

Rovers, B. 2007. "What works; kanttekeningen bij een populair programma." *Tijdschrift voor Veiligheid*, (6) 3, 7-22.

What works; kanttekeningen bij een populair programma

Voor veel onderzoekers en beleidmakers in het justitieveld staat de *What works* benadering voor het streven naar effectieve(re) interventies door middel van onderzoek dat voldoet aan de hoogste methodologische eisen (gerandomiseerde veldexperimenten), door gebruikmaking van kwantitatieve research syntheses (door middel van meta-analyses) en door binnen de interventies te focussen op empirisch onderbouwde methodieken. In dit artikel worden kanttekeningen geplaatst bij deze benadering. Deze kanttekeningen betreffen niet de achterliggende gedachte van *evidence-based* werken, maar de uitwerking die hieraan wordt gegeven. Beoogd wordt dat de hiervoor genoemde aanpak niet in alle gevallen zal leiden tot effectieve(re) interventies. In voorkomende gevallen kan deze aanpak in zijn uitwerking zelfs haaks komen staan op het beoogde doel (effectievere interventies). De auteur doet enkele suggesties voor toekomstig onderzoek naar interventies.

1. Inleiding

Gedragregulering vindt op verschillende manieren plaats. Persoonsgerichte interventies in het veiligheidsveld kunnen worden beschouwd als één van de mogelijke vormen. Deze interventies zijn meestal gericht op het voorkómen van crimineel gedrag of het verminderen van recidive. De wijze waarop deze doelstellingen worden nagestreefd kan sterk variëren, al naar gelang de context waarin de interventie plaatsvindt (binnen of buiten het justitiedomein); zo kan een school een antispijbelprogramma ontwikkelen, een buurthuis kan activiteiten opzetten voor overlastgevendende jongeren, een hulpverleningsorganisatie kan gezinstherapie aanbieden, een inrichting voor jeugdige gedetineerden leert de jongeren vaardigheden aan, een reclasseringsorganisatie biedt nazorg voor gedetineerden, et cetera. Deze interventies hebben gemeen dat ze (mede) beogen de betrokkenen uit de criminaliteit te houden. De mate waarin deze interventies succesvol zijn, is onderwerp van de *What works* benadering.

In het denken over *evidence-based* interventies in het veiligheidsveld neemt de *What works* benadering een belangrijke plaats in. Het is lastig om deze benadering scherp af te bakenen van andere benaderingen, maar we kunnen wel enkele contouren schetsen die we hierna zullen bespreken. De kanttekeningen die ik wil plaatsen bij deze benadering betreffen niet de achterliggende gedachte van *evidence-based* werken, maar de wijze waarop hieraan vorm wordt gegeven door veel onderzoekers en ook beleidmakers, en de ongewenste effecten die hiervan het gevolg kunnen zijn, namelijk dat de onderzoeksinspanningen niet leiden tot het gewenste doel: effectievere interventies.

In paragraaf 2 zal ik kort de achtergronden van de *What works* benadering schetsen en de belangrijkste constituerende elementen beschrijven. Daarna zal ik in de paragrafen 3 en 4 aangeven welke kanttekeningen er zoal te plaatsen zijn bij deze benadering, waarna ik in paragraaf 5 de resultaten zal bespreken en besluit met enkele suggesties voor verder onderzoek.

2. What works – een korte introductie

Het streven naar effectieve interventies is van alle tijden. De wijze waarop hieraan vorm wordt gegeven verschilt naar tijd en plaats en wellicht ook per discipline. Beoefenaren van sociaal-wetenschappelijke disciplines kijken tegenwoordig graag naar de medische sector waar al meer dan 50 jaar geleden het gebruik van *Randomized Controlled Trials* (RCT's) de

gouden standaard werd voor het vaststellen van effecten van medische interventies (Petrosini et al. 2001, Harrington 2002).

In het veiligheidsveld is het denken over *evidence-based* interventies van jongere datum (Farrington & Welsh 2005). De publicaties van Martinson (1974) en Lipton, Martinson en Wilks (1975) figureren in historische overzichten vaak als een soort startpunt voor beschouwingen over de *What Works* benadering. Deze auteurs kwamen na evaluatie van tal van effectstudies tot de pessimistische conclusie dat, op enkele geïsoleerde uitzonderingen na, de pogingen om regelovertreeders te rehabiliteren mislukt waren: men was niet in staat gebleken de recidive van betrokkenen omlaag te krijgen. Deze conclusie is later kortweg samengevat als: *Nothing works*. Een conclusie die door veel criminologen nadien voor waar werd aangenomen zonder zelf aanvullend onderzoek te doen (Cullen & Gendreau 2000).

Diverse ontwikkelingen hebben er voor gezorgd dat het vraagstuk van *evidence-based* werken in het veiligheidsveld toch weer op de agenda is gekomen. De ontwikkeling van nieuwe onderzoeksmethoden, met name de opkomst van meta-analyse, heeft hier waarschijnlijk een belangrijk aandeel in gehad. Ook het feit dat overheden in toenemende mate kritische kanttekeningen zijn gaan plaatsen bij de financiering van allerlei (bij voorbaat zinloze?) interventies, heeft hierbij een rol gespeeld. De positievere uitkomsten van onderzoek naar rehabilitatieprogramma's door een groep Canadese onderzoekers (Cullen & Gendreau 2000) en later ook de meer genuanceerde resultaten van grootschalige research syntheses door Sherman et al. (1997), in opdracht van het Amerikaanse Congres, hebben eraan bijgedragen dat vandaag de dag met meer positivisme wordt aangekeken tegen de mogelijkheden van persoonsgerichte preventie- en rehabilitatieprogramma's voor (jonge) regelovertreeders. De oprichting, in 2000, van de 'Campbell Collaboration' (een forum voor het bevorderen van *evidence-based* kennis) bevestigt het hernieuwde geloof in de mogelijkheid om door middel van hoogwaardig onderzoek te komen tot verbeterde inzichten in de werkzaamheid van justitiële interventies (<http://www.campbellcollaboration.org/>).

Deze ontwikkeling zien we ook terug in het beleid. Beleidmakers zijn hogere eisen gaan stellen aan interventieprogramma's en aan onderzoek dat deze programma's moet toetsen dan wel rechtvaardigen (Beenackers 2000, Wartna 2005, National Research Council 2005, Van der Laan et al. 2006). In 2005 werd in Nederland de Erkenningscommissie Gedragsinterventies geïnstalleerd. Deze commissie adviseert het Ministerie van Justitie over de kwaliteit van justitiële interventies. Dit doet zij op basis van *What works* beginselen. Tot en met eind 2006 werden aan deze commissie 13 interventies ter toetsing voorgelegd. Zes daarvan werden voorlopig erkend (dit betekent dat de interventie niet aan alle kwaliteitscriteria voldoet, maar voldoende potentieel heeft), de overige zeven werden niet erkend. Géén van de dertien voorgelegde interventies werd volledig erkend, dat wil zeggen: had voldaan aan alle kwaliteitscriteria (Erkenningscommissie Gedragsinterventies 2006).¹

Hoe stel je nu vast of interventies werken? Het antwoord op deze vraag brengt ons bij wat ik de constituerende elementen van de *What works* benadering zou willen noemen. Zoals in de inleiding reeds aangegeven, is het lastig om deze benadering scherp af te bakenen van andere benaderingen, omdat er geen officieel protocol bestaat. *What works* staat voor een aantal uitgangspunten die we vaak gehanteerd zien, maar waarover geen consensus hoeft te bestaan bij de betrokkenen. De hierna omschreven elementen zien we vaak terug in de *What works* benadering, in het bijzonder bij onderzoekers en beleidmakers die zich associëren met de uitgangspunten van de Campbell Collaboration. In mijn optiek zijn de volgende drie onderdelen hierbij cruciaal:

¹ Recentelijk is ook een landelijke Erkenningscommissie voor Jeugdinterventies van start gegaan (bron: www.nji.nl).

1. de effecten van interventies zijn onderzocht aan de hand van *research designs die zo hoog mogelijk scoren op interne validiteit*;
2. Om te kunnen spreken over bewezen effecten beschikt men bij voorkeur over *een zo groot mogelijke hoeveelheid data* die verkregen is middels de hiervoor genoemde designs;
3. Binnen de interventies is de aandacht vooral gericht op *bewezen methodieken*.

Toelichting

Om de methodologische kwaliteit van onderzoek vast te stellen, met name op het punt van de interne validiteit, zijn diverse criteria en schalen ontwikkeld. De *Maryland Scientific Methods Scale* (SMS) is wellicht de bekendste in het veiligheidsveld (Farrington et al. 2002). Op een 5-punts schaal wordt weergegeven in hoeverre uitkomstvariabelen kunnen worden toegeschreven aan de gepleegde interventie. Bij een niveau-1 studie gebeurt dit op basis van correlaties tussen interventievariabelen en uitkomstvariabelen, bij een niveau-5 studie (het hoogste niveau) worden effecten vastgesteld door gebruik te maken van een experimentele opzet (vergelijking tussen experimentele en controlegroep), het uitvoeren van voor- en nametingen in beide groepen, controle voor storende variabelen en gerandomiseerde toekenning van de interventie. Er wordt gestreefd naar studies van een zo hoog mogelijk niveau, omdat deze beter in staat zijn de causale effecten van interventies vast te stellen.

Het tweede element, het streven naar empirische validering door gebruikmaking van zoveel mogelijk data, zien we vooral terug in de grote hoeveelheid research syntheses die in de voorbije jaren geproduceerd zijn. Deze studies zijn gericht op het vaststellen van effecten van interventies door de resultaten uit (zoveel mogelijk) verschillende onderzoeken 'te stapelen'. Er zijn verschillende methoden om een research synthese uit te voeren en net als voor de studies die onderwerp zijn van een research synthese, geldt ook voor de research synthese zelf dat er een ordening kan worden aangebracht in de methodologische kwaliteit. *Narrative review*, de *vote count* methode en meta-analyse zijn veel gebruikte methoden om onderzoeksresultaten samen te vatten (Welsh & Farrington 2001). Met name meta-analyses zijn de laatste jaren zeer populair geworden en vormen een belangrijke peiler onder de *What works* benadering. Belangrijke kenmerken zijn: een uitputtende zoektocht naar relevante studies, expliciete inclusie- en exclusiecriteria voor studies, het gebruik van een expliciet codeerschema om studiekekenmerken en resultaten te beschrijven en het gebruik van de beste statistische methoden om de studieresultaten samen te vatten in één simpele kwantitatieve maat; de gemiddelde effectgrootte (Lipsey & Wilson 2001, Petrosini et al. 2001, Wilson 2001). Meta-analyse wordt door velen gezien als een vorm van research synthese die in methodologisch opzicht superieur is aan de andere vormen en daarom te verkiezen.

De voorgaande elementen hebben betrekking op *hoe* je de effecten van interventies het beste kan vaststellen. Het laatste element zoomt in op de inhoud van de interventie zelf. We zien hier een sterke focus op 'bewezen' methodieken. Dit zijn methodieken waaraan een geëxpliciteerde én empirisch gefundeerde beleidstheorie ten grondslag ligt. Een beleidstheorie is een theorie die verklaart of voorspelt hoe de interventie zal uitpakken (Leeuw 2005). Explicitering van de theoretische aannames die aan een interventie ten grondslag liggen en toetsing van deze aannames aan het beschikbare empirisch materiaal wordt gezien als een belangrijke en noodzakelijke stap naar effectieve(re) interventies (Berghuis et al. 2004, Leeuw 2005).

Op basis van het voorgaande laten de uitgangspunten van de *What works* benadering zich als volgt samenvatten: om te kunnen bepalen of interventies werken, hebben we onderzoek nodig. Niet zomaar onderzoek, maar onderzoek dat voldoet aan de hoogste criteria van interne validiteit. Hoe meer van dit soort onderzoek we tot onze beschikking hebben (dus: hoe groter

onze database), des te gefundeerder ons oordeel over de effectiviteit van deze interventies. Meta-analyses zijn volgens velen het beste middel om dit oordeel te vellen. Bij het beoordelen van de werkzame stof in interventies staan bewezen methodieken centraal. Dit zijn methodieken die zijn gebaseerd op beleidstheorieën die voldoende empirisch onderbouwd zijn.

Hierna zal ik mijn kanttekeningen plaatsen bij deze benadering. Deze betreffen enerzijds de gebruikte onderzoeksmethodologie en anderzijds de totstandkoming en toepassing van de *What works* beginselen.

3. Kanttekeningen bij de onderzoeksmethodologie

In de *What works* benadering ligt een sterke nadruk op de methodologische kwaliteit van de gehanteerde onderzoeksdesigns. Bij primaire studies gaat het dan vooral om gerandomiseerde (veld)experimenten en designs die dit model benaderen. Bij research syntheses gaat het vooral om meta-analyses. Bij het gebruik van deze designs wil ik de volgende vier kanttekeningen plaatsen, die ik aansluitend zal uitwerken:

1. Onderzoeksdesigns die hoog scoren op interne validiteit zijn beperkt toepasbaar
2. Hoge interne validiteit gaat vaak ten koste van externe validiteit
3. Nadruk op interne validiteit van designs leidt aandacht af van werkzame stof in interventies
4. Meta-analyses zijn noodzakelijke maar geen voldoende voorwaarde voor realiseren van effectieve interventies

Ad 1) Onderzoeksdesigns die hoog scoren op interne validiteit zijn beperkt toepasbaar

Het aantal studies in het veiligheidsveld dat hoog scoort op de SMS-schaal is –naar verhouding- gering (Lipsey 2001, Farrington & Welsh 2005). Dit is echter geen natuurverschijnsel of een situatie die voortvloeit uit het onwaarschijnlijke gegeven dat onderzoekers ‘slechte designs’ prefereren. Verschillende factoren beperken de mogelijkheden om dit soort designs te gebruiken.

Een belangrijke factor vormen de juridische randvoorwaarden waarbinnen (justitiële) interventies worden uitgevoerd. Experimenteren met interventies kan op gespannen voet staan met juridische beginselen die ten grondslag liggen aan diezelfde interventies, zoals het uitgangspunt dat gelijke gevallen gelijk behandeld moeten worden. Het juridisch kader begrenst aldus de mogelijkheden om vrijelijk te experimenteren (De Roos 2007). In het verlengde daarvan liggen ook ethisch-normatieve argumenten die pleiten tegen willekeurige toekenning van respondenten aan experimentele en controlecondities. Een groot verschil met de medische sector is hier dat de interventies zelf moreel geladen zijn.²

Behalve juridische beperkingen zijn er praktische beperkingen aan het gebruik van deze designs. Deze beperkingen kunnen zich op tal van vlakken voordoen en betreffen zowel de implementatie als de uitvoering (Berk 2005). Hierbij gaat het om zaken als het creëren van goede experimentele condities (randomisering etc.), voldoende deelnemers in beide onderzoekscondities (experimentele en controlegroep), een programma dat lang genoeg loopt om het te kunnen evalueren, de financiële middelen om dit te realiseren, de medewerking van organisaties, professionals en justitiabelen, et cetera. Het succesvol uitvoeren van een

² Overigens roept de toepassing van justitiële interventies waarvan het effect niet bewezen is, ook morele vragen op.

gerandomiseerd veldexperiment vergt een protocol waaraan strak de hand wordt gehouden. Dit blijkt vaak moeilijk te realiseren; in de praktijk ontstaan –om allerlei redenen- vaak afwijkingen van ‘de behandeling zoals gepland’. Weerstand van medewerkers, maar ook onbewuste schendingen van het protocol, vormen veel voorkomende bedreigingen voor het onderzoek (Feder et al. 2000, Raynor 2003, Atkinson, 2004, Berk 2005, Asscher et al. 2007). Zelfs Lipsey (2001), een fervent voorstander van het gebruik van experimentele designs, constateert dat de bronnen en omstandigheden die nodig zijn voor het realiseren van deze designs in de praktijk nogal schaars zijn.

Verschillende auteurs hebben gewezen op het belang van de beleidscontext en hoe deze het gebruik van experimentele designs door onderzoekers beïnvloedt. Pawson (2002) heeft in dit verband gewezen op het feit dat reguliere beleidscycli op gespannen voet staan met het gebruik van experimentele onderzoeksdesigns: beleidsmakers willen resultaten op korte termijn, terwijl onderzoek door middel van experimentele designs vraagt om een lange-termijn blik. Lum & Yang (2005) stellen dat de sterke beleidsoriëntatie van de criminologie ervoor heeft gezorgd dat in deze discipline minder experimenteel onderzoek wordt gedaan dan in andere sociaal-wetenschappelijke disciplines. Men volgt vooral de beleidsagenda.

Samenvattend: de toepassingsmogelijkheden van experimentele designs in het veiligheidsveld worden door verschillende factoren beperkt. Sommige van deze factoren hebben een min of meer structureel karakter.

Ad 2) Hoge interne validiteit gaat vaak ten koste van externe validiteit

Cronbach (1982) stelde zichzelf de vraag ‘Wat heb je aan onderzoeksresultaten als je ze niet kan generaliseren?’ Zijn antwoord laat zich raden.

Randomisering bij onderzoek in het veiligheidsveld stuit op allerhande juridische, praktische en beleidsmatige beperkingen. Echter, wanneer we ondanks al deze beperkingen toch in staat zijn om randomisering toe te passen, zijn we nog niet uit de problemen. Randomisering kan namelijk de natuurlijke situatie aantasten, waardoor de winst die aan de ene kant wordt geboekt (hogere interne validiteit) aan de andere kant weer moet worden afgeboekt (geringere generaliseerbaarheid van de onderzoeksresultaten).

Recent onderzoek in Nederland naar de effecten van HALT kan hiervoor als illustratie dienen (Ferwerda et al. 2006). Bij dit onderzoek werd gebruik gemaakt van een quasi-experimenteel onderzoeksdesign. Toekenning van jongeren aan de experimentele groep dan wel de controlegroep vond plaats door de onderzoekers. Toekenning aan de experimentele groep betekende voor de jongeren het regulier ondergaan van de HALT-afdoening. Om deze jongeren te stimuleren mee te doen aan de nameting, werd hen € 15 euro in het vooruitzicht gesteld. Toekenning aan de controlegroep vond plaats door jongeren een keus voor te leggen: reguliere HALT-afdoening of meewerken aan het onderzoek, hetgeen betekende dat ze geen HALT-afdoening kregen, maar wel verplicht waren mee te werken aan de voor- en nameting (bij niet nakomen hiervan werd alsnog de reguliere HALT-afdoening ten uitvoer gebracht). Het streven naar verhoging van de interne validiteit blijft hier waarschijnlijk niet zonder gevolgen. Immers, welk effect gaat uit van het feit dat jongeren die een ‘straf’ verwachten te horen krijgen dat ze weer ‘naar huis mogen’? En welk effect heeft het in het vooruitzicht stellen van een financiële vergoeding aan jongeren die een HALT-afdoening krijgen vanwege gedrag dat niet door de beugel kan? Met andere woorden, hoe representatief zijn deze jongeren nog voor de jongeren die een HALT-afdoening ondergaan zonder de genoemde ingrepen? (Rovers 2006).

We zagen hiervoor dat onderzoeksdesigns die hoog scoren op interne validiteit in de praktijk slechts zeer beperkt (kunnen) worden toegepast. Dit gegeven creëert ook een bedreiging voor de generaliseerbaarheid van de resultaten. Immers, generaliseerbaarheid houdt in dat we de onderzoeksresultaten kunnen vertalen naar andere soorten individuen en

groepen, andersoortig probleemgedrag, andere omstandigheden, andersoortige instituties, et cetera. Als we onze kennis baseren op onderzoeksdesigns die slechts in een gering aantal contexten zijn toegepast, hebben we mogelijk een probleem: de contexten waarin de experimentele designs zijn toegepast zouden op belangrijke kenmerken wel eens kunnen afwijken van de contexten waarin deze designs niet zijn (of konden worden) toegepast. Extrapolatie van de onderzoeksresultaten mag in deze omstandigheid geen automatisme zijn. Het vereist op zijn minst –meer dan nu het geval is- problematisering van het fenomeen en nader onderzoek naar de generaliseerbaarheid van onderzoeksresultaten in verschillende contexten.

Ad 3) Nadruk op interne validiteit van designs leidt aandacht af van werkzame stof in interventies

De sterke nadruk op interne validiteit van onderzoeksdesigns, gerelateerd aan het streven om de effectiviteit van interventies zo goed mogelijk te kunnen bepalen, heeft als ongewenst bijeffect dat interventies steeds vaker worden benaderd als een louter input-output gegeven. Dit komt voort uit het op zich gerechtvaardigde streven om verschillen in bijvoorbeeld gedrag vóór en na een interventie vast te kunnen stellen. Deze focus reduceert echter in veel gevallen de interventie zelf tot een *black box*. Wat er in die *black box* gebeurt en hoe dit de uitkomsten beïnvloedt, blijft onduidelijk. Wie de effectstudies beschouwt die in de afgelopen jaren zijn verschenen (zowel primaire studies als research syntheses) kan snel vaststellen dat wat er tijdens de interventie precies gebeurt, in de meeste gevallen geheel buiten beschouwing blijft. Al in een vroeg stadium waarschuwden onder andere onderzoekers van de gerenommeerde ‘Canadese school’ voor dit gevaar (Gendreau & Ross 1979) en later ook wel anderen (zie bijv. Pawson 2002, Merrington & Stanley 2004, Leeuw 2005, Klein Haarhuis et al. 2005), maar waarschijnlijk zien we hier de economische wet van schaarste optreden: een euro kan je maar één keer uitgeven. Als gekozen wordt voor het vaststellen van effectiviteit van interventies door middel van designs die hoog scoren op interne validiteit, zien we in de praktijk van veel onderzoek de interventie zelf uit het design verdwijnen.

Wanneer de interventies zelf buiten beeld blijven, is de praktische waarde van de onderzoeksresultaten mijns inziens gering. Als we vaststellen dat de interventie ‘niet’ werkt, kunnen we niet vaststellen waarom dit het geval is. Omgekeerd geldt hetzelfde voor als de interventie (in bepaalde gevallen) wel blijkt te werken: we weten niet waarom dit het geval is. We kunnen hooguit, en dat gebeurt uiteraard vaak, de onderzoeksresultaten statistisch elaboreren, zodat we weten voor wie het gewerkt heeft of onder welke omstandigheden of bij welke organisaties, maar dit geeft nog steeds geen antwoord op de vraag waarom het juist daar gewerkt heeft en bij andere personen of onder andere omstandigheden niet. Uiteraard tast dit ook de generaliseerbaarheid van de resultaten aan, waarover we hiervoor spraken. Replicatie van succes is hoe dan ook lastiger. Om op dit vlak vooruitgang te boeken, hebben we -veel meer dan nu het geval is- onderzoeksdesigns nodig waarin een expliciete verbinding wordt gelegd tussen wat er in een interventie gebeurt en de uitkomst van die interventie.

Pawson (2002, 2006) heeft ervoor gepleit om interventies te beschouwen als theorieën, zodat we kunnen toetsen welke elementen van de interventie op welke manier verbonden zijn met de uitkomsten ervan. Hij noemt dit realistische evaluatie. Voorzover zijn pleidooi al navolging vindt, gebeurt dit mijns inziens op een manier waarin de potentie van zijn benadering niet volledig tot zijn recht komt. Velen stellen de interventie gelijk aan de gebruikte methodiek en gaan vervolgens na in hoeverre deze gestoeld is op aannames die empirisch houdbaar zijn. Het potentieel van Pawson’s pleidooi is mijns inziens veel groter, doordat het uitgaat van het idee dat de interventie moet worden opgevat als een sociale realiteit waarin individuen en groepen in interactie met elkaar en met andere elementen van de interventie (zoals de gehanteerde methodiek) betekenis verlenen aan wat er gebeurt in de

interventie. Deze processen van betekenisverlening beïnvloeden vervolgens het gedrag van betrokkenen. Hierover kun je toetsbare theorieën opstellen. Door de aandacht te beperken tot methodieken, mis je belangrijke sociale en psychologische processen die bepalend kunnen zijn voor de uitkomst van de interventie, zoals de invloed van professionals die de interventie uitvoeren (Dowden & Andrews 2004, Barkham 2007). Onderzoek naar medische en psychotherapeutische interventies laat bijvoorbeeld zien dat sociale interacties en processen van betekenisverlening door behandelaars en patiënten/cliënten een cruciale rol spelen bij de uitkomst van deze interventies en dat deze factoren in veel gevallen minstens zo belangrijk zijn als de gehanteerde behandeling of methodiek (Rovers 2007, Barkham 2007).

Ad 4) Meta-analyses zijn noodzakelijke maar geen voldoende voorwaarde voor realiseren van effectieve interventies

Analoog aan de toegenomen populariteit van gerandomiseerde veldexperimenten bij primaire studies zien we bij de uitvoering van research syntheses dat onderzoekers steeds vaker kiezen voor het gebruik van meta-analyse. De heldere statistiek die eraan ten grondslag ligt, de objectiverende en transparante procedure en de eenvoudige outputmaten zijn enkele van de aantrekkelijke kenmerken van deze vorm van research synthese.

Naast fervente aanhangers zijn er ook fervente critici van deze methodologie. Pawson (2006) heeft uitvoerig beschreven waarom het gebruik van meta-analyses naar zijn idee niet zal leiden tot effectievere interventies. Zijn kritiek komt er samengevat op neer dat meta-analyses, in tegenstelling tot wat de voorstanders beweren, geen nieuwe kennis opleveren. Volgens hem gaan meta-analyses voorbij aan tal van factoren die relevant zijn voor het succes van interventies, doordat in elk stadium van de analyse noodgedwongen simplificaties en selecties moeten worden gemaakt. Wat de voorstanders zien als één van de sterke punten van de meta-analyse, de samenvatting van allerhande onderzoeksresultaten tot één of enkele kwantitatieve maten, zie Pawson juist als zwakte: de complexe en gelaagde interventiewerkelijkheid wordt 'platgeslagen' tot een ééndimensionale samenvatting waarmee theoretisch noch praktisch iets te doen valt.

Net als aan elke methodologie (en statistische techniek) zijn ook aan het gebruik van meta-analyse voor- en nadelen verbonden, kent het model mogelijkheden en beperkingen, en zijn er aannames waaraan voldaan moet worden voor een verantwoord gebruik (Lipsey & Wilson 2001, Wilson 2001, Davies & Crombie 2003). De voorstanders hebben gelijk wanneer ze beweren dat meta-analyse als methode van research synthese objectiever en transparanter is, dat de resultaten preciezer en beter replicerbaar zijn, dat de wijze waarop onderzoeksresultaten worden samengevat statistisch meer geavanceerd is en dat de methode beter in staat is om verborgen interventie-effecten te vinden. De tegenstanders hebben echter gelijk wanneer ze beweren dat in het veiligheidsveld zelden wordt voldaan aan de aannames die aan het model ten grondslag liggen, dat de geïncludeerde studies vaak te heterogeen zijn om te kunnen vergelijken en dat de methode doorgaans leidt tot op zijn best algemene en gesimplificeerde resultaten (waarbij veel informatie uit de originele studies verloren is gegaan) en op zijn slechtst tot misleidende resultaten (Davies & Crombie 2003, Berk 2005, Pawson 2006).

In navolging van Lipsey & Wilson (2001) pleit ik voor het gebruik van verschillende methoden van research synthese naast elkaar, omdat elk van de methoden pro's en con's kent. Dit betekent een meer genuanceerde kijk op het fenomeen research synthese: de resultaten van meta-analyses zijn niet zaligmakend, en zeker niet voldoende om te komen tot effectieve interventies, maar ze vormen wel een zeer bruikbare en ook noodzakelijke schakel in het proces dat leidt tot effectievere interventies. Hiervoor zijn echter ook andere methoden nodig.

4. Kanttekeningen bij de totstandkoming en toepassing van *What works* beginselen

De onderzoeksresultaten die worden verkregen via de hiervoor beschreven onderzoeksmethoden hebben geleid tot algemene *What works* beginselen. In deze paragraaf wil ik drie kanttekeningen plaatsen bij de totstandkoming en toepassing van deze beginselen:

1. De totstandkoming van de algemene *What works* beginselen is helder noch systematisch
2. De *What works* beginselen zijn weinig bruikbaar voor de interventiepraktijk
3. De focus op *What works* beginselen is niet bevorderlijk voor innovatie

Ad 1) De totstandkoming van algemene What works beginselen is helder noch systematisch
Met name onderzoekers van de 'Canadese school' zijn succesvol geweest in het wereldwijd verspreiden van veel gehanteerde *What works* beginselen, zoals het risicobeginsel, het behoeftebeginsel, et cetera (zie overzicht bij Van der Laan 2004). Je ziet deze beginselen, in rijtjes van wisselende lengte, in tal van publicaties terug. Voor de gemiddelde onderzoeker in het veiligheidsveld beginnen ze inmiddels verdacht veel te lijken op mantra's: als je ze maar vaak genoeg herhaalt, verkrijgen ze vanzelf een zekere geldigheid.

Hoewel veel van de *What works* beginselen nogal *straight forward* zijn en weinig discussie zullen oproepen, wil ik toch een paar kanttekeningen plaatsen bij de wijze waarop we tot deze beginselen zijn gekomen. Ondanks verwoede pogingen ben ik er niet in geslaagd te ontdekken op basis van welke (bij voorkeur heldere) criteria de gevonden onderzoeksresultaten nu zijn vertaald naar deze beginselen? Is dit gebaseerd op een bepaalde effectgrootte? Of op een bepaalde hoeveelheid studies waarin deze factoren een rol hebben gespeeld? En welke onderzoeksresultaten hebben het dan NIET tot beginselen geschopt en waarom niet? Of doen we daar niet aan en worden alle 'significante' factoren vertaald in beginselen en zo ja, moeten we dan op enig moment niet bepaalde beginselen boven andere verkiezen om het rijtje werkbaar te houden? En welke criteria gebruiken we hiervoor? Zijn de studies waar deze beginselen op gebaseerd zijn wel voldoende extern valideerbaar, bijvoorbeeld naar andere (nationale) contexten (Haen Marschall 2005)? En hoe is omgegaan met het feit dat je vanuit probabilistische onderzoeksresultaten komt tot min of meer deterministische (immers: algemeen geldende en toe te passen) beginselen? Ik kan me zo voorstellen dat bepaalde beginselen in bepaalde contexten belangrijker zijn dan in andere.

Ik mis, om kort te gaan, reflectie op dit soort mijns inziens belangrijke kwesties. Wat me opvalt is het gebrek aan systematiek op dit vlak, terwijl de onderzoekers in deze benadering zich juist voor laten staan op methodologische 'hardheid': de gemiddelde effectgrootte van interventies wordt in een meta-analyse achter de komma berekend, maar de vertaalslag hiervan naar beginselen voor goede praktijken lijkt meer op een slag in de lucht. Dit roept tenminste de vraag op hoe bruikbaar deze beginselen nu precies zullen zijn voor de praktijk. Dit is ook het onderwerp van de volgende kanttekening.

Ad 2) De What works beginselen zijn weinig bruikbaar voor de interventiepraktijk
Mijn voornaamste bezwaar tegen de toepassing van de *What works* beginselen is dat deze dermate algemeen zijn dat hun praktische relevantie bijzonder beperkt is. Hier doet zich een dilemma voor dat we niet zomaar kunnen opheffen. Immers, kennis over effectieve interventies vereist dat we onderzoeksresultaten 'stapelen', maar die gestapelde kennis wordt vervolgens zo algemeen, zeker bij het gebruik van meta-analyses, dat deze in specifieke interventiecontexten zijn toepassingskwaliteit verliest. Neem *What works* beginselen als 'responsiviteit' of 'professionaliteit'. De omschrijving van deze beginselen is dermate vaag en algemeen dat er alleen open deuren mee geopend kunnen worden. Ze kunnen als zodanig nauwelijks een sturende functie vervullen voor de praktijk (zie bijv. Merrington & Stanley

2000, Ferguson 2002). De kloof tussen algemene richtlijnen voor effectieve interventies en praktische aanwijzingen die nodig zijn om specifieke interventies succesvol te maken is groot. Deze kloof wordt door de *What works* benadering niet overbrugd, integendeel: de focus ligt hier eerder op het veralgemenen van de onderzoeksresultaten.

Als men in enige interventiepraktijk al een adequate operationaliseringslag kan maken met betrekking tot deze beginselen, stuit men vervolgens op het probleem dat tal van situaties nopen tot afwijken van het algemene beginsel, bijvoorbeeld omdat anders een ander algemeen beginsel in het gedrang komt. Zo vraagt het beginsel ‘programma-integriteit’ om uitvoering van het programma zoals bedoeld, terwijl het beginsel ‘responsiviteit’ erom kan vragen om hiervan af te wijken als bijvoorbeeld blijkt dat bepaalde justitiabelen niet bereikt kunnen worden door de van tevoren bepaalde methode. Responsiviteit vraagt uiteindelijk om individuele afstemming van een interventie op het individu en zijn mogelijkheden. Dit kan impliceren dat moet worden afgeweken van de interventie zoals bedoeld. Hoe hiermee om te gaan?

De beginselen kunnen ook op gespannen voet staan met juridische of andersoortige randvoorwaarden waarbinnen interventies zich (moeten) voltrekken. Veel interventieprogramma’s hebben om juridische, financiële of andersoortige redenen bijvoorbeeld een beperkte doorlooptijd. Dit enkele gegeven kan het al moeilijk of zelfs onmogelijk maken om bepaalde beginselen in de praktijk te brengen.

Het academisch discours over de *What works* beginselen kenmerkt zich soms door onverschilligheid of onwetendheid over de wijze waarop in concrete praktijken vorm kan worden gegeven aan interventies en de praktische beperkingen (zie hiervoor) die hierbij onvermijdelijk aan de orde zijn. Door hiermee géén of onvoldoende rekening te houden formuleert de wetenschap, vanuit de optiek van de praktijk, eerder utopische (wenselijke) dan praktisch bruikbare en toepasbare kennis. Terwijl het streven nu juist is om te komen tot effectievere interventies in de praktijk.

Ad 3) De focus op What works beginselen is niet bevorderlijk voor innovatie

What works beginselen zijn tot stand gekomen door de resultaten van grote hoeveelheden (bestaand) onderzoek te ‘stapelen’ en te vertalen naar algemene beginselen voor *good practices*. Deze vormen weer de basis voor de kwaliteitscriteria die beleidsmakers gebruiken om nieuwe interventies goed te keuren. Deze werkwijze geeft een sterke focus op kennis die in het verleden is gegenereerd en neemt deze als maatstaf voor de toekomst. Hierdoor wordt mijns inziens weinig of geen ruimte gelaten voor innovatie en bovendien wordt bestaande kennis impliciet verabsoluteerd. We mogen aannemen dat onze huidige inzichten in de toekomst plaats zullen maken voor andere, betere inzichten. De hiervoor beschreven benadering is niet bepaald een geschikte randvoorwaarde voor dergelijke innovatie. Innoveren vraagt om risico’s nemen, nieuwe (theoretische) paden durven bewandelen. De *What works* benadering is daarentegen juist geënt op principes als zekerheid en gebaande paden.

5. Bespreking van resultaten

De *What works* benadering legt een zware nadruk op de methodologische kwaliteit van onderzoeksdesigns, meer in het bijzonder op het aspect van de interne validiteit. Daar zijn goede redenen voor, omdat dergelijke designs beter dan andere in staat zijn om vast te stellen of de gepleegde interventie heeft geleid tot de gewenste resultaten. Onderzoek laat bovendien zien dat er een correlatie bestaat tussen kenmerken van het research design en uitkomsten in effectstudies: hoe ‘sterker’ het design des te geringer zijn doorgaans de effecten die worden gevonden (Weisburd et al. 2001). Bovenstaande bespreking laat echter zien dat de medaille

ook een keerzijde heeft. Om het in medische termen te formuleren: mooie designs zijn geen panacee voor alle kwalen. Ze kunnen weliswaar een belangrijke kwaal verhelpen (problemen bij vaststellen van causale effecten van interventies), maar de behandeling is niet zonder bijwerkingen. Deze bijwerkingen kunnen in voorkomende gevallen zelfs erger zijn dan de kwaal (de onderzoeksresultaten zijn bijvoorbeeld niet meer generaliseerbaar). In navolging van Eck (2006) pleit ik dan ook voor een ‘gemengde portfolio van invasieve, beperkt invasieve en niet-invasieve onderzoeksmethoden’ (2006:356-357). Door de hele range van onderzoeksmethoden in te zetten, kunnen we ook de respectieve voordelen benutten. Effectstudies beperken tot grootschalige gerandomiseerde veldexperimenten betekent dat we onszelf veel nuttige kennis ontzeggen.

De nadruk die sommigen tegenwoordig leggen op het belang van goede designs kan doorschieten. Zo constateert Farrington (2006) dat de criteria die de National Research Council in de VS hanteert om interventieprogramma’s in aanmerking te laten komen voor evaluatiestudies (de zogenaamde evalueerbaarheidscriteria) zodanig zijn dat circa 96% van de programma’s niet in aanmerking komt. Slechts 4% van de programma’s leent zich voor een evaluatiestudie die voldoet aan de hoogste methodologische eisen. Farrington, toch geen tegenstander van goede designs, vraagt zich af of het verstandig is om de onderzoeksgelden te beperken tot ‘Rolls-Royce evaluaties van Rolls-Royce programma’s’ en daardoor de overgrote meerderheid van de interventieprogramma’s ongeëvalueerd te laten (2006:330). Voorgaande bespreking toont aan dat er een spanning kan optreden tussen enerzijds het realiseren van research designs die hoog scoren op interne validiteit en anderzijds het streven naar en vormgeven in de praktijk van effectieve(re) interventies.

Dit spanningsveld tussen het realiseren van goed onderzoek enerzijds en het realiseren van effectieve interventies anderzijds is niet onvermijdelijk, maar ligt wel op de loer, zoals ook de volgende twee voorbeelden illustreren. Deze hebben betrekking op het vaststellen van de effectiviteit van ‘kleine interventies’ en de invloed van programma-ontwikkelaars/onderzoekers op de effectiviteit van interventies.

Kenmerkend voor de interventiepraktijk is dat er vaak meerdere interventies tegelijkertijd plaatsvinden, waarbij elke interventie begrensd wordt door juridische, organisatorische, financiële en andere randvoorwaarden, zodat sprake is van een lappendeken van ‘kleine interventies’ die naast elkaar bestaan. Bijvoorbeeld een jongere die een leerstraf ondergaat voor ongeoorloofd spijbelen, heeft vaak ook (al) te maken met andersoortige interventies (op school, thuis, bij de politie, bij de Raad voor de Kinderbescherming, et cetera). Eén van de interventies, bijvoorbeeld de leerstraf, eruit lichten en onderzoeken of deze van invloed is op bijvoorbeeld recidive is in deze situatie een weinig zinvolle strategie. Toch gebeurt dit de hele tijd. Vanuit het streven naar onderzoeksdesigns die hoog scoren op interne validiteit is dit ook goed te begrijpen: je hebt immers een heldere, afgebakende interventie nodig en bij voorkeur random toekenning van justitiabelen aan deze interventie. Deze door onderzoekstechnische redenen ingegeven behoefte gaat echter voorbij aan het nuchtere feit dat de betreffende (kleine) interventie in *the grand scheme of things* hoogstwaarschijnlijk een verwaarloosbaar effect zal hebben op recidive (dit blijkt ook steevast). Er is in de sociale realiteit van de betreffende jongere immers zoveel meer aan de hand dat de recidive kan beïnvloeden, maar voor deze factoren is geen plaats in het design. Rekening houden met deze factoren, en dus met andere interventies, betekent echter dat het experimentele design niet langer gebruikt kan worden. Wat te doen?

Ander voorbeeld: onderzoek toont aan dat er een correlatie bestaat tussen de rol van programma-ontwikkelaars/onderzoekers bij een interventie en de gerapporteerde effecten: hoe meer de onderzoekers betrokken zijn bij de opzet, implementatie en uitvoering van de interventie, des te sterker zijn de gerapporteerde effecten van die interventie. Het gaat om sterke effecten (Petrosino & Soydan 2005, Sherman 2006). Er zijn aanwijzingen dat deze

correlatie veroorzaakt wordt door de grotere betrokkenheid van de betreffende onderzoekers. Deze zou er voor zorgen dat er met meer aandacht, betrokkenheid, gedrevenheid, etc. gewerkt wordt aan de interventie en dit zou zich vervolgens vertalen in een 'betere' interventie. Vanuit de optiek van onderzoek is dit een te vermijden situatie omdat er ongewenste vertekening in de resultaten kan sluipen. Vanuit de optiek van de praktijk echter is het juist interessant om deze verbinding nader te onderzoeken en na te gaan welke factoren hier spelen en op welke wijze deze ingezet kunnen worden teneinde de effecten van de interventie te maximaliseren. Wat te doen?

Het streven naar *evidence-based* interventies is een streven dat ik voluit ondersteun. Ik signaleer tegelijkertijd dat in de huidige *What works* benadering de focus wel erg sterk ligt op het realiseren van 'mooie' designs en ik constateer tegelijkertijd dat dit streven in zijn uitwerking soms haaks kan komen staan op het beoogde doel, het bereiken van effectievere interventies. Hierin zal een middenweg bewandeld moeten worden: de kloof tussen organisaties en professionals die 'maar wat doen' aan de ene kant en wetenschappers die kennis produceren zonder praktijkrelevantie aan de andere kant zal van beide kanten gedicht moeten worden.

Onderzoekers kunnen mijns inziens bijdragen aan effectievere interventies door een paar zaken in acht te nemen. Zij kunnen bijvoorbeeld actiever meedenken over de vertaalslag van wetenschappelijke kennis naar toepassing in interventiepraktijken en hierbij geen afstand nemen van de praktische randvoorwaarden die hierbij aan de orde zijn, deze zijn immers onlosmakelijk met effectiviteit en succes verbonden. Ook is het verstandig als zij het gehele arsenaal aan beschikbare onderzoeksmethodologieën benutten en daaruit de beste kennis destilleren. Effectonderzoek moet niet beperkt worden tot onderzoeksdesigns die hoog scoren op interne validiteit en interventieprogramma's die op deze manier bestudeerd kunnen worden. Deze werkwijze belemmert kennisgroei. Ten slotte is het belangrijk om in effectstudies een expliciete verbinding te leggen tussen inhoud en uitkomst van interventies en binnen de interventies meer aandacht besteden aan sociale interacties en betekenisverlening door betrokkenen actoren. Andere disciplines laten ons zien dat hier winst te boeken valt.

Literatuur

- Asscher, J., Dekovic, M., Laan, P. v. d., Prins, P., & Arum, S. v. (2007). De effectiviteit van multisysteemtherapie (MST). Knelpunten en kansen bij de implementatie van een randomized controlled trial. *Tijdschrift voor Criminologie*, 49(1), 57-67.
- Atkinson, D. (2004). The What Works Debate: Keeping a Human Perspective. *Probation Journal*, 51(3), 248-252.
- Barkham, M. (2007). The role of practitioners in implementing, designing and utilizing practice based evidence. In W. Bruinsma & C. Konijn & M. v. d. Steege (Eds.), *Jeugdzorg in onderzoek: De professional en effectiviteit* (pp. 41-54). Den Haag: Ministerie van VWS.
- Beenackers, E. M. T. (2000). *Effectiviteit van sanctieprogramma's: bouwstenen voor een toetsingskader : een literatuuronderzoek* (WODC onderzoeksnotities ; 2000/1). Den Haag: Wetenschappelijk Onderzoek- en Documentatiecentrum.
- Berghuis, B., Leeuw, F., & Scheepmaker, M. (2004). Beleidstheorieën - Voorwoord. *Justitiële Verkenningen*, 30(5), 5-9.
- Berk, R. E. (2005). Randomized experiments as the bronze standard. *Journal of Experimental Criminology*, 1, 417-433.
- Cronbach, L. J. (1982). *Designing evaluation of educational and social programs*. San Fransisco: Jossey-Bass.
- Cullen, F. T., & Gendreau, P. (2000). Assessing Correctional Rehabilitation: Policy, Practice, and Prospects. In J. Horney (Ed.), *Processes, and Decisions of the Criminal Justice System; Criminal Justice 2000* (Vol. 3, pp. 109-175). Rockville, MD: National Institute of Justice/NCJRS.
- Davies, H. T. O., & Crombie, I. K. (2007). *What is meta-analysis?* Hayward Medical Communications

- (download op: <http://www.evidence-based-medicine.co.uk/ebmfiles/WhatisMetaAn.pdf> [2007, 13-7-2007].
- Dowden, G., & Andrews, D. A. (2004). The Importance of Staff Practice in Delivering Correctional Treatment: A Meta-Analytic Review of Core Correctional Practice. *International Journal of Offender Therapy and Comparative Criminology*, 48(2), 203-214.
- Eck, J. E. (2006). When is a Bologna sandwich better than sex? A defense of small n-case study evaluations. *Journal of Experimental Criminology*, 2, 345-362.
- Erkenningscommissie-Gedragsinterventies. (2006). *Jaarverslag 2006*. Den Haag: Ministerie van Justitie.
- Farrington, D. P., Gottfredson, D. C., Sherman, L. W., & Welsh, B. C. (2002). The Maryland Scientific Methods Scale. In L. W. Sherman & D. P. Farrington & B. C. Welsh & D. L. MacKenzie (Eds.), *Evidence-Based Crime Prevention*. New York: 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, 9-38.
- Farrington, D. P. (2006). Methodological quality and the evaluation of anti-crime programs. *Journal of Experimental Criminology*, 2, 329-337.
- Feder, L., Jolin, A., & Feyerherm, W. (2000). Lessons from two randomized experiments in criminal justice settings. *Crime & Delinquency*, 46(3), 380-400.
- Ferguson, J. L. (2002). Putting the "What works" Research into Practice: An Organizational Perspective. *Criminal Justice and Behavior*, 29(4), 472-492.
- Ferwerda, H. B., Leiden, I. M. G. G. v., Arts, N. A. M., & Hauber, A. R. (2006). *Halt: het Alternatief? De effecten van HALT beschreven (reeks Onderzoek en Beleid, nr 244)*. Den Haag/Meppel: Wetenschappelijk Onderzoek- en Documentatiecentrum (WODC) / Boom Juridische uitgevers.
- Gendreau, P., & Ross, R. R. (1979). Effective correctional treatment: Bibliotherapy for cynics. *Crime & Delinquency*, 25(October), 463-489.
- Haen-Marshall, I. (2005). Introduction: International Transferability of the practice of evidence-based crime policies. *European Journal on Criminal Policy and Research*, 11, 225-231.
- Harrington, A. (2002). "Seeing" the Placebo-effect: Historical Legacies and Present Opportunities. In H. A. Guess & A. Kleinman & J. W. Kusek & L. Engel, W. (Eds.), *The Science of the Placebo; Toward an Interdisciplinary Research Agenda* (pp. 35-52). London UK: BMJ Books.
- Klein-Haarhuis, C., Ooyen-Houben, M. v., Kleemans, E., & Leeuw, F. (2005). Rechtshandhaving geevalueerd: een synthese van 31 onderzoeken. *Justitiele Verkenningen*, 31(8), 54-71.
- Laan, P. H. v. d. (2004). Over straffen, effectiviteit en erkenning; de wetenschappelijke onderbouwing van preventie en strafrechtelijke interventie. *Justitiele Verkenningen*, 30(5), 31-48.
- Laan, P. H. v. d., Krooi, H., Voort, P. v. d., & Wijkman, M. (2006). *Toetsing en certificering van preventieve, curatieve en repressieve interventies in de jeugdzorg*. Leiden: NSCR.
- Leeuw, F. L. (2005). Trends and Developments in Program Evaluations in General and Criminal Justice Programs in Particular. *European Journal on Criminal Policy and Research*, 11, 233-258.
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical Meta-Analysis*. Thousand Oaks [etc.]: Sage.
- Lipsey, M. W. (2001). Unsolved problems and unfinished business. *American Journal of Evaluation*, 22(3), 325-328.
- Lipton, D., Martinson, R., & Wilks, J. (1975). *The effectiveness of correctional treatment: A survey of treatment evaluation studies*. New York: Praeger.
- Lum, C., & Yang, S.-M. (2005). Why do evaluation researchers in crime and justice choose non-experimental methods? *Journal of Experimental Criminology*, 1, 191-213.
- Martinson, R. (1974). What works? Questions and answers about prison reform. *The Public Interest*, 35(Spring), 22-54.
- Merrington, S., & Stanley, S. (2004). What Works? Revisiting the Evidence in England and Wales. *Probation Journal*, 51(1), 7-20.
- National-Research-Council. (2005). *Improving evaluation of anticrime programs*. Washington D.C.: National Academies Press.
- Pawson, R. (2002). Evidence-based Policy: The promise of 'Realist Synthesis'. *Evaluation*, 8(3), 340-358.
- Pawson, R. (2006). *Evidence-based Policy. A Realist Perspective*. London [etc.]: Sage.
- Petrosino, A., Boruch, R. F., Soydan, H., Duggan, L., & Sanchez-Meca, J. (2001). Meeting the Challenges of Evidence-Based Policy: The Campbell Collaboration. *The Annals of the American Academy of Political and Social Research*, 578(November), 14-34.
- Petrosino, A., & Soydan, H. (2005). The impact of program developers as evaluators on criminal recidivism: Results from meta-analyses of experimental and quasi-experimental research. *Journal of Experimental Criminology*, 1, 435-450.
- Raynor, P. (2003). Evidence-Based Probation and its Critics. *Probation Journal*, 50(4), 334-345.
- Roos, T. d. (2007). Experimenteren met sancties? Strafrechtstheoretische kanttekeningen bij een empirisch-

- wetenschappelijke trend. *Tijdschrift voor Criminologie*, 49(1), 12-23.
- Rovers, B. (2006). Boekbespreking: Ferwerda, Henk B., Leiden, I.M.G.G. van, Arts, N.A.M. en A.R. Hauber (2006). *Halt: Het alternatief? De effecten van Halt beschreven*, Den Haag: Boom Juridische Uitgevers / WODC (Onderzoek en Beleid, deel 244). *Tijdschrift voor Veiligheid*, 5(4), 74-76.
- Rovers, B. (2007, 14 juni). *Preventie van jeugdcriminaliteit en de placeborespons; Toepassing van het belief effect in het veiligheidsveld*. Paper presented at the NVK-Marktdag, Leiden.
- Sherman, L. W., Gottfredson, D., MacKenzie, D., Eck, J., Reuter, P., & Bushway, S. (1997). *Preventing crime: What works, what doesn't, what's promising? A Report to the United States Congress*. Washington D.C.: Office of Justice Programs, U.S. Department of Justice.
- Sherman, L. W. (2006). "To develop and test": The inventive difference between evaluation and experimentation. *Journal of Experimental Criminology*, 2, 393-406.
- Wartna, B. S. J. (2005). *Evaluatie van daderprogramma's : een wegwijzer voor onderzoek naar de effecten van strafrechtelijke interventies speciaal gericht op het terugdringen van recidive*. Den Haag: Wetenschappelijk onderzoek- en documentatiecentrum (WODC) / Boom.
- Weisburd, D., Lum, C. M., & Petrosino, A. (2001). Does Research Design Affect Study Outcomes in Criminal Justice? *The Annals of the American Academy of Political and Social Research*, 578, 50-70.
- Welsh, B. C., & Farrington, D. P. (2001). Towards an Evidence-Based Approach to Preventing Crime. *The Annals of the American Academy of Political and Social Research*, 578, 158-173.
- Wilson, D. B. (2001). Meta-Analytic Methods for Criminology. *The Annals of the American Academy of Political and Social Research*, 578, 71-89.

Summary

‘What works’ has become an important research tradition within Criminology. It aims at more effective interventions by using only research designs that satisfy the highest methodological criteria (particularly randomized experimental designs), by focusing on quantitative research syntheses (by means of meta-analyses) and by focusing on the use of empirically proven methodologies in interventions. The article argues that this approach will not necessarily lead to more effective interventions.

Seven short comments are given on this approach. They refer to the research methodology (1-4) and the construction and use of the ‘What works’ principles (5-7). In short: 1) The potential use of randomized experimental designs in the Justice field is limited by various factors, some of which are more or less structural. 2) The strive for higher internal validity comes with a price, i.e. lower external validity (less possibilities to generalize results to other contexts). 3) The focus on internal validity, combined with the need to establish the effectiveness of interventions, has in practice reduced the intervention itself to a black box. As a result we don’t have a clue about the causal factors within interventions. 4) The role of meta-analyses in realizing more effective interventions is more limited than some think or hope. 5) The way in which research results are translated into ‘What works’ principles is not very clear and far from systematic. 6) The ‘What works’ principles have very limited practical value. They are too general for practical guidance and following different principles may in practice lead to opposing intervention strategies. 7) The use of the ‘What works’ principles inhibits innovative ideas, theories and interventions. The focus is on existing knowledge, implicitly making this knowledge absolute.

It is concluded that the approach taken by many researchers in the ‘What works’ tradition is adding to our knowledge, but at the same time the limitations of this approach are clear. The author makes some suggestions for future research on interventions.

Rovers, B. 2007. "What works; kanttekeningen bij een populair programma." *Tijdschrift voor Veiligheid*, (6) 3, 7-22.

What works; kanttekeningen bij een populair programma

Voor veel onderzoekers en beleidmakers in het justitieveld staat de *What works* benadering voor het streven naar effectieve(re) interventies door middel van onderzoek dat voldoet aan de hoogste methodologische eisen (gerandomiseerde veldexperimenten), door gebruikmaking van kwantitatieve research syntheses (door middel van meta-analyses) en door binnen de interventies te focussen op empirisch onderbouwde methodieken. In dit artikel worden kanttekeningen geplaatst bij deze benadering. Deze kanttekeningen betreffen niet de achterliggende gedachte van *evidence-based* werken, maar de uitwerking die hieraan wordt gegeven. Beoogd wordt dat de hiervoor genoemde aanpak niet in alle gevallen zal leiden tot effectieve(re) interventies. In voorkomende gevallen kan deze aanpak in zijn uitwerking zelfs haaks komen staan op het beoogde doel (effectievere interventies). De auteur doet enkele suggesties voor toekomstig onderzoek naar interventies.

1. Inleiding

Gedragregulering vindt op verschillende manieren plaats. Persoonsgerichte interventies in het veiligheidsveld kunnen worden beschouwd als één van de mogelijke vormen. Deze interventies zijn meestal gericht op het voorkómen van crimineel gedrag of het verminderen van recidive. De wijze waarop deze doelstellingen worden nagestreefd kan sterk variëren, al naar gelang de context waarin de interventie plaatsvindt (binnen of buiten het justitiedomein); zo kan een school een antispjbelprogramma ontwikkelen, een buurthuis kan activiteiten opzetten voor overlastgevendende jongeren, een hulpverleningsorganisatie kan gezinstherapie aanbieden, een inrichting voor jeugdige gedetineerden leert de jongeren vaardigheden aan, een reclasseringsorganisatie biedt nazorg voor gedetineerden, et cetera. Deze interventies hebben gemeen dat ze (mede) beogen de betrokkenen uit de criminaliteit te houden. De mate waarin deze interventies succesvol zijn, is onderwerp van de *What works* benadering.

In het denken over *evidence-based* interventies in het veiligheidsveld neemt de *What works* benadering een belangrijke plaats in. Het is lastig om deze benadering scherp af te bakenen van andere benaderingen, maar we kunnen wel enkele contouren schetsen die we hierna zullen bespreken. De kanttekeningen die ik wil plaatsen bij deze benadering betreffen niet de achterliggende gedachte van *evidence-based* werken, maar de wijze waarop hieraan vorm wordt gegeven door veel onderzoekers en ook beleidmakers, en de ongewenste effecten die hiervan het gevolg kunnen zijn, namelijk dat de onderzoeksinspanningen niet leiden tot het gewenste doel: effectievere interventies.

In paragraaf 2 zal ik kort de achtergronden van de *What works* benadering schetsen en de belangrijkste constituerende elementen beschrijven. Daarna zal ik in de paragrafen 3 en 4 aangeven welke kanttekeningen er zoal te plaatsen zijn bij deze benadering, waarna ik in paragraaf 5 de resultaten zal bespreken en besluit met enkele suggesties voor verder onderzoek.

2. What works – een korte introductie

Het streven naar effectieve interventies is van alle tijden. De wijze waarop hieraan vorm wordt gegeven verschilt naar tijd en plaats en wellicht ook per discipline. Beoefenaren van sociaal-wetenschappelijke disciplines kijken tegenwoordig graag naar de medische sector waar al meer dan 50 jaar geleden het gebruik van *Randomized Controlled Trials* (RCT's) de

gouden standaard werd voor het vaststellen van effecten van medische interventies (Petrosini et al. 2001, Harrington 2002).

In het veiligheidsveld is het denken over *evidence-based* interventies van jongere datum (Farrington & Welsh 2005). De publicaties van Martinson (1974) en Lipton, Martinson en Wilks (1975) figureren in historische overzichten vaak als een soort startpunt voor beschouwingen over de *What Works* benadering. Deze auteurs kwamen na evaluatie van tal van effectstudies tot de pessimistische conclusie dat, op enkele geïsoleerde uitzonderingen na, de pogingen om regelovertreeders te rehabiliteren mislukt waren: men was niet in staat gebleken de recidive van betrokkenen omlaag te krijgen. Deze conclusie is later kortweg samengevat als: *Nothing works*. Een conclusie die door veel criminologen nadien voor waar werd aangenomen zonder zelf aanvullend onderzoek te doen (Cullen & Gendreau 2000).

Diverse ontwikkelingen hebben er voor gezorgd dat het vraagstuk van *evidence-based* werken in het veiligheidsveld toch weer op de agenda is gekomen. De ontwikkeling van nieuwe onderzoeksmethoden, met name de opkomst van meta-analyse, heeft hier waarschijnlijk een belangrijk aandeel in gehad. Ook het feit dat overheden in toenemende mate kritische kanttekeningen zijn gaan plaatsen bij de financiering van allerlei (bij voorbaat zinloze?) interventies, heeft hierbij een rol gespeeld. De positievere uitkomsten van onderzoek naar rehabilitatieprogramma's door een groep Canadese onderzoekers (Cullen & Gendreau 2000) en later ook de meer genuanceerde resultaten van grootschalige research syntheses door Sherman et al. (1997), in opdracht van het Amerikaanse Congres, hebben eraan bijgedragen dat vandaag de dag met meer positivisme wordt aangekeken tegen de mogelijkheden van persoonsgerichte preventie- en rehabilitatieprogramma's voor (jonge) regelovertreeders. De oprichting, in 2000, van de 'Campbell Collaboration' (een forum voor het bevorderen van *evidence-based* kennis) bevestigt het hernieuwde geloof in de mogelijkheid om door middel van hoogwaardig onderzoek te komen tot verbeterde inzichten in de werkzaamheid van justitiële interventies (<http://www.campbellcollaboration.org/>).

Deze ontwikkeling zien we ook terug in het beleid. Beleidmakers zijn hogere eisen gaan stellen aan interventieprogramma's en aan onderzoek dat deze programma's moet toetsen dan wel rechtvaardigen (Beenackers 2000, Wartna 2005, National Research Council 2005, Van der Laan et al. 2006). In 2005 werd in Nederland de Erkenningscommissie Gedragsinterventies geïnstalleerd. Deze commissie adviseert het Ministerie van Justitie over de kwaliteit van justitiële interventies. Dit doet zij op basis van *What works* beginselen. Tot en met eind 2006 werden aan deze commissie 13 interventies ter toetsing voorgelegd. Zes daarvan werden voorlopig erkend (dit betekent dat de interventie niet aan alle kwaliteitscriteria voldoet, maar voldoende potentieel heeft), de overige zeven werden niet erkend. Géén van de dertien voorgelegde interventies werd volledig erkend, dat wil zeggen: had voldaan aan alle kwaliteitscriteria (Erkenningscommissie Gedragsinterventies 2006).¹

Hoe stel je nu vast of interventies werken? Het antwoord op deze vraag brengt ons bij wat ik de constituerende elementen van de *What works* benadering zou willen noemen. Zoals in de inleiding reeds aangegeven, is het lastig om deze benadering scherp af te bakenen van andere benaderingen, omdat er geen officieel protocol bestaat. *What works* staat voor een aantal uitgangspunten die we vaak gehanteerd zien, maar waarover geen consensus hoeft te bestaan bij de betrokkenen. De hierna omschreven elementen zien we vaak terug in de *What works* benadering, in het bijzonder bij onderzoekers en beleidmakers die zich associëren met de uitgangspunten van de Campbell Collaboration. In mijn optiek zijn de volgende drie onderdelen hierbij cruciaal:

¹ Recentelijk is ook een landelijke Erkenningscommissie voor Jeugdinterventies van start gegaan (bron: www.nji.nl).

1. de effecten van interventies zijn onderzocht aan de hand van *research designs die zo hoog mogelijk scoren op interne validiteit*;
2. Om te kunnen spreken over bewezen effecten beschikt men bij voorkeur over *een zo groot mogelijke hoeveelheid data* die verkregen is middels de hiervoor genoemde designs;
3. Binnen de interventies is de aandacht vooral gericht op *bewezen methodieken*.

Toelichting

Om de methodologische kwaliteit van onderzoek vast te stellen, met name op het punt van de interne validiteit, zijn diverse criteria en schalen ontwikkeld. De *Maryland Scientific Methods Scale* (SMS) is wellicht de bekendste in het veiligheidsveld (Farrington et al. 2002). Op een 5-punts schaal wordt weergegeven in hoeverre uitkomstvariabelen kunnen worden toegeschreven aan de gepleegde interventie. Bij een niveau-1 studie gebeurt dit op basis van correlaties tussen interventievariabelen en uitkomstvariabelen, bij een niveau-5 studie (het hoogste niveau) worden effecten vastgesteld door gebruik te maken van een experimentele opzet (vergelijking tussen experimentele en controlegroep), het uitvoeren van voor- en nametingen in beide groepen, controle voor storende variabelen en gerandomiseerde toekenning van de interventie. Er wordt gestreefd naar studies van een zo hoog mogelijk niveau, omdat deze beter in staat zijn de causale effecten van interventies vast te stellen.

Het tweede element, het streven naar empirische validering door gebruikmaking van zoveel mogelijk data, zien we vooral terug in de grote hoeveelheid research syntheses die in de voorbije jaren geproduceerd zijn. Deze studies zijn gericht op het vaststellen van effecten van interventies door de resultaten uit (zoveel mogelijk) verschillende onderzoeken 'te stapelen'. Er zijn verschillende methoden om een research synthese uit te voeren en net als voor de studies die onderwerp zijn van een research synthese, geldt ook voor de research synthese zelf dat er een ordening kan worden aangebracht in de methodologische kwaliteit. *Narrative review*, de *vote count* methode en meta-analyse zijn veel gebruikte methoden om onderzoeksresultaten samen te vatten (Welsh & Farrington 2001). Met name meta-analyses zijn de laatste jaren zeer populair geworden en vormen een belangrijke peiler onder de *What works* benadering. Belangrijke kenmerken zijn: een uitputtende zoektocht naar relevante studies, expliciete inclusie- en exclusiecriteria voor studies, het gebruik van een expliciet codeerschema om studiekekenmerken en resultaten te beschrijven en het gebruik van de beste statistische methoden om de studieresultaten samen te vatten in één simpele kwantitatieve maat; de gemiddelde effectgrootte (Lipsey & Wilson 2001, Petrosini et al. 2001, Wilson 2001). Meta-analyse wordt door velen gezien als een vorm van research synthese die in methodologisch opzicht superieur is aan de andere vormen en daarom te verkiezen.

De voorgaande elementen hebben betrekking op *hoe* je de effecten van interventies het beste kan vaststellen. Het laatste element zoomt in op de inhoud van de interventie zelf. We zien hier een sterke focus op 'bewezen' methodieken. Dit zijn methodieken waaraan een geëxpliciteerde én empirisch gefundeerde beleidstheorie ten grondslag ligt. Een beleidstheorie is een theorie die verklaart of voorspelt hoe de interventie zal uitpakken (Leeuw 2005). Explicitering van de theoretische aannames die aan een interventie ten grondslag liggen en toetsing van deze aannames aan het beschikbare empirisch materiaal wordt gezien als een belangrijke en noodzakelijke stap naar effectieve(re) interventies (Berghuis et al. 2004, Leeuw 2005).

Op basis van het voorgaande laten de uitgangspunten van de *What works* benadering zich als volgt samenvatten: om te kunnen bepalen of interventies werken, hebben we onderzoek nodig. Niet zomaar onderzoek, maar onderzoek dat voldoet aan de hoogste criteria van interne validiteit. Hoe meer van dit soort onderzoek we tot onze beschikking hebben (dus: hoe groter

onze database), des te gefundeerder ons oordeel over de effectiviteit van deze interventies. Meta-analyses zijn volgens velen het beste middel om dit oordeel te vellen. Bij het beoordelen van de werkzame stof in interventies staan bewezen methodieken centraal. Dit zijn methodieken die zijn gebaseerd op beleidstheorieën die voldoende empirisch onderbouwd zijn.

Hierna zal ik mijn kanttekeningen plaatsen bij deze benadering. Deze betreffen enerzijds de gebruikte onderzoeksmethodologie en anderzijds de totstandkoming en toepassing van de *What works* beginselen.

3. Kanttekeningen bij de onderzoeksmethodologie

In de *What works* benadering ligt een sterke nadruk op de methodologische kwaliteit van de gehanteerde onderzoeksdesigns. Bij primaire studies gaat het dan vooral om gerandomiseerde (veld)experimenten en designs die dit model benaderen. Bij research syntheses gaat het vooral om meta-analyses. Bij het gebruik van deze designs wil ik de volgende vier kanttekeningen plaatsen, die ik aansluitend zal uitwerken:

1. Onderzoeksdesigns die hoog scoren op interne validiteit zijn beperkt toepasbaar
2. Hoge interne validiteit gaat vaak ten koste van externe validiteit
3. Nadruk op interne validiteit van designs leidt aandacht af van werkzame stof in interventies
4. Meta-analyses zijn noodzakelijke maar geen voldoende voorwaarde voor realiseren van effectieve interventies

Ad 1) Onderzoeksdesigns die hoog scoren op interne validiteit zijn beperkt toepasbaar

Het aantal studies in het veiligheidsveld dat hoog scoort op de SMS-schaal is –naar verhouding- gering (Lipsey 2001, Farrington & Welsh 2005). Dit is echter geen natuurverschijnsel of een situatie die voortvloeit uit het onwaarschijnlijke gegeven dat onderzoekers ‘slechte designs’ prefereren. Verschillende factoren beperken de mogelijkheden om dit soort designs te gebruiken.

Een belangrijke factor vormen de juridische randvoorwaarden waarbinnen (justitiële) interventies worden uitgevoerd. Experimenteren met interventies kan op gespannen voet staan met juridische beginselen die ten grondslag liggen aan diezelfde interventies, zoals het uitgangspunt dat gelijke gevallen gelijk behandeld moeten worden. Het juridisch kader begrenst aldus de mogelijkheden om vrijelijk te experimenteren (De Roos 2007). In het verlengde daarvan liggen ook ethisch-normatieve argumenten die pleiten tegen willekeurige toekenning van respondenten aan experimentele en controlecondities. Een groot verschil met de medische sector is hier dat de interventies zelf moreel geladen zijn.²

Behalve juridische beperkingen zijn er praktische beperkingen aan het gebruik van deze designs. Deze beperkingen kunnen zich op tal van vlakken voordoen en betreffen zowel de implementatie als de uitvoering (Berk 2005). Hierbij gaat het om zaken als het creëren van goede experimentele condities (randomisering etc.), voldoende deelnemers in beide onderzoekscondities (experimentele en controlegroep), een programma dat lang genoeg loopt om het te kunnen evalueren, de financiële middelen om dit te realiseren, de medewerking van organisaties, professionals en justitiabelen, et cetera. Het succesvol uitvoeren van een

² Overigens roept de toepassing van justitiële interventies waarvan het effect niet bewezen is, ook morele vragen op.

gerandomiseerd veldexperiment vergt een protocol waaraan strak de hand wordt gehouden. Dit blijkt vaak moeilijk te realiseren; in de praktijk ontstaan –om allerlei redenen- vaak afwijkingen van ‘de behandeling zoals gepland’. Weerstand van medewerkers, maar ook onbewuste schendingen van het protocol, vormen veel voorkomende bedreigingen voor het onderzoek (Feder et al. 2000, Raynor 2003, Atkinson, 2004, Berk 2005, Asscher et al. 2007). Zelfs Lipsey (2001), een fervent voorstander van het gebruik van experimentele designs, constateert dat de bronnen en omstandigheden die nodig zijn voor het realiseren van deze designs in de praktijk nogal schaars zijn.

Verschillende auteurs hebben gewezen op het belang van de beleidscontext en hoe deze het gebruik van experimentele designs door onderzoekers beïnvloedt. Pawson (2002) heeft in dit verband gewezen op het feit dat reguliere beleidscycli op gespannen voet staan met het gebruik van experimentele onderzoeksdesigns: beleidsmakers willen resultaten op korte termijn, terwijl onderzoek door middel van experimentele designs vraagt om een lange-termijn blik. Lum & Yang (2005) stellen dat de sterke beleidsoriëntatie van de criminologie ervoor heeft gezorgd dat in deze discipline minder experimenteel onderzoek wordt gedaan dan in andere sociaal-wetenschappelijke disciplines. Men volgt vooral de beleidsagenda.

Samenvattend: de toepassingsmogelijkheden van experimentele designs in het veiligheidsveld worden door verschillende factoren beperkt. Sommige van deze factoren hebben een min of meer structureel karakter.

Ad 2) Hoge interne validiteit gaat vaak ten koste van externe validiteit

Cronbach (1982) stelde zichzelf de vraag ‘Wat heb je aan onderzoeksresultaten als je ze niet kan generaliseren?’ Zijn antwoord laat zich raden.

Randomisering bij onderzoek in het veiligheidsveld stuit op allerhande juridische, praktische en beleidsmatige beperkingen. Echter, wanneer we ondanks al deze beperkingen toch in staat zijn om randomisering toe te passen, zijn we nog niet uit de problemen. Randomisering kan namelijk de natuurlijke situatie aantasten, waardoor de winst die aan de ene kant wordt geboekt (hogere interne validiteit) aan de andere kant weer moet worden afgeboekt (geringere generaliseerbaarheid van de onderzoeksresultaten).

Recent onderzoek in Nederland naar de effecten van HALT kan hiervoor als illustratie dienen (Ferwerda et al. 2006). Bij dit onderzoek werd gebruik gemaakt van een quasi-experimenteel onderzoeksdesign. Toekenning van jongeren aan de experimentele groep dan wel de controlegroep vond plaats door de onderzoekers. Toekenning aan de experimentele groep betekende voor de jongeren het regulier ondergaan van de HALT-afdoening. Om deze jongeren te stimuleren mee te doen aan de nameting, werd hen € 15 euro in het vooruitzicht gesteld. Toekenning aan de controlegroep vond plaats door jongeren een keus voor te leggen: reguliere HALT-afdoening of meewerken aan het onderzoek, hetgeen betekende dat ze geen HALT-afdoening kregen, maar wel verplicht waren mee te werken aan de voor- en nameting (bij niet nakomen hiervan werd alsnog de reguliere HALT-afdoening ten uitvoer gebracht). Het streven naar verhoging van de interne validiteit blijft hier waarschijnlijk niet zonder gevolgen. Immers, welk effect gaat uit van het feit dat jongeren die een ‘straf’ verwachten te horen krijgen dat ze weer ‘naar huis mogen’? En welk effect heeft het in het vooruitzicht stellen van een financiële vergoeding aan jongeren die een HALT-afdoening krijgen vanwege gedrag dat niet door de beugel kan? Met andere woorden, hoe representatief zijn deze jongeren nog voor de jongeren die een HALT-afdoening ondergaan zonder de genoemde ingrepen? (Rovers 2006).

We zagen hiervoor dat onderzoeksdesigns die hoog scoren op interne validiteit in de praktijk slechts zeer beperkt (kunnen) worden toegepast. Dit gegeven creëert ook een bedreiging voor de generaliseerbaarheid van de resultaten. Immers, generaliseerbaarheid houdt in dat we de onderzoeksresultaten kunnen vertalen naar andere soorten individuen en

groepen, andersoortig probleemgedrag, andere omstandigheden, andersoortige instituties, et cetera. Als we onze kennis baseren op onderzoeksdesigns die slechts in een gering aantal contexten zijn toegepast, hebben we mogelijk een probleem: de contexten waarin de experimentele designs zijn toegepast zouden op belangrijke kenmerken wel eens kunnen afwijken van de contexten waarin deze designs niet zijn (of konden worden) toegepast. Extrapolatie van de onderzoeksresultaten mag in deze omstandigheid geen automatisme zijn. Het vereist op zijn minst –meer dan nu het geval is- problematisering van het fenomeen en nader onderzoek naar de generaliseerbaarheid van onderzoeksresultaten in verschillende contexten.

Ad 3) Nadruk op interne validiteit van designs leidt aandacht af van werkzame stof in interventies

De sterke nadruk op interne validiteit van onderzoeksdesigns, gerelateerd aan het streven om de effectiviteit van interventies zo goed mogelijk te kunnen bepalen, heeft als ongewenst bijeffect dat interventies steeds vaker worden benaderd als een louter input-output gegeven. Dit komt voort uit het op zich gerechtvaardigde streven om verschillen in bijvoorbeeld gedrag vóór en na een interventie vast te kunnen stellen. Deze focus reduceert echter in veel gevallen de interventie zelf tot een *black box*. Wat er in die *black box* gebeurt en hoe dit de uitkomsten beïnvloedt, blijft onduidelijk. Wie de effectstudies beschouwt die in de afgelopen jaren zijn verschenen (zowel primaire studies als research syntheses) kan snel vaststellen dat wat er tijdens de interventie precies gebeurt, in de meeste gevallen geheel buiten beschouwing blijft. Al in een vroeg stadium waarschuwden onder andere onderzoekers van de gerenommeerde ‘Canadese school’ voor dit gevaar (Gendreau & Ross 1979) en later ook wel anderen (zie bijv. Pawson 2002, Merrington & Stanley 2004, Leeuw 2005, Klein Haarhuis et al. 2005), maar waarschijnlijk zien we hier de economische wet van schaarste optreden: een euro kan je maar één keer uitgeven. Als gekozen wordt voor het vaststellen van effectiviteit van interventies door middel van designs die hoog scoren op interne validiteit, zien we in de praktijk van veel onderzoek de interventie zelf uit het design verdwijnen.

Wanneer de interventies zelf buiten beeld blijven, is de praktische waarde van de onderzoeksresultaten mijns inziens gering. Als we vaststellen dat de interventie ‘niet’ werkt, kunnen we niet vaststellen waarom dit het geval is. Omgekeerd geldt hetzelfde voor als de interventie (in bepaalde gevallen) wel blijkt te werken: we weten niet waarom dit het geval is. We kunnen hooguit, en dat gebeurt uiteraard vaak, de onderzoeksresultaten statistisch elaboreren, zodat we weten voor wie het gewerkt heeft of onder welke omstandigheden of bij welke organisaties, maar dit geeft nog steeds geen antwoord op de vraag waarom het juist daar gewerkt heeft en bij andere personen of onder andere omstandigheden niet. Uiteraard tast dit ook de generaliseerbaarheid van de resultaten aan, waarover we hiervoor spraken. Replicatie van succes is hoe dan ook lastiger. Om op dit vlak vooruitgang te boeken, hebben we -veel meer dan nu het geval is- onderzoeksdesigns nodig waarin een expliciete verbinding wordt gelegd tussen wat er in een interventie gebeurt en de uitkomst van die interventie.

Pawson (2002, 2006) heeft ervoor gepleit om interventies te beschouwen als theorieën, zodat we kunnen toetsen welke elementen van de interventie op welke manier verbonden zijn met de uitkomsten ervan. Hij noemt dit realistische evaluatie. Voorzover zijn pleidooi al navolging vindt, gebeurt dit mijns inziens op een manier waarin de potentie van zijn benadering niet volledig tot zijn recht komt. Velen stellen de interventie gelijk aan de gebruikte methodiek en gaan vervolgens na in hoeverre deze gestoeld is op aannames die empirisch houdbaar zijn. Het potentieel van Pawson’s pleidooi is mijns inziens veel groter, doordat het uitgaat van het idee dat de interventie moet worden opgevat als een sociale realiteit waarin individuen en groepen in interactie met elkaar en met andere elementen van de interventie (zoals de gehanteerde methodiek) betekenis verlenen aan wat er gebeurt in de

interventie. Deze processen van betekenisverlening beïnvloeden vervolgens het gedrag van betrokkenen. Hierover kun je toetsbare theorieën opstellen. Door de aandacht te beperken tot methodieken, mis je belangrijke sociale en psychologische processen die bepalend kunnen zijn voor de uitkomst van de interventie, zoals de invloed van professionals die de interventie uitvoeren (Dowden & Andrews 2004, Barkham 2007). Onderzoek naar medische en psychotherapeutische interventies laat bijvoorbeeld zien dat sociale interacties en processen van betekenisverlening door behandelaars en patiënten/cliënten een cruciale rol spelen bij de uitkomst van deze interventies en dat deze factoren in veel gevallen minstens zo belangrijk zijn als de gehanteerde behandeling of methodiek (Rovers 2007, Barkham 2007).

Ad 4) Meta-analyses zijn noodzakelijke maar geen voldoende voorwaarde voor realiseren van effectieve interventies

Analoog aan de toegenomen populariteit van gerandomiseerde veldexperimenten bij primaire studies zien we bij de uitvoering van research syntheses dat onderzoekers steeds vaker kiezen voor het gebruik van meta-analyse. De heldere statistiek die eraan ten grondslag ligt, de objectiverende en transparante procedure en de eenvoudige outputmaten zijn enkele van de aantrekkelijke kenmerken van deze vorm van research synthese.

Naast fervente aanhangers zijn er ook fervente critici van deze methodologie. Pawson (2006) heeft uitvoerig beschreven waarom het gebruik van meta-analyses naar zijn idee niet zal leiden tot effectievere interventies. Zijn kritiek komt er samengevat op neer dat meta-analyses, in tegenstelling tot wat de voorstanders beweren, geen nieuwe kennis opleveren. Volgens hem gaan meta-analyses voorbij aan tal van factoren die relevant zijn voor het succes van interventies, doordat in elk stadium van de analyse noodgedwongen simplificaties en selecties moeten worden gemaakt. Wat de voorstanders zien als één van de sterke punten van de meta-analyse, de samenvatting van allerhande onderzoeksresultaten tot één of enkele kwantitatieve maten, zie Pawson juist als zwakte: de complexe en gelaagde interventiewerkelijkheid wordt 'platgeslagen' tot een ééndimensionale samenvatting waarmee theoretisch noch praktisch iets te doen valt.

Net als aan elke methodologie (en statistische techniek) zijn ook aan het gebruik van meta-analyse voor- en nadelen verbonden, kent het model mogelijkheden en beperkingen, en zijn er aannames waaraan voldaan moet worden voor een verantwoord gebruik (Lipsey & Wilson 2001, Wilson 2001, Davies & Crombie 2003). De voorstanders hebben gelijk wanneer ze beweren dat meta-analyse als methode van research synthese objectiever en transparanter is, dat de resultaten preciezer en beter replicerbaar zijn, dat de wijze waarop onderzoeksresultaten worden samengevat statistisch meer geavanceerd is en dat de methode beter in staat is om verborgen interventie-effecten te vinden. De tegenstanders hebben echter gelijk wanneer ze beweren dat in het veiligheidsveld zelden wordt voldaan aan de aannames die aan het model ten grondslag liggen, dat de geïncludeerde studies vaak te heterogeen zijn om te kunnen vergelijken en dat de methode doorgaans leidt tot op zijn best algemene en gesimplificeerde resultaten (waarbij veel informatie uit de originele studies verloren is gegaan) en op zijn slechtst tot misleidende resultaten (Davies & Crombie 2003, Berk 2005, Pawson 2006).

In navolging van Lipsey & Wilson (2001) pleit ik voor het gebruik van verschillende methoden van research synthese naast elkaar, omdat elk van de methoden pro's en con's kent. Dit betekent een meer genuanceerde kijk op het fenomeen research synthese: de resultaten van meta-analyses zijn niet zaligmakend, en zeker niet voldoende om te komen tot effectieve interventies, maar ze vormen wel een zeer bruikbare en ook noodzakelijke schakel in het proces dat leidt tot effectievere interventies. Hiervoor zijn echter ook andere methoden nodig.

4. Kanttekeningen bij de totstandkoming en toepassing van *What works* beginselen

De onderzoeksresultaten die worden verkregen via de hiervoor beschreven onderzoeksmethoden hebben geleid tot algemene *What works* beginselen. In deze paragraaf wil ik drie kanttekeningen plaatsen bij de totstandkoming en toepassing van deze beginselen:

1. De totstandkoming van de algemene *What works* beginselen is helder noch systematisch
2. De *What works* beginselen zijn weinig bruikbaar voor de interventiepraktijk
3. De focus op *What works* beginselen is niet bevorderlijk voor innovatie

Ad 1) De totstandkoming van algemene What works beginselen is helder noch systematisch
Met name onderzoekers van de 'Canadese school' zijn succesvol geweest in het wereldwijd verspreiden van veel gehanteerde *What works* beginselen, zoals het risicobeginsel, het behoeftebeginsel, et cetera (zie overzicht bij Van der Laan 2004). Je ziet deze beginselen, in rijtjes van wisselende lengte, in tal van publicaties terug. Voor de gemiddelde onderzoeker in het veiligheidsveld beginnen ze inmiddels verdacht veel te lijken op mantra's: als je ze maar vaak genoeg herhaalt, verkrijgen ze vanzelf een zekere geldigheid.

Hoewel veel van de *What works* beginselen nogal *straight forward* zijn en weinig discussie zullen oproepen, wil ik toch een paar kanttekeningen plaatsen bij de wijze waarop we tot deze beginselen zijn gekomen. Ondanks verwoede pogingen ben ik er niet in geslaagd te ontdekken op basis van welke (bij voorkeur heldere) criteria de gevonden onderzoeksresultaten nu zijn vertaald naar deze beginselen? Is dit gebaseerd op een bepaalde effectgrootte? Of op een bepaalde hoeveelheid studies waarin deze factoren een rol hebben gespeeld? En welke onderzoeksresultaten hebben het dan NIET tot beginselen geschopt en waarom niet? Of doen we daar niet aan en worden alle 'significante' factoren vertaald in beginselen en zo ja, moeten we dan op enig moment niet bepaalde beginselen boven andere verkiezen om het rijtje werkbaar te houden? En welke criteria gebruiken we hiervoor? Zijn de studies waar deze beginselen op gebaseerd zijn wel voldoende extern valideerbaar, bijvoorbeeld naar andere (nationale) contexten (Haen Marschall 2005)? En hoe is omgegaan met het feit dat je vanuit probabilistische onderzoeksresultaten komt tot min of meer deterministische (immers: algemeen geldende en toe te passen) beginselen? Ik kan me zo voorstellen dat bepaalde beginselen in bepaalde contexten belangrijker zijn dan in andere.

Ik mis, om kort te gaan, reflectie op dit soort mijns inziens belangrijke kwesties. Wat me opvalt is het gebrek aan systematiek op dit vlak, terwijl de onderzoekers in deze benadering zich juist voor laten staan op methodologische 'hardheid': de gemiddelde effectgrootte van interventies wordt in een meta-analyse achter de komma berekend, maar de vertaalslag hiervan naar beginselen voor goede praktijken lijkt meer op een slag in de lucht. Dit roept tenminste de vraag op hoe bruikbaar deze beginselen nu precies zullen zijn voor de praktijk. Dit is ook het onderwerp van de volgende kanttekening.

Ad 2) De What works beginselen zijn weinig bruikbaar voor de interventiepraktijk
Mijn voornaamste bezwaar tegen de toepassing van de *What works* beginselen is dat deze dermate algemeen zijn dat hun praktische relevantie bijzonder beperkt is. Hier doet zich een dilemma voor dat we niet zomaar kunnen opheffen. Immers, kennis over effectieve interventies vereist dat we onderzoeksresultaten 'stapelen', maar die gestapelde kennis wordt vervolgens zo algemeen, zeker bij het gebruik van meta-analyses, dat deze in specifieke interventiecontexten zijn toepassingskwaliteit verliest. Neem *What works* beginselen als 'responsiviteit' of 'professionaliteit'. De omschrijving van deze beginselen is dermate vaag en algemeen dat er alleen open deuren mee geopend kunnen worden. Ze kunnen als zodanig nauwelijks een sturende functie vervullen voor de praktijk (zie bijv. Merrington & Stanley

2000, Ferguson 2002). De kloof tussen algemene richtlijnen voor effectieve interventies en praktische aanwijzingen die nodig zijn om specifieke interventies succesvol te maken is groot. Deze kloof wordt door de *What works* benadering niet overbrugd, integendeel: de focus ligt hier eerder op het veralgemenen van de onderzoeksresultaten.

Als men in enige interventiepraktijk al een adequate operationaliseringslag kan maken met betrekking tot deze beginselen, stuit men vervolgens op het probleem dat tal van situaties nopen tot afwijken van het algemene beginsel, bijvoorbeeld omdat anders een ander algemeen beginsel in het gedrang komt. Zo vraagt het beginsel ‘programma-integriteit’ om uitvoering van het programma zoals bedoeld, terwijl het beginsel ‘responsiviteit’ erom kan vragen om hiervan af te wijken als bijvoorbeeld blijkt dat bepaalde justitiabelen niet bereikt kunnen worden door de van tevoren bepaalde methode. Responsiviteit vraagt uiteindelijk om individuele afstemming van een interventie op het individu en zijn mogelijkheden. Dit kan impliceren dat moet worden afgeweken van de interventie zoals bedoeld. Hoe hiermee om te gaan?

De beginselen kunnen ook op gespannen voet staan met juridische of andersoortige randvoorwaarden waarbinnen interventies zich (moeten) voltrekken. Veel interventieprogramma’s hebben om juridische, financiële of andersoortige redenen bijvoorbeeld een beperkte doorlooptijd. Dit enkele gegeven kan het al moeilijk of zelfs onmogelijk maken om bepaalde beginselen in de praktijk te brengen.

Het academisch discours over de *What works* beginselen kenmerkt zich soms door onverschilligheid of onwetendheid over de wijze waarop in concrete praktijken vorm kan worden gegeven aan interventies en de praktische beperkingen (zie hiervoor) die hierbij onvermijdelijk aan de orde zijn. Door hiermee géén of onvoldoende rekening te houden formuleert de wetenschap, vanuit de optiek van de praktijk, eerder utopische (wenselijke) dan praktisch bruikbare en toepasbare kennis. Terwijl het streven nu juist is om te komen tot effectievere interventies in de praktijk.

Ad 3) De focus op What works beginselen is niet bevorderlijk voor innovatie

What works beginselen zijn tot stand gekomen door de resultaten van grote hoeveelheden (bestaand) onderzoek te ‘stapelen’ en te vertalen naar algemene beginselen voor *good practices*. Deze vormen weer de basis voor de kwaliteitscriteria die beleidsmakers gebruiken om nieuwe interventies goed te keuren. Deze werkwijze geeft een sterke focus op kennis die in het verleden is gegenereerd en neemt deze als maatstaf voor de toekomst. Hierdoor wordt mijns inziens weinig of geen ruimte gelaten voor innovatie en bovendien wordt bestaande kennis impliciet verabsoluteerd. We mogen aannemen dat onze huidige inzichten in de toekomst plaats zullen maken voor andere, betere inzichten. De hiervoor beschreven benadering is niet bepaald een geschikte randvoorwaarde voor dergelijke innovatie. Innoveren vraagt om risico’s nemen, nieuwe (theoretische) paden durven bewandelen. De *What works* benadering is daarentegen juist geënt op principes als zekerheid en gebaande paden.

5. Bespreking van resultaten

De *What works* benadering legt een zware nadruk op de methodologische kwaliteit van onderzoeksdesigns, meer in het bijzonder op het aspect van de interne validiteit. Daar zijn goede redenen voor, omdat dergelijke designs beter dan andere in staat zijn om vast te stellen of de gepleegde interventie heeft geleid tot de gewenste resultaten. Onderzoek laat bovendien zien dat er een correlatie bestaat tussen kenmerken van het research design en uitkomsten in effectstudies: hoe ‘sterker’ het design des te geringer zijn doorgaans de effecten die worden gevonden (Weisburd et al. 2001). Bovenstaande bespreking laat echter zien dat de medaille

ook een keerzijde heeft. Om het in medische termen te formuleren: mooie designs zijn geen panacee voor alle kwalen. Ze kunnen weliswaar een belangrijke kwaal verhelpen (problemen bij vaststellen van causale effecten van interventies), maar de behandeling is niet zonder bijwerkingen. Deze bijwerkingen kunnen in voorkomende gevallen zelfs erger zijn dan de kwaal (de onderzoeksresultaten zijn bijvoorbeeld niet meer generaliseerbaar). In navolging van Eck (2006) pleit ik dan ook voor een ‘gemengde portfolio van invasieve, beperkt invasieve en niet-invasieve onderzoeksmethoden’ (2006:356-357). Door de hele range van onderzoeksmethoden in te zetten, kunnen we ook de respectieve voordelen benutten. Effectstudies beperken tot grootschalige gerandomiseerde veldexperimenten betekent dat we onszelf veel nuttige kennis ontzeggen.

De nadruk die sommigen tegenwoordig leggen op het belang van goede designs kan doorschieten. Zo constateert Farrington (2006) dat de criteria die de National Research Council in de VS hanteert om interventieprogramma’s in aanmerking te laten komen voor evaluatiestudies (de zogenaamde evalueerbaarheidscriteria) zodanig zijn dat circa 96% van de programma’s niet in aanmerking komt. Slechts 4% van de programma’s leent zich voor een evaluatiestudie die voldoet aan de hoogste methodologische eisen. Farrington, toch geen tegenstander van goede designs, vraagt zich af of het verstandig is om de onderzoeksgelden te beperken tot ‘Rolls-Royce evaluaties van Rolls-Royce programma’s’ en daardoor de overgrote meerderheid van de interventieprogramma’s ongeëvalueerd te laten (2006:330). Voorgaande bespreking toont aan dat er een spanning kan optreden tussen enerzijds het realiseren van research designs die hoog scoren op interne validiteit en anderzijds het streven naar en vormgeven in de praktijk van effectieve(re) interventies.

Dit spanningsveld tussen het realiseren van goed onderzoek enerzijds en het realiseren van effectieve interventies anderzijds is niet onvermijdelijk, maar ligt wel op de loer, zoals ook de volgende twee voorbeelden illustreren. Deze hebben betrekking op het vaststellen van de effectiviteit van ‘kleine interventies’ en de invloed van programma-ontwikkelaars/onderzoekers op de effectiviteit van interventies.

Kenmerkend voor de interventiepraktijk is dat er vaak meerdere interventies tegelijkertijd plaatsvinden, waarbij elke interventie begrensd wordt door juridische, organisatorische, financiële en andere randvoorwaarden, zodat sprake is van een lappendeken van ‘kleine interventies’ die naast elkaar bestaan. Bijvoorbeeld een jongere die een leerstraf ondergaat voor ongeoorloofd spijbelen, heeft vaak ook (al) te maken met andersoortige interventies (op school, thuis, bij de politie, bij de Raad voor de Kinderbescherming, et cetera). Eén van de interventies, bijvoorbeeld de leerstraf, eruit lichten en onderzoeken of deze van invloed is op bijvoorbeeld recidive is in deze situatie een weinig zinvolle strategie. Toch gebeurt dit de hele tijd. Vanuit het streven naar onderzoeksdesigns die hoog scoren op interne validiteit is dit ook goed te begrijpen: je hebt immers een heldere, afgebakende interventie nodig en bij voorkeur random toekenning van justitiabelen aan deze interventie. Deze door onderzoekstechnische redenen ingegeven behoefte gaat echter voorbij aan het nuchtere feit dat de betreffende (kleine) interventie in *the grand scheme of things* hoogstwaarschijnlijk een verwaarloosbaar effect zal hebben op recidive (dit blijkt ook steevast). Er is in de sociale realiteit van de betreffende jongere immers zoveel meer aan de hand dat de recidive kan beïnvloeden, maar voor deze factoren is geen plaats in het design. Rekening houden met deze factoren, en dus met andere interventies, betekent echter dat het experimentele design niet langer gebruikt kan worden. Wat te doen?

Ander voorbeeld: onderzoek toont aan dat er een correlatie bestaat tussen de rol van programma-ontwikkelaars/onderzoekers bij een interventie en de gerapporteerde effecten: hoe meer de onderzoekers betrokken zijn bij de opzet, implementatie en uitvoering van de interventie, des te sterker zijn de gerapporteerde effecten van die interventie. Het gaat om sterke effecten (Petrosino & Soydan 2005, Sherman 2006). Er zijn aanwijzingen dat deze

correlatie veroorzaakt wordt door de grotere betrokkenheid van de betreffende onderzoekers. Deze zou er voor zorgen dat er met meer aandacht, betrokkenheid, gedrevenheid, etc. gewerkt wordt aan de interventie en dit zou zich vervolgens vertalen in een 'betere' interventie. Vanuit de optiek van onderzoek is dit een te vermijden situatie omdat er ongewenste vertekening in de resultaten kan sluipen. Vanuit de optiek van de praktijk echter is het juist interessant om deze verbinding nader te onderzoeken en na te gaan welke factoren hier spelen en op welke wijze deze ingezet kunnen worden teneinde de effecten van de interventie te maximaliseren. Wat te doen?

Het streven naar *evidence-based* interventies is een streven dat ik voluit ondersteun. Ik signaleer tegelijkertijd dat in de huidige *What works* benadering de focus wel erg sterk ligt op het realiseren van 'mooie' designs en ik constateer tegelijkertijd dat dit streven in zijn uitwerking soms haaks kan komen staan op het beoogde doel, het bereiken van effectievere interventies. Hierin zal een middenweg bewandeld moeten worden: de kloof tussen organisaties en professionals die 'maar wat doen' aan de ene kant en wetenschappers die kennis produceren zonder praktijkrelevantie aan de andere kant zal van beide kanten gedicht moeten worden.

Onderzoekers kunnen mijns inziens bijdragen aan effectievere interventies door een paar zaken in acht te nemen. Zij kunnen bijvoorbeeld actiever meedenken over de vertaalslag van wetenschappelijke kennis naar toepassing in interventiepraktijken en hierbij geen afstand nemen van de praktische randvoorwaarden die hierbij aan de orde zijn, deze zijn immers onlosmakelijk met effectiviteit en succes verbonden. Ook is het verstandig als zij het gehele arsenaal aan beschikbare onderzoeksmethodologieën benutten en daaruit de beste kennis destilleren. Effectonderzoek moet niet beperkt worden tot onderzoeksdesigns die hoog scoren op interne validiteit en interventieprogramma's die op deze manier bestudeerd kunnen worden. Deze werkwijze belemmert kennisgroei. Ten slotte is het belangrijk om in effectstudies een expliciete verbinding te leggen tussen inhoud en uitkomst van interventies en binnen de interventies meer aandacht besteden aan sociale interacties en betekenisverlening door betrokkenen actoren. Andere disciplines laten ons zien dat hier winst te boeken valt.

Literatuur

- Asscher, J., Dekovic, M., Laan, P. v. d., Prins, P., & Arum, S. v. (2007). De effectiviteit van multisysteemtherapie (MST). Knelpunten en kansen bij de implementatie van een randomized controlled trial. *Tijdschrift voor Criminologie*, 49(1), 57-67.
- Atkinson, D. (2004). The What Works Debate: Keeping a Human Perspective. *Probation Journal*, 51(3), 248-252.
- Barkham, M. (2007). The role of practitioners in implementing, designing and utilizing practice based evidence. In W. Bruinsma & C. Konijn & M. v. d. Steege (Eds.), *Jeugdzorg in onderzoek: De professional en effectiviteit* (pp. 41-54). Den Haag: Ministerie van VWS.
- Beenackers, E. M. T. (2000). *Effectiviteit van sanctieprogramma's: bouwstenen voor een toetsingskader : een literatuuronderzoek* (WODC onderzoeksnotities ; 2000/1). Den Haag: Wetenschappelijk Onderzoek- en Documentatiecentrum.
- Berghuis, B., Leeuw, F., & Scheepmaker, M. (2004). Beleidstheorieën - Voorwoord. *Justitiële Verkenningen*, 30(5), 5-9.
- Berk, R. E. (2005). Randomized experiments as the bronze standard. *Journal of Experimental Criminology*, 1, 417-433.
- Cronbach, L. J. (1982). *Designing evaluation of educational and social programs*. San Fransisco: Jossey-Bass.
- Cullen, F. T., & Gendreau, P. (2000). Assessing Correctional Rehabilitation: Policy, Practice, and Prospects. In J. Horney (Ed.), *Processes, and Decisions of the Criminal Justice System; Criminal Justice 2000* (Vol. 3, pp. 109-175). Rockville, MD: National Institute of Justice/NCJRS.
- Davies, H. T. O., & Crombie, I. K. (2007). *What is meta-analysis?* Hayward Medical Communications

- (download op: <http://www.evidence-based-medicine.co.uk/ebmfiles/WhatisMetaAn.pdf> [2007, 13-7-2007].
- Dowden, G., & Andrews, D. A. (2004). The Importance of Staff Practice in Delivering Correctional Treatment: A Meta-Analytic Review of Core Correctional Practice. *International Journal of Offender Therapy and Comparative Criminology*, 48(2), 203-214.
- Eck, J. E. (2006). When is a Bologna sandwich better than sex? A defense of small n-case study evaluations. *Journal of Experimental Criminology*, 2, 345-362.
- Erkenningscommissie-Gedragsinterventies. (2006). *Jaarverslag 2006*. Den Haag: Ministerie van Justitie.
- Farrington, D. P., Gottfredson, D. C., Sherman, L. W., & Welsh, B. C. (2002). The Maryland Scientific Methods Scale. In L. W. Sherman & D. P. Farrington & B. C. Welsh & D. L. MacKenzie (Eds.), *Evidence-Based Crime Prevention*. New York: 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, 9-38.
- Farrington, D. P. (2006). Methodological quality and the evaluation of anti-crime programs. *Journal of Experimental Criminology*, 2, 329-337.
- Feder, L., Jolin, A., & Feyerherm, W. (2000). Lessons from two randomized experiments in criminal justice settings. *Crime & Delinquency*, 46(3), 380-400.
- Ferguson, J. L. (2002). Putting the "What works" Research into Practice: An Organizational Perspective. *Criminal Justice and Behavior*, 29(4), 472-492.
- Ferwerda, H. B., Leiden, I. M. G. G. v., Arts, N. A. M., & Hauber, A. R. (2006). *Halt: het Alternatief? De effecten van HALT beschreven (reeks Onderzoek en Beleid, nr 244)*. Den Haag/Meppel: Wetenschappelijk Onderzoek- en Documentatiecentrum (WODC) / Boom Juridische uitgevers.
- Gendreau, P., & Ross, R. R. (1979). Effective correctional treatment: Bibliotherapy for cynics. *Crime & Delinquency*, 25(October), 463-489.
- Haen-Marshall, I. (2005). Introduction: International Transferability of the practice of evidence-based crime policies. *European Journal on Criminal Policy and Research*, 11, 225-231.
- Harrington, A. (2002). "Seeing" the Placebo-effect: Historical Legacies and Present Opportunities. In H. A. Guess & A. Kleinman & J. W. Kusek & L. Engel, W. (Eds.), *The Science of the Placebo; Toward an Interdisciplinary Research Agenda* (pp. 35-52). London UK: BMJ Books.
- Klein-Haarhuis, C., Ooyen-Houben, M. v., Kleemans, E., & Leeuw, F. (2005). Rechtshandhaving geevalueerd: een synthese van 31 onderzoeken. *Justitiele Verkenningen*, 31(8), 54-71.
- Laan, P. H. v. d. (2004). Over straffen, effectiviteit en erkenning; de wetenschappelijke onderbouwing van preventie en strafrechtelijke interventie. *Justitiele Verkenningen*, 30(5), 31-48.
- Laan, P. H. v. d., Krooi, H., Voort, P. v. d., & Wijkman, M. (2006). *Toetsing en certificering van preventieve, curatieve en repressieve interventies in de jeugdzorg*. Leiden: NSCR.
- Leeuw, F. L. (2005). Trends and Developments in Program Evaluations in General and Criminal Justice Programs in Particular. *European Journal on Criminal Policy and Research*, 11, 233-258.
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical Meta-Analysis*. Thousand Oaks [etc.]: Sage.
- Lipsey, M. W. (2001). Unsolved problems and unfinished business. *American Journal of Evaluation*, 22(3), 325-328.
- Lipton, D., Martinson, R., & Wilks, J. (1975). *The effectiveness of correctional treatment: A survey of treatment evaluation studies*. New York: Praeger.
- Lum, C., & Yang, S.-M. (2005). Why do evaluation researchers in crime and justice choose non-experimental methods? *Journal of Experimental Criminology*, 1, 191-213.
- Martinson, R. (1974). What works? Questions and answers about prison reform. *The Public Interest*, 35(Spring), 22-54.
- Merrington, S., & Stanley, S. (2004). What Works? Revisiting the Evidence in England and Wales. *Probation Journal*, 51(1), 7-20.
- National-Research-Council. (2005). *Improving evaluation of anticrime programs*. Washington D.C.: National Academies Press.
- Pawson, R. (2002). Evidence-based Policy: The promise of 'Realist Synthesis'. *Evaluation*, 8(3), 340-358.
- Pawson, R. (2006). *Evidence-based Policy. A Realist Perspective*. London [etc.]: Sage.
- Petrosino, A., Boruch, R. F., Soydan, H., Duggan, L., & Sanchez-Meca, J. (2001). Meeting the Challenges of Evidence-Based Policy: The Campbell Collaboration. *The Annals of the American Academy of Political and Social Research*, 578(November), 14-34.
- Petrosino, A., & Soydan, H. (2005). The impact of program developers as evaluators on criminal recidivism: Results from meta-analyses of experimental and quasi-experimental research. *Journal of Experimental Criminology*, 1, 435-450.
- Raynor, P. (2003). Evidence-Based Probation and its Critics. *Probation Journal*, 50(4), 334-345.
- Roos, T. d. (2007). Experimenteren met sancties? Strafrechtstheoretische kanttekeningen bij een empirisch-

- wetenschappelijke trend. *Tijdschrift voor Criminologie*, 49(1), 12-23.
- Rovers, B. (2006). Boekbespreking: Ferwerda, Henk B., Leiden, I.M.G.G. van, Arts, N.A.M. en A.R. Hauber (2006). *Halt: Het alternatief? De effecten van Halt beschreven*, Den Haag: Boom Juridische Uitgevers / WODC (Onderzoek en Beleid, deel 244). *Tijdschrift voor Veiligheid*, 5(4), 74-76.
- Rovers, B. (2007, 14 juni). *Preventie van jeugdcriminaliteit en de placeborespons; Toepassing van het belief effect in het veiligheidsveld*. Paper presented at the NVK-Marktdag, Leiden.
- Sherman, L. W., Gottfredson, D., MacKenzie, D., Eck, J., Reuter, P., & Bushway, S. (1997). *Preventing crime: What works, what doesn't, what's promising? A Report to the United States Congress*. Washington D.C.: Office of Justice Programs, U.S. Department of Justice.
- Sherman, L. W. (2006). "To develop and test": The inventive difference between evaluation and experimentation. *Journal of Experimental Criminology*, 2, 393-406.
- Wartna, B. S. J. (2005). *Evaluatie van daderprogramma's : een wegwijzer voor onderzoek naar de effecten van strafrechtelijke interventies speciaal gericht op het terugdringen van recidive*. Den Haag: Wetenschappelijk onderzoek- en documentatiecentrum (WODC) / Boom.
- Weisburd, D., Lum, C. M., & Petrosino, A. (2001). Does Research Design Affect Study Outcomes in Criminal Justice? *The Annals of the American Academy of Political and Social Research*, 578, 50-70.
- Welsh, B. C., & Farrington, D. P. (2001). Towards an Evidence-Based Approach to Preventing Crime. *The Annals of the American Academy of Political and Social Research*, 578, 158-173.
- Wilson, D. B. (2001). Meta-Analytic Methods for Criminology. *The Annals of the American Academy of Political and Social Research*, 578, 71-89.

Summary

‘What works’ has become an important research tradition within Criminology. It aims at more effective interventions by using only research designs that satisfy the highest methodological criteria (particularly randomized experimental designs), by focusing on quantitative research syntheses (by means of meta-analyses) and by focusing on the use of empirically proven methodologies in interventions. The article argues that this approach will not necessarily lead to more effective interventions.

Seven short comments are given on this approach. They refer to the research methodology (1-4) and the construction and use of the ‘What works’ principles (5-7). In short: 1) The potential use of randomized experimental designs in the Justice field is limited by various factors, some of which are more or less structural. 2) The strive for higher internal validity comes with a price, i.e. lower external validity (less possibilities to generalize results to other contexts). 3) The focus on internal validity, combined with the need to establish the effectiveness of interventions, has in practice reduced the intervention itself to a black box. As a result we don’t have a clue about the causal factors within interventions. 4) The role of meta-analyses in realizing more effective interventions is more limited than some think or hope. 5) The way in which research results are translated into ‘What works’ principles is not very clear and far from systematic. 6) The ‘What works’ principles have very limited practical value. They are too general for practical guidance and following different principles may in practice lead to opposing intervention strategies. 7) The use of the ‘What works’ principles inhibits innovative ideas, theories and interventions. The focus is on existing knowledge, implicitly making this knowledge absolute.

It is concluded that the approach taken by many researchers in the ‘What works’ tradition is adding to our knowledge, but at the same time the limitations of this approach are clear. The author makes some suggestions for future research on interventions.