overcrowding litigation. *The Quarterly Journal of Economics (111,2)*: 319-351.

Loon, D. van der, Meulen, B. van der & Minnaert, A. (2011). *Effectonderzoek in de gedragswetenschappen. Methodologische moeilijkheden en mogelijkheden.* Den Haag: Boom Lemma uitgevers.

McCulloch, P., Altman, D.G., Campbell, W.B., Flum, D.R., Glasziou, P., Marschall, J.C., Nicoll, J. (2009). Surgical Innovation and Evaluation III: No surgical innovation without evaluation: the IDEAL recommendations. *The Lancet, 374,* 1105-1112.

Nesselroade, J.R., Baltes, P.B. 1979. Longitudinal research in the study of behavior and development. Academic Press, San Diego, CA.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. Londen: Sage.

Pawson, R. (2002). Evidence-based policy: The promise of 'realist synthesis'. *Evaluation*, *8*, 340-358.

Pearson, D., Torgerson, D., McDougall, C. & Bowles, R. (2010): Parable of two agencies, one of which randomizes. *The ANNALS of the American Academy of Political and Social Science* 2010 628: 11. DOI: 10.1177/0002716209351500.

Posner, M.A., Ash, A.S., Freund, K.M., Moskowitz, M.A. & Shwartz, M. (2002). Comparing standard regression, propensity score matching, and instrumental variables methods for determining the influence of mammography on stage of diagnosis. *Health Services & Outcomes Research Methodology (2),* 279–290.

Sherman, L.W. (2010). An Introduction to Experimental Criminology. In: *Handbook of Quantitative Criminology (Eds A.R. Piquero & D. Weisburd).* 595-612.

Sherman, L.W., Gottfredson, D.C., MacKenzie, D.L., Eck, J., Reuter, P. & Bushway, S.D. (1998). Preventing Crime: What Works, What Doesn't, What's Promising. *National Institute of Justice: Research in brief.* https://www.ncjrs.gov/pdffiles/171676.PDF

Stukel, T.A., Fisher, E.S., Wennberg, D.E., Alter, D.A., Gottlieb, D.J. & Vermeulen, M.J. (2007). Analysis of observational studies in the presence of treatment selection bias: Effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *Journal of American Medical Association (297,3)*, 278-285.

Sturmans, F. (1997). The Prescription of Heroin to Heroin Addicts: A proposal for the Netherlands. In: Bammer, G. (Ed.); *International Perspectives on he Prescription of Heroin to Dependent Users: A collection of papers from the United Kingdom, Switzerland, the Netherlands and Australia*. Australian Institute of Criminology.

Tollenaar, N., Laan, A.M. van der (2012). Effecten van de ISD-maatregel. Fact sheets 2012-01. Den Haag: WODC.

Torgerson, D.J. & Roland, M. (1998). What is Zelen's design? *British Medical Journal 316(7131)*, 606.

Vollaard, B.A. (2010). Het effect van langdurige opsluiting van veelplegers op de maatschappelijke veiligheid: lessen van een natuurlijk experiment in twaalf stedelijke gebieden. Tilburg: Politie & Wetenschap.

Zaalberg, A., Nijman, H., Bulten, E., Stroosma, L., Staak, C. van der (2009). Voeding en agressieregulatie. Cahier: 2009-05. Den Haag: WODC.

Zelen, M. (1979). A new design for randomized clinical trials. *New England Journal of Medicine (300)*, 1242-1245.

# Bijlage 1 Voorbeelden van designs gebruikt in WODC onderzoek

*HALT onderzoek*[2]
In 2006 is een effectonderzoek gedaan naar de Halt-afdoening voor jeugdigen. In het onderzoek is een groep van bijna duizend jongeren gedurende een jaar op verschillende momenten gevolgd. Het betreft jongeren die zijn opgepakt nadat ze een strafbaar feit hebben gepleegd en door de politie naar Halt zijn verwezen. Bij Halt zijn de jongeren – nadat ze hadden ingestemd met deelname aan het onderzoek – verdeeld over twee onderzoeksgroepen. De helft vervolgt het traject van de Haltafdoening, de andere helft wordt hiervan in het kader van het onderzoek vrijgesteld. Hierbij heeft overigens uiteindelijk om praktische redenen (beperkte omvang van de steekproef per Halt-bureau, groepsgewijze toewijzing van jongeren die delicten in groepsverband hadden gepleegd) geen toevalstoewijzing kunnen plaatsvinden. Er is wel gezorgd voor vergelijkbaarheid met de controlegroep. Er was dus sprake van een *quasi-experimenteel design.* Het onderzoek wees uit dat Halt geen invloed heeft op het terugdringen van criminaliteit en gedragsproblemen bij jongeren. Het opgepakt worden door de politie zou wel eens belangrijker kunnen zijn dan het daadwerkelijk krijgen van de straf. Spijtbetuiging blijkt een belangrijk recidiveverminderend element in de Halt-afdoening. Halt blijkt het beste te werken bij bepaalde (lichte) groepen van *first offenders*.

*SOV onderzoek*[3]
De effecten van het programma Strafrechtelijke Opvang Verslaafden (SOV) is onderzocht in een *quasi-experimenteel design* met twee controlegroepen die een andere vorm van behandeling voor hun verslaving kregen in justitieel kader en een controlegroep met reguliere detentie. De drie doelstellingen waarop het effect van de SOV werd getoetst zijn vermindering criminaliteit, vermindering verslavingsproblematiek en bevordering terugkeer naar de maatschappij. De nameting vond plaats een jaar na uitstroom. Bij zo'n 80% van de respondenten kon een follow-up meting worden gedaan. Dit is een hoog responspercentage dat tot stand is gekomen door zeer veel investering van tijd en moeite in het blijven volgen van deelnemers, ook tijdens de tussenperiode waarin geen metingen werden gedaan. De conclusies van de onderzoekers luiden: "Op basis van de bevindingen van ons onderzoek levert de SOV significant en substantieel betere uitkomsten op het gebied van criminaliteit, verslaving en maatschappelijk functioneren dan reguliere detentie en vergelijkbare uitkomsten als de twee binnen het dranginterventiespectrum intensievere dranginterventies. Onze eindconclusie luidt dan ook dat de SOV voldoet aan de gestelde succescriteria. De succespercentages zijn echter bescheiden en nemen in de loop van de tijd af" (p.20, Koeter & Bakker, 2007).

*Effectiviteit van Stay in Love*[4]
Recentelijk is een effectstudie afgerond waarin een programma voor VMBO-scholieren werd getoetst op effectiviteit. Het programma genaamd *Stay in*

[2] Ferwerda et al. (2006).
[3] Koeter en Bakker (2007)
[4] Kempes et al. (2010).

*Love* beoogt partnergeweld te voorkomen dan wel te verminderen. Om de effecten ervan te onderzoeken werden vragenlijsten m.b.t. attitude, kennis en sociale vaardigheden gebruikt. Alle jongeren werden op vier meetmomenten gevraagd deze vragenlijsten in te vullen. Het moment waarop jongeren de trainingen kregen was verschillend. Klassen werden ingedeeld in 3 condities. De klassen in de eerste conditie kregen de trainingen tussen de meetmomenten één en twee, klassen in de tweede conditie kregen de trainingen tussen de meetmomenten twee en drie, en klassen in conditie drie kregen de trainingen tussen de meetmomenten drie en vier. Zodoende vonden tussen de één en drie meetmomenten voorafgaand aan de uitvoering van het lesprogramma plaats en tussen de één en drie meetmomenten na afloop van het programma. Het model waar gebruik van gemaakt is, heet het *cohort sequential latent growth model* (of geaccelereerd cohort model).

In dit model is de power (kracht van het design) om effecten aan te tonen niet minder dan bij meer traditionele designs met externe controlegroepen in scholen zonder enige interventie (Boom, 2008). Er zijn nu tot wel drie voormetingen bij exact dezelfde deelnemers zonder interventie. Met een externe controlegroep zijn er maar twee metingen die alleen door randomisatie van individuen zo goed mogelijk vergelijkbaar gemaakt kunnen worden. Goede randomisatie vereist grotere aantallen deelnemers en is logistiek moeilijk haalbaar in het onderwijs (immers leerlingen zijn niet aselect toegewezen aan klassen en scholen). Dit betekent dus dat, ook als de effecten van het programma klein zijn, deze bij het voorgestelde design eerder opgemerkt zullen worden en tegen aanzienlijk lagere kosten dan bij meer traditionele designs met externe controlegroepen. In dit onderzoek werd dus geen externe controlegroep gevormd. Doordat per leerling meerdere metingen aanwezig waren, vormden zij als het ware hun eigen controle. Door de gegevens van de condities modelmatig te combineren ontstaat er een design met drie voormetingen en drie nametingen. In Tabel 1 wordt dit model visueel gemaakt door de condities aan de hand van het moment waarop de lessen plaatsvinden met elkaar gelijk te stellen.

Tabel 1. *Weergave model zoals afgeleid van de onderzoeksopzet. M1 t/m M4 stellen de momenten voor waarop leerlingen vragenlijsten invullen. Deze metingen vormen de basis voor T1 t/mT6 de zogenaamde pseudometingen (niet elk individu draagt hieraan bij).*

| | T1 | T2 | T3 | Trainingen | T4 | T5 | T6 |
|---|---|---|---|---|---|---|---|
| Conditie 1 | - | - | M1 | X | M2 | M3 | M4 |
| Conditie 2 | - | M1 | M2 | X | M3 | M4 | - |
| Conditie 3 | M1 | M2 | M3 | X | M4 | - | - |

In het model zoals weergegeven in tabel 1 dragen de afzonderlijke condities elk maar vier meetmomenten per individu bij. Zoals de streepjes aangeven zijn niet van elke leerling alle waarden bekend op alle metingen. Omdat er van genoeg andere leerlingen op dat moment wel een waarde bekend is, mag verwacht worden dat er een goede schatting te maken is van de ontbrekende waarden en daarmee van algemene trends voor en na de interventie. Bij dit design zijn voldoende aantallen dus essentieel.

De eindconclusies van de onderzoekers waren: "Algemeen genomen kan gesteld worden dat Stay in Love + een klein, kortdurend effect op attitude

en in mindere mate op kennis en sociale vaardigheden m.b.t. partnergeweld heeft. De matige programma-integriteit en de door scholieren gerapporteerde onveilige sfeer kunnen hebben bijgedragen aan het feit dat dit effect alleen klein en kortdurend was. Betere programma-integriteit hing samen met minder snelle afname van attitude en kennis twee maanden na Stay in Love+. Vervolgonderzoek moet vaststellen of de effecten van Stay in Love+ ook daadwerkelijk leiden tot een vermindering in partnergeweld." (p.3, Kempes et al. 2010).

*ISD-maatregel effectonderzoek*[5]
Om de effecten van de ISD-maatregel vast te stellen is er een *quasi-experimentele onderzoeksopzet* gebruikt. Zeer actieve veelplegers (ZAVP's, n=554) die een ISD-maatregel kregen opgelegd zijn vergeleken met twee vergelijkbare controlegroepen van veelplegers (beide n=554) die een standaardvrijheidsstraf opgelegd hebben gekregen. Randomiseren was niet mogelijk, omdat het om een maatregel gaat die door de rechter wordt opgelegd. De controlegroepen zijn op zo veel mogelijk kenmerken vergelijkbaar gemaakt met ZAVP's in de ISD-groep. Naarmate er meer kenmerken worden gebruikt die gerelateerd zijn aan recidive, zijn de interventie- en controlegroepen beter vergelijkbaar. Om de groepen te vergelijken zijn demografische en criminele carrièrekenmerken gebruikt, evenals kenmerken van de uitgangszaak en de aanwezigheid van problematiek op verschillende leef gebieden. Voor iedere ZAVP die een ISD-maatregel opgelegd heeft gekregen, is een op deze kenmerken vergelijkbare ZAVP gevonden die een standaardvrijheidsstraf heeft gekregen.

Om de effecten van de ISD-maatregel op recidive te kunnen schatten, zijn zoals gezegd
naast de interventiegroep twee controlegroepen samengesteld uit de database van het gevangeniswezen. Een '*propensity score matching'* (PSM)[6] met twintig covariaten is gebruikt. De PSM is gedaan met de 'nearest neighbor' matching zonder cases meerdere keren te matchen (zonder teruglegging). Als er meerdere vergelijkbare cases zijn in de controlegroepen dan is de selectie van de gematchte case willekeurig. Achteraf is nagegaan in hoeverre de matching is gelukt door de groepen op de covariaten te vergelijken. Daarvoor zijn t-toetsen gehanteerd, met een alpha van 5% om de verschillen tussen controlegroep en ISD-groep te bepalen. Ook de '*standardized bias*' (SB) geeft aan in welke mate de matching in balans is op de covariaten, met andere woorden in welke mate de matching is geslaagd. De groepen bleken na matching goed vergelijkbaar. Van de controlegroepen was niet bekend of zij tussentijds gedragsinterventies hebben gehad. Ook zijn er geen vergelijkingen gemaakt tussen de groepen in termen van de programmadoelen van de ISD-maatregel, alleen in termen van recidive.

*Effect van voeding op gedrag*[7]
In deze evaluatie is het randomiseren wel gelukt. De vraag was of de mate van agressiviteit en de psychische conditie van jongvolwassen gedetineerden in positieve zin beïnvloed konden worden door in te grijpen

[5] Tollenaar & Van der Laan (2012).
[6] Zie bijvoorbeeld Apel & Sweeten (2010)
[7] Zaalberg et al. (2009)

in hun voedingsstatus. Hiervoor slikten 221 jongvolwassen gedetineerden, gedurende minimaal één en maximaal drie maanden, voedingssupplementen dan wel placebo's. De actieve capsules bevatten zowel essentiële vetzuren (ω-3 en ω-6) als een veelheid aan vitaminen en mineralen. Gegevens werden verzameld via registraties en ondervraging. Het onderzoek was *double-blind en placebo-controlled* opgezet. De vergelijking tussen experimentele en controlegroep werd wel bemoeilijkt doordat het randomiseren niet op het niveau van gevangenisafdelingen was gedaan, maar over de hele groep. Daardoor konden op afdelingen scheve verhoudingen bestaan tussen het aantal in de experimentele en controleconditie.

De resultaten wijzen erop dat het aantal incidentenregistraties gedaald was tijdens de nameting voor de groep die voedingssupplementen had gekregen (n = 115) ten opzichte van de placebogroep (n = 106). Deze bevinding wat betreft de ontwikkeling van het aantal gerapporteerde incidenten is in lijn met de resultaten van een eerder Brits onderzoek van Gesch en collega's (2002a). Aangezien echter op een aantal andere agressievragenlijsten geen significante verbeteringen werden gevonden, kan niet zonder meer gesteld worden dat deze studie een antiagressief effect van voedingsupplementen heeft aangetoond. Er werd ook geen verbetering gevonden met betrekking tot de psychische conditie.

*Effectiviteit cognitieve vaardigheidstraining*[8]
Op dit moment loopt het effectonderzoek naar een cognitieve vaardigheidstraining voor justitiabelen die zowel intra- als extramuraal wordt gegeven. De trainingen worden in 20 bijeenkomsten van ongeveer 2 uur gegeven en worden groepsgewijs aangeboden (10-12 deelnemers intramuraal; 12-14 deelnemers extramuraal) door twee trainers per groep. Randomiseren was niet mogelijk. Er wordt gewerkt met een *quasi-experimenteel design*. Naar verwachtingen zal de experimentele groep uit ongeveer 400 deelnemers bestaan. Aanvankelijk leek er een grote controlegroep te bestaan, maar dit bleek in de praktijk niet het geval. Vrijwel iedereen die voor de training in aanmerking komt, krijgt deze ook. Het streven is een controlegroep te vormen, die op basis van praktische redenen de training niet kreeg. Bij de controlegroep (n=100) wordt een voormeting gedaan om zicht te krijgen op de vergelijkbaarheid van de controlegroep met de experimentele groep.

Om de effectiviteit van deze training te toetsen, worden bij deelnemers aan de trainingen voor- en nametingen gedaan op programmadoelen. Bij de controlegroep is alleen een voormeting haalbaar en nodig. Het opsporen van ex-gedetineerden is een dermate kostbare en lastige zaak, dat de baten niet opwegen tegen de kosten. Bovendien stellen de onderzoekers dat er geen aanleiding is te verwachten dat de cognitieve vaardigheden zullen verbeteren zonder een training daarvoor. Een oplossing voor het missen van een tweede meting bij de controlegroep, is het doen van een *theory-based evaluation* (Leeuw, 2012). Daarbij kan het empirische onderzoek aangevuld worden met een analyse van de vermoedelijke validiteit van de interventietheorie. Die validiteit wordt getoetst door na te gaan wat over de mechanismen die in de theorie benoemd zijn, bekend is. Hoe sterker die interventietheorie staat, des te groter de kans dat de idee

[8] Buysse & Loef (*in press).*

dat de interventie werkt, is. Daarbij geldt wel dat er geen aanleiding is om te veronderstellen dat er allemaal vreselijke implementatieproblemen zijn. Om de experimentele en controlegroep te vergelijken op recidive wordt de controlegroep, uitgebreid tot een vergelijkbaar aantal als in de experimentele groep. Hierbij zal gebruikt gemaakt worden van *propensity score matching*.

*Effectiviteit van een pleegzorg programma*[9]
Een ander effectonderzoek dat op dit moment loopt, betreft de effectevaluatie van een zeer intensieve gedragsinterventie voor jeugdige justitiabelen. Bij dit programma is sprake van plaatsing in een pleeggezin gedurende zes à negen maanden, ter vervanging of bekorting van gesloten behandeling in justitiële jeugdinrichtingen (JJI's) en jeugdzorgplus-instellingen (Jz+). In 2011 was er sprake van elf deelnemers (en elf opvoedgezinnen). In 2012 wordt eenzelfde aantal verwacht. Een dergelijk aantal deelnemers en complex selectie- en plaatsingsproces stelt bijzondere eisen aan de onderzoeksopzet. In deze studie maken wij gebruik van *het gestapelde n=1 design*. In dit design wordt elke jongere als een *case-study* gevolgd. Bij voorkeur worden per jongere meerder nulmetingen, op verschillende tijdstippen voor aanvang van de interventie, verricht, zodat de jongere als zijn of haar eigen controleconditie fungeert. Binnen het programma is een aantal fasen te onderscheiden, die opeenvolgend doorlopen dienen te worden. Tussen elke fase worden metingen verricht, waarbij meerdere personen uit het netwerk van de jongere dienen als bron. Naast de vergelijking over tijd binnen deelnemers zelf, wordt ook een vergelijking gemaakt met een controlegroep die *care as usual* krijgt. Belangrijk is dat het gaat om jongeren met eenzelfde recidiverisico en vergelijkbare scores op risico- en beschermende factoren. Overigens zullen jongeren die uitvallen tijdens de behandeling gemonitord blijven worden volgens het 'intention to treat'- principe.

Per deelnemer kan bepaald worden wat betekenisvolle vooruitgang is. Als er sprake is van vergelijkbare patronen in de veranderingen op gedragsdoelen bij verschillende leden van de doelgroep, dan onderbouwt dit de bewijskracht voor de effectiviteit van de interventie (Van Yperen et al., 2008). Om te kunnen spreken van effectiviteit, dienen de case studies wel te voldoen aan een aantal voorwaarden (Van Yperen et al., 2008). Ten eerste dienen de cases representatief te zijn voor de doelgroep van de interventie. Gedegen selectie en registratie van kenmerken van deelnemers zijn essentieel. Ten tweede is de validiteit van de effectmaten zeer belangrijk en deze moeten goed te meten zijn. Ten derde dient er sprake te zijn van voldoende metingen om het verloop goed in kaart te brengen, ook omdat uitgesloten moet worden dat natuurlijke ups en downs gezien worden als effecten. Tenslotte dient er een duidelijk veranderingsmodel ten grondslag te liggen aan de interventie en de beoogde effecten, zodat duidelijk is wat kan worden verwacht.

[9] Nog geen publicatie: onderzoek in uitvoering.