Systematische reviews in de managementpraktijk

Hoe beoordeelt men de kwaliteit en toepasbaarheid?

Eric Barends, Rudolf W. Poolman, Dirk T. Ubbink en Steven ten Have

SAMENVATTING De wereld van management en organisatie kent een rijke traditie van omvangrijke onderzoeken naar succes en falen van organisaties en hun managementaanpakken. Een ikoon van deze traditie vormt het werk van Peters en Waterman naar excellente ondernemingen. In het op het onderzoek onder 43 als succesvol omschreven ondernemingen gebaseerde boek claimen Peter en Waterman het recept voor 'excellentie' te hebben. Dit en vergelijkbare onderzoeken hebben met elkaar gemeen dat zij populair zijn (geweest), grootschalig van opzet, de nodige pretentie hebben *en* de nodige kritiek hebben gekregen. Deze kritiek richt zich vooral op de methodologische tekortkomingen die maken dat de claim niet waargemaakt kan worden. Voor managers is het belangrijk dergelijke onderzoeken goed te beoordelen alvorens ze eventueel toe te laten tot de eigen handelingspraktijk. In dit artikel wordt in lijn met de zich ontwikkelende traditie van het evidence-based management aandacht besteed aan het kunnen beoordelen van systematische reviews, waartoe vaak populaire onderzoeken toe behoren. Er wordt een referentiekader geboden voor het kunnen beoordelen en dit kader wordt ook geïllustreerd aan de hand van een voorbeeld van een concreet en populair onderzoek.

RELEVANTIE VOOR DE PRAKTIJK Managers hebben in hun praktijk behoefte aan en baat bij in hun context goed werkende modellen, concepten en interventies. Deze worden in grote aantallen aangeboden in de vorm van populaire boeken en artikelen. Dikwijls zijn deze gebaseerd op indrukwekkend ogend onderzoek. Voor managers is het essentieel niet klakkeloos aan te haken bij een intuïtief aantrekkelijk concept dat extra verleidt omdat het gebaseerd zou zijn op grootschalig onderzoek. Om te voorkomen dat fouten worden gemaakt en tijd en middelen worden verspild, moet een manager hier het kaf van het koren kunnen scheiden. Het kunnen beoordelen van onderzoek kan daaraan een belangrijke bijdrage leveren. Hier worden handvatten en criteria aangereikt om deze beoordeling verantwoord en professioneel te kunnen doen.

1 Inleiding

De wereld van management en organisatie kent een rijke traditie van omvangrijke onderzoeken naar succes en falen van organisaties en hun managementaanpakken. Een ikoon van deze traditie vormt het werk van Peters en Waterman naar excellente ondernemingen (Peters en Waterman, 1982). Een paar jaar na het onderzoek bleken 14 van de 43 onderzochte ondernemingen in grote moeilijkheden te verkeren, onder hen IBM dat 'excellent' was in de 'mainframe' markt, maar de opkomst van de personal computer (mede daardoor) in eerste instantie miste. Later hebben de onderzoekers toegegeven dat hun selectie op relatief oppervlakkige wijze tot stand is gekomen; aan collega-consultants is eenvoudigweg gevraagd welke ondernemingen in de hen bekende sectoren tot de uitblinkers gerekend moeten worden. Andere, meer onder managers en adviseurs zeer populaire onderzoeken zijn die van Collins en Porras naar visionaire ondernemingen (Collins en Porras, 1995), het werk van Collins over leiderschap (Collins, 2001) en meer recent het onderzoek van De Waal naar 'high performance organizations' (De Waal, 2006a). Het gaat steeds om onderzoeken die in de publiciteit en onder managers het nodige teweeg hebben gebracht en vervolgens stevige kritiek hebben gekregen. Deze kritiek varieert van het plaatsen van kanttekeningen bij de houdbaarheid van de conclusies, zoals in het geval van de excellente ondernemingen (Business Week, 1984), methodologische kanttekeningen bij het onderzoek naar de high performance organizations (Ten Have, 2007), tot vergaande methologische kritiek op de genoemde onderzoeken naar leiderschap en visionaire ondernemingen (Rosenzweig, 2007). Onderzoeken naar succes en falen, wat werkt en wat niet werkt is, zeker als de specifieke contexten daar onderdeel van zijn, zijn van groot belang voor de managementpraktijk. Dit evidence-based practice paradigma, nl. het gebruik maken van evidence uit wetenschappelijk onderzoek ter verantwoording van de gekozen aanpak, wordt in de (klinische) gezondheidszorg steeds meer toegepast om de kwaliteit van de zorg te verbeteren (Ubbink en Legemate,

2004). Maar het is ook van groot belang dat die onderzoeken goed worden uitgevoerd en alleen dat claimen wat ze ook waar kunnen maken (Poolman et al, 2007). Onderzoeken die managers verleiden door een intuïtief aantrekkelijke boodschap te koppelen aan de verwijzing naar grootschalig onderzoek maar methodologisch tekort schieten, moeten met het oog op een gezond en hygiënisch vakgebied kritisch bejegend worden. Positiever geformuleerd: door misstanden aan de kaak te stellen kunnen 'afnemers' van managementideeën geholpen worden kritisch en alert te zijn. Door bij te dragen aan de verspreiding van kennis die nodig is om onderzoek goed te kunnen doen en te beoordelen wordt een impuls gegeven aan een gezond vakgebied en een vitale vakgemeenschap van managers.

Positief is dat de aandacht voor de empirische onderbouwing van theorieën, modellen en interventies ook op het gebied van management in de afgelopen jaren langzaam maar zeker lijkt toe te nemen. Vooral auteurs als Rousseau en Pfeffer en Sutton hebben door hun pleidooien voor evidence based management hieraan een belangrijke impuls gegeven (Pfeffer en Sutton, 2006; Rousseau, 2006). In ons artikel 'Op weg naar evidence-based verandermanagement' (Barends en Ten Have, 2008) wordt beschreven hoe binnen het vakgebied verandermanagement de fase van 'vrije' theorievorming die gekenmerkt wordt door een grote verscheidenheid aan scholen, ideologisch gekleurde opvattingen, goeroes en aanhangers, plaats aan het maken is voor een fase waarin 'voorzichtig' onderzoek gedaan wordt. Daarbij gaat het vooral om observationeel en retrospectief onderzoek. Onder observationeel onderzoek verstaat men onderzoek waarbij de onderzoeker alleen waarneemt en niet intervenieert, met de bedoeling om verbanden te vinden tussen de waargenomen gegevens. Een bekend type observationeel onderzoek is het cohort-onderzoek, waarbij grote groepen mensen of bedrijven gedurende een lange periode gevolgd worden om te kijken (prospectief) of er verschil ontstaat tussen de groepen. Een ander bekend type observationeel onderzoek is het case-control onderzoek, waarbij een groep bedrijven met een bepaalde uitkomst (bijvoorbeeld bovengemiddeld goede prestaties) achteraf (retrospectief) vergeleken wordt met een groep die deze uitkomst niet heeft. Bekende case-control onderzoeken zijn 'In Search of Excellence' van Tom Peters en 'Good to Great' van Jim Collins (Collins, 2001; Peters en Waterman, 1982). Dergelijk onderzoek is in deze fase weliswaar noodzakelijk voor de ontwikkeling van (verander)management tot een 'volwassen' wetenschappelijk onderbouwd vakgebied, maar de werkwijze die daarbij gehanteerd wordt is vanuit methodologisch oogpunt beperkt. De twee meest in het oog springende beperkingen betreft het niet kunnen valideren van de volledigheid en betrouwbaarheid van de gegevens, en de onmogelijkheid tot het bewijzen van causale verbanden tussen de waargenomen variabelen. Zo wordt in de onderzoeken van zowel Peters als Collins in een grote verzameling gegevens willekeurig gezocht naar statistische verbanden en kritische succesfactoren1. Ditzelfde geldt voor een andere vorm van onderzoek die steeds vaker gepubliceerd wordt, die van de systematische review (Arthur *et al*, 2003; Collins en Holton, 2004). Vaak is een zogenaamde meta-analyse, waarbij de uitkomst van meerdere studies gecombineerd worden, onderdeel van een systematische review. Systematische reviews met een meta-analyse worden, mits goed uitgevoerd en gebaseerd op goed uitgevoerde primaire studies, beschouwd als een krachtige vorm van bewijsvoering. Maar de uitkomsten van systematische reviews met een meta-analyse kunnen een sterk vertekend beeld opleveren doordat gebruik gemaakt wordt van studies die onderling verschillen qua opzet en methodologische kwaliteit. Gezien het gewicht dat in het algemeen aan meta-analyses wordt toegekend is het van belang dat managers in staat zijn om dergelijk onderzoek kritisch tegen het licht te houden en te beoordelen wat de waarde is voor de managementpraktijk.

Het doel van dit artikel is om managers een kader te bieden bij het beoordelen en interpreteren van systematische reviews. Hiervoor gaan we in op de belangrijkste aspecten van een systematische review zoals beschreven in de richtlijnen van de Cochrane Collaboration (Higgins en Green, 2006), de MOOSE Guidelines (Stroup *et al*, 2000) en de QUOROM Statement checklist (Moher *et al*, 1999). Hoewel deze richtlijnen ontwikkeld zijn binnen de geneeskunde zijn de methodologische uitgangspunten en principes van toepassing op elk vakgebied waar wetenschappelijk onderzoek wordt gedaan, van astronomie tot zoölogie (Petticrew, 2001). Ter illustratie wordt in dit artikel een bekende systematische review beoordeeld. Verondersteld wordt dat het kennen en begrijpen van de basisterminologie en het concept van de systematische review, managers kan helpen bij het beoordelen van de kwaliteit van het onderzoek en de relevantie van de uitkomst voor de dagelijkse praktijk van de manager. Daardoor wordt een bijdrage geleverd aan een verantwoorde en professionele ontwikkeling van het vak. Onderzoeken worden dan op waarde geschat en degenen die omarmd worden vormen de basis voor professioneel handelen in plaats van mooie, maar ongefundeerde beloftes.

2 Het concept van de systematische review

2.1 Beschrijvende vs systematische reviews

Er bestaan verschillende soorten overzichtsartikelen of reviews. Het meest gezaghebbende overzichtsartikel, dat wil zeggen de review met de meest krachtige bewijsvoering, is de systematische review. Een systematische review beoogt zo volledig mogelijk alle relevante wetenschappelijke studies met betrekking tot een specifiek onderwerp te

Tabel 1 Hiërarchie van bewijskracht

Level	Onderzoeksdesign	Conclusie
A1	Systematische review van tenminste level A3 studies	Het is aangetoond dat …
A2	Gerandomiseerd en gecontroleerd onderzoek (RCT)	
A3	Gecontroleerd onderzoek zonder randomisatie	
B1	Systematische review van tenminste level B3 studies	Het is aannemelijk dat …
B2	Cohort onderzoek	
B3	Case control onderzoek	
C1	Systematische review van tenminste level C3 studies	Er zijn aanwijzingen dat …
C2	Niet-vergelijkend onderzoek met een voormeting, vergelijkend onderzoek zonder randomisatie en zonder voormeting	
C3	Niet-vergelijkend onderzoek zonder voormeting, cross-sectioneel onderzoek en case-studies	
D	Mening van deskundigen	Deskundigen zijn van mening dat …
Bovenstaande indeling is gebaseerd op de Levels of Evidence van het Oxford Centre for Evidence-based Medicine en de indeling van Campbell (Shadish, Cook and Campbell, 2002, Experimental and Quasi-Experimental Designs for Generalized Causal Inference)		

Tabel 2 Geschiktheid van verschillende onderzoeksdesigns voor verschillende onderzoeksvragen

Onderzoeksvraag	Kwalitatief onderzoek	Enquêtes	Cohort en case-control onderzoek	RCT's
Effectiviteit: werkt het?, werkt A beter dan B?			+	++
Proces: hoe werkt het?	++	+		
Veiligheid: wegen de positieve effecten op tegen de negatieve (bij)effecten?	+		+	++
Acceptatie: is de doelgroep bereid de nieuwe werkwijze over te nemen?	++	+		+
Kosteffectiviteit: leidt het tot lagere kosten?, is A goedkoper dan B?				++
Geschiktheid: is dit de beste interventie / werkwijze voor deze doelgroep?	++	++		
Tevredenheid: zijn medewerkers en/of klanten tevreden over de nieuwe werkwijze?	++	++	+	
Gebaseerd op Petticrew & Roberts, Evidence, hierarchies and typologies: Horses for Courses. Journal of Epidemiology and Community Healthcare, 2003, 57: 527-9				

identificeren en de validiteit en kracht van de bewijsvoering van iedere studie afzonderlijk te beoordelen. Zoals de naam al aangeeft wordt bij een systematische review bij het zoeken naar studies op een systematische manier te werk gegaan en de methodologische kwaliteit door meerdere onderzoekers onafhankelijk van elkaar kritisch beoordeeld, waardoor de review transparant, controleerbaar en reproduceerbaar is. Het gebruik van statistische analysetechnieken in een systematische review om de uitkomsten van de individuele studies getalsmatig te combineren om daarmee tot een nauwkeuriger schatting van het effect te komen, wordt een 'meta-analyse' genoemd.

Naast de systematische review komen ook andere soorten reviews voor, zoals de beschrijvende review. Een beschrijvende review wordt ook wel narratieve- of literatuur review genoemd. Een bekend voorbeeld van een beschrijvende review is die waarin een overzicht wordt gegeven van theorieën en wetenschappelijk onderzoek op het gebied van organisatieverandering in de jaren '90 (Armenakis en Bedeian, 1999). In dit overzichtsartikel wordt aan de hand van vier thema's (content, context, process en outcome assessment) een aantal studies besproken die volgens Armenakis representatief zijn voor al het onderzoek binnen het thema. Op deze wijze passeren op basis van een beperkt aantal studies een groot aantal onderwerpen de revue. Maar een systematische review beperkt zich in de regel tot een specifiek aspect van organisatieverandering zoals bijvoorbeeld de invloed van directe en indirecte participatie van medewerkers op de arbeidssatisfactie en betrokkenheid bij de organisatie (Macy, Peterson en Norton, 1989). In een beschrijvende review wordt wel een overzicht gegeven van wetenschappelijke studies, maar wordt niet op een systematische wijze gezocht. Bovendien worden studies vaak geselecteerd op basis van het gezichtspunt van de reviewer. Ook vindt bij een beschrijvende review geen meta-analyse plaats. Door de subjectieve manier waarop het zoeken en selecteren van de studies plaatsvindt, is de kans op vertekening van de uitkomst groot (Antman, 1992; Bushman en Wells, 2001). Om deze reden wordt een beschrijvende review beschouwd als een zwakkere vorm van bewijsvoering dan een systematische review.

2.2 Hiërarchie van bewijsvoering: hoe sterk is het bewijs?

In de Engelstalige literatuur wordt bewust gesproken van 'evidence' en niet van 'proof'. Evidence is dan ook niet hetzelfde als bewijs, maar kunnen aanwijzingen zijn die zo zwak zijn dat ze nauwelijks overtuigen of zo sterk dat niemand twijfelt over de juistheid' (Offringa, Assendelft en Scholten, 2008). Het is daarom belangrijk om te kunnen bepalen welke evidence de meeste bewijskracht heeft. Hiervoor wordt gebruik gemaakt van zogenaamde 'levels

of evidence' die een hiërarchische ordening aangeven van de verschillende onderzoeksdesigns op basis van de interne validiteit (Guyatt *et al*, 1995; Phillips *et al*, 2001). De interne validiteit geeft aan in hoeverre de resultaten van het onderzoek vertekend kunnen zijn en zegt dus iets over de mate waarin alternatieve verklaringen voor de gevonden uitkomst mogelijk zijn. Het zuivere experiment in de vorm van een gerandomiseerd en gecontroleerd vergelijkend onderzoek2 (RCT, randomised controlled trial) geldt als 'gouden standaard', gevolgd door niet-gerandomiseerd gecontroleerd onderzoek (quasi-experiment) en observationele studies zoals cohort- en case-control studies (zie tabel 1). Enquêtes en case studies worden beschouwd als onderzoeksdesigns waarbij de kans op vertekening van de uitkomst het grootst is en staan daarom laag in de hiërarchie. Helemaal onderaan staan beweringen die uitsluitend zijn gebaseerd op de persoonlijke mening van deskundigen. Systematische reviews met een meta-analyse van gerandomiseerde en gecontroleerde studies worden daarom beschouwd als het hoogste level of evidence.

2.3 Hiërarchie van bewijsvoering: welk onderzoek voor welke vraag?

In de sociale wetenschappen en het vakgebied management zullen voornamelijk systematische reviews voorkomen van niet-gerandomiseerd en observationeel onderzoek, casestudies en enquêtes. Hoewel de interne validiteit van dit type onderzoek minder groot is dan van gerandomiseerd onderzoek, betekent dit niet dat dergelijke vormen van onderzoek minder bruikbaar zijn. Immers, als alleen dit type onderzoek is verricht, blijft dit het best beschikbare bewijsmateriaal. Bovendien doen de 'levels of evidence' geen uitspraak over de externe validiteit (generaliseerbaarheid) van een studie. Een RCT kan minder goed generaliseerbaar zijn, wat de praktische bruikbaarheid beperkt. Observationele studies en niet-vergelijkend onderzoek kennen daarentegen een lagere interne validiteit, maar kunnen voor de managementpraktijk heel bruikbaar zijn, mits men daarbij in het achterhoofd houdt dat de resultaten van dergelijke studiedesigns meer vatbaar zijn voor vertekening. Een bezwaar tegen de levels of evidence is dan ook dat deze onvoldoende rekening houden met de methodologische geschiktheid van het gekozen onderzoeksdesign in relatie tot de onderzoeksvraag (Guyatt *et al*, 2008). Verschillende typen onderzoeksvragen vereisen verschillende typen wetenschappelijk onderzoek: zo komt evidence met betrekking tot het effect van interventies van RCT's, evidence met betrekking tot bijeffecten en risicofactoren van observationeel onderzoek, en evidence met betrekking tot de wijze waarop het effect tot stand is gekomen komt vaak van kwalitatief onderzoek (Petticrew en Roberts, 2003). Bij de beoordeling van de kwaliteit van een systematische review zal dus tevens gekeken moeten worden of met betrekking tot de onderzoeksvraag naar de juiste typen onderzoek is gezocht.

Tabel 3 Beoordeling van een systematische review

1	Gaat de review uit van een expliciete vraagstelling?
2	Is de zoekactie systematisch en reproduceerbaar?
3	Is publicatiebias zoveel mogelijk voorkomen?
4	Heeft de selectie plaatsgevonden door ≥ 2 reviewers onafhankelijk van elkaar?
5	Heeft de selectie plaatsgevonden aan de hand van expliciete in- en exclusiecriteria?
6	Is de kwaliteit van de studies beoordeeld door ≥ 2 reviewers onafhankelijk van elkaar?
7	Is de kwaliteit beoordeeld aan de hand van expliciete criteria?
8	Is beschreven hoe data-extractie heeft plaatsgevonden?
9	Zijn de kenmerken van de oorspronkelijke studies beschreven?
10	Is voldoende rekening gehouden met heterogeniteit?
11	Is de meta-analyse correct uitgevoerd?
12	Is de conclusie in lijn met de bewijskracht van de geincludeerde studies?
13	Is de uitkomst generaliseerbaar naar de eigen praktijk?
14	Is de uitkomst toepasbaar in de eigen praktijk?

Bovenstaande vragenlijst is gebaseerd op de gevalideerde index van Oxman and Guyatt (Oxman, A. D., Guyatt, G. H. 1991. Validation of an index of the quality of review articles. Journal of Clinical Epidemiology; 44(11):1271-8.

In tabel 2 wordt een overzicht gegeven van de geschiktheid van verschillende soorten onderzoek met betrekking tot een specifieke onderzoeksvraag.

3 Beoordeling van de kwaliteit van een systematische review

Om de kwaliteit van een systematische review te kunnen vaststellen, moet deze kritisch en deskundig worden beoordeeld. Om dit te doen kan gebruik gemaakt worden van een aantal essentiële vragen gebaseerd op de gevalideerde index van Oxman en Guyat (zie tabel 3). Deze vragen worden hierna behandeld.

3.1 Gaat de review uit van een expliciete vraagstelling?

Een systematische review dient uit te gaan van een expliciete vraagstelling met voldoende focus. In de medische wereld wordt hiervoor gebruik gemaakt van een handig ezelsbruggetje. Artsen en medische studenten hebben geleerd om eerst een zogenaamde PICO te maken voordat ze op zoek gaan naar evidence. In dit acroniem staat de P voor Patiënt, Problem of Population, de I voor Intervention, de C voor Comparison en de O voor Outcome. In de sociale wetenschappen is dit ezelsbruggetje overgenomen en de C van Context toegevoegd (PICOC). De gedachte hierachter is dat alle vijf elementen van belang zijn bij het gericht zoeken naar evidence en elke wijziging in de P, I, C, O of C tot een andere evidence en dus ook een andere uitkomst leidt. Zo leidt een zoekactie op basis van een vraag als 'Is appreciative inquiry een effectieve interventie bij veranderingen?' tot een uitkomst met een beperkte praktische waarde omdat in de vraagstelling alleen de I (appreciative inquiry) geadresseerd wordt, zonder rekening te houden met:
de C: in vergelijking met business proces redesign is het effect

mogelijk anders dan in vergelijking met survey-feedback;
de O: het effect op de productie is mogelijk anders dan het effect op de veranderbereidheid bij middle-managers;
en de C: bij een financiële dienstverlener is het effect mogelijk anders dan bij een academisch ziekenhuis.

De lezer zal bij de beoordeling van een systematische review dus goed moeten kijken of deze uitgaat van een vraag die van toepassing is op de dagelijkse praktijk van de manager. In dat kader is een antwoord op de vraag 'wat werkt' voor de managementpraktijk minder relevant dan het antwoord op de vraag 'wat werkt, voor wie, bij welk probleem, in welke context'. Een duidelijke beschrijving van de organisatie, de doelgroep, het probleem, de interventie (of methodiek, onafhankelijke variabele of succesfactor), de context en de uitkomstparameters is daarvoor een vereiste.

3.2 Is de zoekactie systematisch en reproduceerbaar en is publicatiebias zoveel mogelijk voorkomen?

Op basis van de vraagstelling dient in de internationale literatuur op een gestructureerde wijze gezocht te worden naar alle relevante onderzoeksartikelen. In eerste instantie zal daarbij gezocht worden in bibliografische databases op basis van duidelijk omschreven zoektermen. De meest relevante bibliografische databases in ons vakgebied zijn ABI/INFORM van ProQuest, Business Source Premier van EBSCO en Science Direct van Elsevier. Met deze databases kunnen op basis van trefwoorden of woorden in de titel of de samenvatting in meer dan 2500 journals gezocht worden naar (onderzoeks)artikelen op het gebied van bedrijfskunde, management, marketing, financiën en economie. Naast deze twee databases dient tevens gezocht te worden in databases die zich richten op aanpalende disciplines zoals psychologie (PsycINFO) en educatie (ERIC). Indien niet in ABI/INFORM én Business Source Premier is gezocht is een review niet valide omdat de kans groot is dat veel relevante studies gemist zijn. Ditzelfde geldt voor een zoekactie die zich alleen beperkt tot elektronische databases. Om deze reden dient ook handmatig gezocht te worden in indexen van journals en referenties van gevonden onderzoeksartikelen, met name naar ongepubliceerde studies.

De reden dat ook naar ongepubliceerde studies gezocht dient te worden is gelegen in het feit dat studies die een positief effect aantonen een grotere kans hebben om gepubliceerd te worden dan studies met een negatieve uitkomst, hetzij doordat dergelijke studies door de redactie van een journal geweigerd wordt, hetzij doordat de auteur van de studie het onderzoek überhaupt niet voor publicatie aanbiedt (Olson *et al*, 2002). Al sinds de jaren 50 is gebleken dat studies die een negatief effect aantonen ondervertegenwoordigd zijn in journals op zowel medisch, psychologisch als sociaal wetenschappelijk gebied, dus aangenomen mag worden dat dit voor het vakgebied management niet anders is (Dickersin, 1990; Egger en Smith, 1998; Egger *et al*, 1997; Rosenthal, 1979; Sterling, 1959)
Als een systematische review alleen gebaseerd is op gepubliceerde studies is de kans groot dat vertekening door publicatiebias optreedt. Bij publicatiebias wordt de uitkomst vertekend doordat studies met een positief effect oververtegenwoordigd zijn, waardoor de uitkomst van de meta-analyse sterk vertekend wordt, vaak in de vorm van een overschatting van de onderzochte relatie of het positieve effect. Uit een studie naar meta-analyses leidde de exclusie van de ongepubliceerde studies tot een toename van het gevonden effect van 15 procent (McAuley *et al*, 2000). Om deze reden zal de lezer bij de beoordeling van een systematische review moeten nagaan of er sprake is van publicatiebias.
Tot slot moet tevens rekening gehouden worden met vertekening van de uitkomst ten gevolge van taalrestricties. De kans is daardoor aanwezig dat relevante studies uit andere taalgebieden gemist worden. Daarnaast lijken in Engelstalige journals studies met een positieve uitkomst oververtegenwoordigd te zijn. Dit wordt waarschijnlijk veroorzaakt doordat onderzoekers studies met een positieve uitkomst eerder in internationale, Engelstalige literatuur publiceren en studies met een negatieve uitkomst bij voorkeur in een tijdschrift in hun eigen land (Egger en Smith, 1998).

3.3 Is de selectie van studies systematisch en reproduceerbaar?

In de regel zal een zoekactie leiden tot een groot aantal 'hits', soms enkele duizenden. Een deel van de gevonden studies zal niet direct relevant zijn voor de onderzoeksvraag. Om deze reden zal men vervolgens beoordelen welke studies uitgesloten kunnen worden van de systematische review. Om selectiebias te voorkomen dient deze selectie van studies bij voorkeur uitgevoerd te worden door minimaal twee reviewers, onafhankelijk van elkaar en op basis van duidelijk omschreven in- en exclusiecriteria. Het selectieproces dient bovendien helder gedocumenteerd te worden, bijvoorbeeld in de vorm van een stroomdiagram waarin zichtbaar wordt hoeveel studies geëxcludeerd zijn en op grond van welke criteria. In- en exclusiecriteria dienen nauw aan te sluiten op de onderzoeksvraag. Dit kunnen criteria zijn met betrekking tot het type organisatie, de doelgroep, de soort interventie of te onderzoeken factor en uitkomstparameter (PICOC), maar ook criteria met betrekking tot het onderzoeksdesign of level of evidence. Zoals reeds aangegeven, is de vraag welk onderzoeksdesign het meest geschikt is voor beantwoording van de onderzoeksvraag daarbij van belang.

3.4 Is de methodologische kwaliteit van de studies adequaat beoordeeld?

De kwaliteit van de systematische review wordt sterk beïnvloed door de methodologische kwaliteit van de geïncludeerde primaire studies. Het meeste onderzoek op het gebied

van management heeft methodologische beperkingen die de uitkomst in meer of mindere mate vertekenen, meestal in de vorm van een overschatting van het effect of de gevonden relatie. Bij de interpretatie van de uitkomsten dient hiermee rekening gehouden te worden. Beoordeling van de kwaliteit van de geïncludeerde studies moet bij voorkeur door tenminste twee reviewers, onafhankelijk van elkaar plaats te vinden. De eerste stap daarbij is het rangschikken van de studies op basis van het level of evidence. Vervolgens dient de methodologische kwaliteit beoordeeld te worden. Daarbij kan gebruik gemaakt worden van criterialijsten of checklisten passend bij het onderzoeksdesign. In veel studies op het gebied van management wordt gebruik gemaakt van vragenlijsten als meetinstrument. Bij deze studies dient uiteraard ook beoordeeld te worden of de gebruikte vragenlijsten gevalideerd zijn en dus meten wat ze zeggen te meten. Daarnaast dient bij deze studies beoordeeld te worden of de onderzoekspopulatie en het responspercentage representatief zijn voor de doelgroep. Indien geen of een onvolledige beoordeling van de methodologische kwaliteit heeft plaatsgevonden zijn conclusies en aanbevelingen van de review niet valide en dus onbruikbaar voor de managementpraktijk.

3.5 Is adequaat beschreven hoe data-extractie heeft plaatsgevonden en zijn de belangrijkste kenmerken van de oorspronkelijke studies beschreven?

Data-extractie is het verzamelen van de resultaten van de geselecteerde studies. Concreet gaat het daarbij om informatie over de onderzochte organisaties en de context (P + C), de aard van de interventie, model of methodiek (I) en de uitkomst (O), bij voorkeur in de vorm van een kwantificeerbare uitkomstmaat zoals een puntschatting (bijvoorbeeld gemiddelde of percentage) met een 95% betrouwbaarheidsinterval of standaardfout. Ook data-extractie dient bij voorkeur plaats te vinden door twee reviewers, onafhankelijk van elkaar. Nadat data-extractie heeft plaatsgevonden dienen de belangrijkste kenmerken van de studies beschreven te worden op een manier die aansluit bij de onderzoeksvraag, bij voorkeur in de vorm van een overzichtelijke tabel.

3.6 Is er voldoende rekening gehouden met heterogeniteit?

Om te bepalen of er sprake is van verschillen tussen de studies (heterogeniteit) zijn drie perspectieven van belang. Het eerste perspectief betreft de vraagstelling waarop een studie gebaseerd is (praktische heterogeniteit). Combinatie van de uitkomsten is alleen zinvol wanneer de PICOC van de onderzoeksvragen vergelijkbaar zijn en er dus sprake is van dezelfde type organisatie, doelgroep, interventie, meetinstrumenten en uitkomstmaten. Hierbij dient tevens kritisch gekeken te worden naar de begrippen en concepten waarvan de studies uitgaan. Zo wordt onder het begrip 'aspirine' in de meeste landen en culturen hetzelfde verstaan, maar onder het begrip 'focus' en 'actiegericht' waarschijnlijk niet (Petticrew en Roberts, 2006). De constatering dat studies in hun onderzoeksvraag dezelfde begrippen hanteren is daarom onvoldoende, ook zal beoordeeld moeten worden of de studies onder deze begrippen allemaal hetzelfde verstaan.

Het tweede perspectief betreft de onderzoeksmethodologie van de studies (methodologische heterogeniteit). Een combinatie van observationele studies, niet vergelijkende studies en enquêtes zijn methodologisch te verschillend om met elkaar te combineren en de uitkomsten van deze studies mogen dus niet gecombineerd worden (Hatala *et al*, 2005). In dat geval kunnen alleen de uitkomsten van studies met hetzelfde onderzoeksdesign gecombineerd worden en dient dit in de meta-analyse duidelijk beschreven te worden.

Het derde perspectief betreft de statistische vergelijkbaarheid van de uitkomst van de studies (statistische heterogeniteit). Zelfs wanneer de studies uitgaan van dezelfde onderzoeksvraag en hetzelfde onderzoeksdesign hanteren kunnen de uitkomsten niet zondermeer gecombineerd worden. Om de statistische vergelijkbaarheid te kunnen beoordelen kan gebruik gemaakt worden van een relatief eenvoudige toets, zoals een zogenoemde forest-plot waarin het gevonden effect en het 95 procent betrouwbaarheidsinterval van de individuele studies grafisch wordt weergegeven (Lewis en Clarke, 2001)

3.7 Is de meta-analyse correct uitgevoerd?

Wanneer de geïncludeerde studies voldoende vergelijkbaar zijn kunnen statistische technieken gebruikt worden om de uitkomsten van de individuele studies te combineren. Hierbij maakt het niet uit of het gaat om RCT's, observationeel onderzoek of niet-vergelijkend onderzoek, zolang de uitkomst van het onderzoek een puntschatting heeft en er een standaardfout berekend kan worden is combinatie van de afzonderlijke uitkomsten van de studies mogelijk.

Een voor de hand liggende manier om de resultaten van verschillende studies te combineren is het tellen van het aantal studies dat een positief effect van een interventie of succesfactor aantoont en het aantal dat een negatief effect aantoont, en vervolgens de twee uitkomsten met elkaar vergelijken. Van deze techniek, die bekend staat als 'vote counting', is sinds de jaren '70 bekend dat deze niet erg valide is omdat geen rekening gehouden wordt met essentiële zaken zoals sample size (kleine studies tellen even zwaar als grote studies), onderzoeksdesign (er wordt geen onderscheid gemaakt tussen vergelijkend onderzoek, enquêtes en casestudies) en methodologische kwaliteit (slecht uitgevoerde studies of studies met veel bias tellen even zwaar als kwalitatief goede studies) (Light en Smith, 1971). Ook wanneer voor deze verschillen wordt gecorrigeerd door gebruik te maken van weging leidt vote counting vaak tot onbetrouwbare uitkomsten (Hedges en Olkin, 1980; Hunter en Schmidt, 1990). Helaas komt vote counting nog steeds veel voor, vaak op een impliciete

manier, te herkennen aan zinnen in de conclusie als 'de meeste onderzoeken tonen aan ....'

3.8 Zijn de conclusies van de systematische review valide en toepasbaar?

De lezer zal op basis van de hierboven genoemde aspecten een inschatting moeten maken van de validiteit van de systematische review. De belangrijkste vragen die daarbij aan de orde komen zijn weergegeven in tabel 3. Indien meerdere van deze vragen met nee worden beantwoord kan er sprake van ernstige methodologische beperkingen die grote gevolgen kunnen hebben voor de validiteit.

Als de validiteit van de systematische review voldoende is, kunnen de resultaten worden beoordeeld. Zo niet, dan zijn de resultaten niet betrouwbaar. Vervolgens moet de lezer beoordelen of de conclusies en aanbevelingen van de auteur aansluiten bij de bewijskracht van de onderzochte studies. Hierbij moet opnieuw rekening gehouden worden met het onderzoeksdesign van de onderzochte studies en de hiërarchie van bewijsvoering. Zo kan op basis van de uitkomst van een meta-analyse van observationeel onderzoek geen antwoord geven worden op de vraag of de onderzochte interventie of succesfactor de oorzaak is van het gevonden effect. Hooguit kan de conclusie getrokken worden dat er tussen deze twee een verband bestaat. Of dat een causaal verband is en wat nu precies oorzaak en gevolg is, daarover kan en mag de auteur ten gevolge van de beperkte bewijskracht van observationeel onderzoek geen uitspraken doen. Ditzelfde geldt voor een systematische review van enquêtes of casestudies. Op basis van dergelijke vormen van onderzoek kunnen alleen conclusies getrokken worden over de beleving en toepasbaarheid van de onderzochte interventie of succesfactor, niet over de effectiviteit van deze twee. De auteur dient hiermee rekening te houden bij de formulering van zijn conclusie (zie tabel 1).

Tot slot dient de lezer te beoordelen of de uitkomst van de systematische review generaliseerbaar is en toepasbaar is binnen zijn eigen managementpraktijk. Voor de beoordeling van de generaliseerbaarheid dient gekeken te worden naar de vijf PICOC-elementen waarop de onderzoeksvraag van de systematische review gebaseerd is. Zo zal de lezer zelf moeten beoordelen of het resultaat van een systematische review van studies naar het effect van participatieve besluitvorming in non-profitorganisaties met veel hoog opgeleide professionals, zoals in een ziekenhuis, generaliseerbaar is naar een commerciële organisatie met voornamelijk administratieve medewerkers zoals een verzekeringsmaatschappij. Algemene richtlijnen zijn hiervoor niet te geven. Met betrekking tot de toepasbaarheid dient de lezer vooral kritisch te zijn wanneer in de conclusie en aanbevelingen geen sprake is van objectiveerbare en meetbare factoren maar van algemene termen zoals 'focus', 'alignment' of 'actiegerichtheid'. Door het hoge abstractieniveau zijn dergelijke termen weliswaar goed generaliseerbaar maar leidt het gebrek aan eenduidigheid tot een sterk verminderde toepasbaarheid. De lezer zal daarom kritisch moeten beoordelen of dergelijke abstracties voldoende handvaten bieden om toe te kunnen passen in de eigen managementpraktijk.

4 Goed beoordelen: het voorbeeld van onderzoek naar de High Performance Organisation

Om de kwaliteitsbeoordeling van een systematische review te illustreren wordt het onderzoek naar high performance organisaties van De Waal genomen. De keuze om deze systematische review als voorbeeld te nemen is vooral gelegen in de media-aandacht die naar aanleiding van de uitkomst is ontstaan. Zo is er veel over deze review gepubliceerd en heeft een aantal bekende managers het belang van de uitkomst voor de managementpraktijk benadrukt[3]. De auteur van het artikel doet sinds een aantal jaren onderzoek naar de kenmerken van zogenaamde High Performance Organisaties (HPO's): organisaties die gedurende een periode van ten minste vijf tot tien jaar betere resultaten behaalt dan concurrenten of vergelijkbare organisatie. Deel van dit onderzoek is een systematische review en meta-analyse van 91 studies waarin dergelijke organisaties zijn onderzocht. Voor de beoordeling van de kwaliteit is gebruik gemaakt van de informatie in het artikel 'Karakteristiek van de high performance organisation' gepubliceerd in Holland Management Review, een online working paper en informatie afkomstig van de auteur zelf (De Waal, 2006a; De Waal, 2006b).

4.1 De vraagstelling

Hoewel deze nergens expliciet vermeld wordt, lijkt de review uit te gaan van de volgende onderzoeksvraag: Wat zijn kenmerken van een organisatie die langdurige groei realiseert en gedurende vijf jaar betere financiële en niet-financiële prestaties behaalt dan concurrenten of vergelijkbare organisaties? Opvallend is dat de elementen van de PICOC in deze vraagstelling breed geformuleerd zijn. Zo is er sprake van 'een organisatie' (P) en wordt de context (C) niet nader gespecificeerd. De auteur gaat er dus van uit dat de kenmerken van een succesvolle autofabrikant dezelfde zouden kunnen zijn als die van een succesvol ziekenhuis. Ditzelfde geldt voor de Outcome, namelijk 'langdurige groei en gedurende vijf jaar betere financiële en niet-financiële prestaties'. Hier wordt niet geëxpliciteerd om wat voor groei (omzet?, winst?, marktaandeel?) en welke prestaties (tevredenheid van medewerkers?, kwaliteit van de dienstverlening?, aandeelhouderswaarde?) het nu precies gaat. Een dergelijk breed geformuleerde vraagstelling heeft gevolgen voor de toepasbaarheid van de uitkomst: ondui-

delijk is of deze review antwoord geeft op specifiekere vragen uit de managementpraktijk. Verder is het van belang te constateren dat niet gezocht wordt naar 'succesfactoren', waarbij sprake is van een oorzaak-gevolg relatie, maar naar 'kenmerken'. Er is dus sprake van explorerend in plaats van verklarend onderzoek. Dit betekent dat aan de uitkomst van de review geen vergaande conclusies verbonden kunnen worden.

4.2 De zoekactie

Omdat in het artikel niet vermeld staat waar precies gezocht is, is navraag gedaan bij de auteur zelf. Daarbij is gebleken dat vooral gezocht is in Business Source Premier en ScienceDirect. Omdat er geen algemeen geaccepteerde definitie van een HPO bestaat is gezocht op zoektermen als 'high performance', 'high performance workorganizations', 'high results' en 'flexible organizations'. Om selectiebias te voorkomen is zo breed mogelijk gezocht, van boeken tot journals (peer-reviewed en niet peer-reviewed) en in de Engelse, Nederlandse en Duitse taal. Daarnaast is gezocht via Google en zijn Nederlandse en buitenlandse collega's gevraagd naar studies, ook ongepubliceerd. Hieruit blijkt dat voor deze review grondig en systematisch is gezocht naar studies. Wel is de zoekactie van de auteur door het tekort aan informatie moeilijk reproduceerbaar. Een ander punt is dat niet in ABI/INFORM is gezocht, waardoor de kans aanwezig is dat studies zijn gemist.

4.3 De selectie

In het artikel worden drie criteria genoemd op basis waarvan de selectie van studies heeft plaatsgevonden. Twee daarvan betreffen de methodologische kwaliteit van een studie en worden in de meta-analyse gebruikt om de uitkomst van de afzonderlijke studies te wegen. Criteria met betrekking tot het soort organisatie, de context of het type onderzoek zijn verder niet geformuleerd. Uit informatie afkomstig van de auteur blijkt dat de selectie niet is gedaan door twee onafhankelijke reviewers. Dit is jammer, want het verhoogt de kans op selectiebias.

4.4 De kwaliteit

Ook de methodologische kwaliteit is in dit artikel door één persoon beoordeeld, wat eveneens de kans op bias vergroot. Voor de beoordeling van de methodologische kwaliteit worden twee criteria gehanteerd. Als eerste criterium wordt genoemd 'een dusdanig aantal respondenten dat de resultaten als redelijk representatief kunnen worden aangemerkt'. Hier worden twee begrippen door elkaar gehaald, namelijk de representativiteit en de betrouwbaarheid. Het aantal respondenten heeft namelijk geen directe relatie met de representativiteit, ook een grote steekproef kan onvoldoende representatief zijn. Het aantal respondenten is echter wel van invloed op de betrouwbaarheid van de resultaten, hiervoor is een eenvoudige formule beschikbaar waarmee op basis van objectieve normen de betrouwbaarheid berekend kan worden. Indien van deze formule gebruik gemaakt zou zijn zou dit de objectiviteit van de kwaliteitsbeoordeling ten goede zijn gekomen. Het tweede kwaliteitscriterium betreft de aanwezigheid van een onderzoeksverslag waarin verantwoording wordt afgelegd over 'de onderzoeksmethode, de onderzoeksaanpak en de wijze van selectie van de onderzoekspopulatie...' Helaas is geen gebruik gemaakt van objectieve normen, criterialijsten of checklisten om de kans op bias te beperken.

4.5 Data-extractie en presentatie

De data-extractie is door twee personen gedaan, onafhankelijk van elkaar. Een overzicht met de belangrijkste gegevens van de oorspronkelijke studies ontbreekt in het artikel in Holland Management Review, in de online workingpaper zijn deze wel opgenomen. Voor de bepaling van de HPO-kenmerken zijn uit de geselecteerde studies 'die elementen gehaald die volgens de onderzoekers voor organisaties belangrijk zijn om een HPO te kunnen worden'. Uit het overzicht blijkt dat deze elementen niet gekwantificeerd zijn in de vorm van een spreidingsmaat, een percentage of een getal. Meta-analyse is hierdoor helaas niet mogelijk.

4.6 Heterogeniteit

Uit het overzicht van de 91 studies blijkt dat er sprake is van grote verschillen tussen de studies, zowel qua studiedesign (case studies, surveys, kwalitatief onderzoek en zelfs persoonlijke ervaringen) als qua onderzoekspopulatie (Aziatische multinationals, Nederlandse non-profit organisaties, Amerikaanse high-tech organisaties, kleine en middelgrote Duitse bedrijven). Dit laatste hoeft op zich geen probleem te zijn, heterogeniteit met betrekking tot de onderzoekspopulatie kan de generaliseerbaarheid van de uitkomst zelfs vergroten. Voorwaarde daarbij is wel dat er met betrekking tot de gevonden uitkomst geen verschillen bestaan tussen de subgroepen. Zo is het niet ondenkbaar dat de HPO-kenmerken van een Nederlands ziekenhuis verschillen van de HPO-kenmerken van een Amerikaanse beursgenoteerde fabrikant van microchips. Een subgroepanalyse kan hierover uitsluitsel geven, maar uit het artikel wordt niet duidelijk of deze heeft plaatsgevonden. De resultaten van de 91 studies hadden dus niet gecombineerd mogen worden tot een totaalresultaat. Uit de conclusie van de online working-paper blijkt dat de auteur zich hiervan bewust is: 'There is also the issue of "apples and pears": studies of a different kind have all been lumped together, making the results of the comparison potentially incomparable' (De Waal, 2006b). Het is jammer dat deze belangrijke opmerking in het Nederlandse artikel achterwege is gebleven.

4.7 Meta-analyse

Omdat er geen statistische technieken zijn gebruikt om de uitkomsten van de individuele studies te combineren is in deze review strikt genomen geen sprake van een meta-analyse. Wel zijn de uitkomsten van de geselecteerde studies op een kwalitatieve manier gecombineerd tot een totaalresultaat. Hiervoor zijn 'de elementen' die uit de studies naar voren zijn gekomen geclusterd en gewogen op basis van de twee eerdergenoemde (subjectieve) kwaliteitscriteria. Vervolgens is op basis van de weging voor ieder cluster een totaalscore berekend en zijn de clusters met de hoogste score door de auteur aangemerkt als HPO-kenmerk. Dit lijkt een voor de hand liggende manier om de uitkomst van verschillende studies te combineren, maar is feitelijk een subjectieve vorm van 'vote counting' en dus niet erg betrouwbaar.

4.8 Conclusie

Uit de bovenstaande beoordeling blijkt dat een groot aantal vragen zoals weergegeven in tabel 3 met 'nee' beantwoord moet worden. De validiteit van deze review is dus onvoldoende om de resultaten als betrouwbaar te kunnen interpreteren. Om deze reden moet geconcludeerd worden dat de uitkomst van de systematische review en de aanbevelingen van de auteur vooralsnog beperkt bruikbaar zijn voor de managementpraktijk. Of zoals de auteur zelf terecht opmerkt: 'Further research should focus on validating the characteristics found in this study...' (De Waal, 2006b)

5 Tot slot

Het onderzoek in het voorbeeld verdient grote waardering omdat er een serieuze en toegewijde poging is gedaan de kennis en inzichten rond een voor managers relevant onderwerp bijeen te brengen. Dit neemt niet weg dat het onderzoek ook beoordeeld moet worden met de kennis en kunde die daarvoor voorhanden is. Zeker omdat dit onderzoek impact heeft op managers en in de publiciteit gekoppeld wordt aan een forse claim; veel managers lopen er mee weg en er wordt gesproken over 'de heilige graal van het management'. Het vak komt verder als meer auteurs bereid en in staat zijn dergelijke inspanningen te leveren en deze te koppelen aan de regels die gelden voor de productie en beoordeling van systematische reviews. Als dit verder vorm krijgt wordt ook het managementvak verder geprofessionaliseerd. Dan wordt door wetenschappers een verdere invulling gegeven aan de opdracht die Drucker aan dat vak heeft meegegeven: 'systematisch en methodisch te doen wat vroeger op gevoel en intuïtief werd gedaan, tot beginselen en begrippen te herleiden wat aan de ervaring werd overgelaten en incidentele kennis te vervangen door een logisch schema, dat innerlijke samenhang vertoont' (Drucker, 1957, p. 313). ■

E.G.R. Barends is seniorresearcher bij Ten Have Change Management en werkzaam als verandermanager in de gezondheidszorg.
Dr. R.W. Poolman is orthopedisch chirurg en werkzaam in het Onze Lieve Vrouwe Gasthuis te Amsterdam, hij is medeinitiator van de International Evidence Based Orthopaedic Surgery Working Group en als reviewer verbonden aan o.a. BMJ en Biomed Central.
Dr. D.T. Ubbink is arts en epidemioloog en als programmaleider werkzaam bij de afdeling Kwaliteit & Proces Innovatie, expertisegroep Evidence-based practice & Implementatie van het Academisch Medisch Centrum te Amsterdam en o.a. editor bij de Wounds Review Group van de Cochrane Collaboration.
Prof. dr. mr. S. ten Have is organisatieadviseur en partner van Ten Have Change Management, hij is als hoogleraar Strategie en Verandering verbonden aan de Vrije Universiteit Amsterdam.

Literatuur

■ Antman, E.M. 1992. A comparison of results of meta-analyses of randomized controlled trials and reccommendations of clinical experts. *Journal of the American Medical Association*, 286(2): pp. 240-8.
■ Armenakis, A., Bedeian, A. 1999. Organizational Change: A Review of Theory and Research in the 1990s. *Journal of Management*, 25(3): pp. 293-315.
■ Arthur, W.J., Bennett, W.J., Edens, P., Bell, S.T. 2003. Effectiveness of training in organizations: A meta-analysis of design and evaluation features. *Journal of Applied Psychology*, 88: pp. 234-45.
■ Barends, E., Ten Have, S. 2008. Op weg naar evidence-based management. *Holland Management Review*(120): pp. 45-51.
■ Bushman, B., Wells, G. 2001. Narrative impressions of literature: The availability bias and corrective properties of meta-analytic approaches. *Personality and Social Psycholgy Bulletin*, 27(9): pp. 1123-30.
■ *Business Week*. Business Week. 1984. Oops. Who's excellent now?, November 5: 76-88.
■ Collins, D.B., Holton, E.F. 2004. Human Resource Development Quarterly. *The effectiveness of managerial leadership development programs: A meta-analysis of studies from 1982 to 2001*, 15: pp. 217-48.
■ Collins, J., Porras, J. 1995. *Built to Last: Succesful Habits of Visionary Companies*. New York: Harper Business.
■ Collins, J.C. 2001. *Good to Great: Why Some Companies Make the Leap...and Others Don't*. London: Random House Books.
■ De Waal, A. 2006a. Karakteristiek van de high performance organisation. *Holland Management Review* (107): pp. 18-25.
■ De Waal, A. 2006b. The Characteristics of a High Performance Organisation.: Social Science Research Network, Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=931873.
■ Dickersin, K. 1990. The existence of publication bias and risk factors for its occurrence. *JAMA*, 263(10): pp. 1385-9.

■ Drucker, P. 1957. *Management in de praktijk*. Bussum: G.J.A. Ruys Uitgeversmaatschappij.

■ Egger, M., Smith, G. 1998. Bias in location and selection of studies. *British Medical Journal*, 316(7124): pp. 61-6.

■ Egger, M., Smith, G., Schneider, M., Minder, C. 1997. Bias in meta-analysis detected by a simple, graphical test *British Medical Journal*, 315(7109): pp. 629-34.

■ Guyatt, G., Oxman, A.D., Kunz, R., Vist, G.E., Falck-Itter, Y., Schünemann, H.J., for the GRADE Working Group. 2008. What is 'quality of evidence' and why is it important to clinicians? *British Medical Journal*, 336(7651): pp. 995-98.

■ Guyatt, G., Sackett, D., Sinclair, J., Hayward, R., Cook D., and Cook, R. 1995. User's guides to the medical literature. IX. A method for grading health care recommendations. *Journal of the American Medical Association*, 274: pp. 1800-4.

■ Hatala, R., Keitz, S., Wyer, P., Guyatt, G. for the Evidence Based Medicine Teaching Tips Working Group. 2005. Tips for learners of evidence-based medicine: 4. Assessing heterogeneity of primary studies in systematic reviews and wether to combine their results. *Canadian Medical Association Journal*, 172(5): pp. 661-5.

■ Hedges, L., Olkin, I. 1980. Vote-counting methods in research synthesis. *Psychological Bulletin*, 88(2): pp. 359-69.

■ Higgins, J., Green, S., editor. 2006. *Cochrane Handbook for Systematic Reviews of Interventions 4.2.6 [updated September 2006]*. Chichester, UK: John Wiley & Sons, Ltd.

■ Hunter, J., Schmidt, F. 1990. *Methods of meta-analysis. Correcting error and bias in research findings.* California: Sage Publications.

■ Lewis, S., Clarke, M. 2001. Forest plots: trying to see the wood and the trees. *British Medical Journal*, 322.

■ Light, R., Smith, P. 1971. Accumulating evidence: Procedures for resolving contradictions among different research studies. *Harvard Educational Review*, 41(4): pp. 429-71.

■ Macy, B. A., Peterson, M. F., Norton, L. W. 1989. A test of participation theory in a work re-design field setting: Degree of participation and comparison site contrasts. *Human Relations, , No. 12*, 42(12): 1095-165

■ McAuley, L., Pham, B., Tugwell, P., Moher, D. 2000. Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? *The Lancet*, 356(9237): pp. 1228-31.

■ Moher, D., Cook, D., Eastwood, S., Olkin, I., Rennie, D., Stroup, D., for the QUORUM Group. 1999. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. *The Lancet*, 354(9193): pp. 1896-900.

■ Offringa, M., Assendelft, W, Scholten, R. 2008. *Inleiding in evidence-bases medicine, klinisch handelen gebaseerd op bewijsmateriaal*. Houten / Antwerpen: Bohn Stafleu van Loghum.

■ Olson, C., Rennie, D., Cook, D., Dickersin, K., Flanagin, A., Hogan, J. W., Zhu, Q., Reiling, J., Pace, B. 2002. Publication Bias in Editorial Decision Making. *JAMA*, 287(21): pp. 2825-28.

■ Peters, T.J., Waterman, R.H. 1982. *In Search of Excellence: Lessons from America's Best-Run Companies*. New York: Harper & Row.

■ Petticrew, M. 2001. Systematic reviews from astronomy to zoology: myths and misconceptions. *British Medical Journal*, 322(13): 98-101.

■ Petticrew, M., Roberts, H. 2003. Evidence, hierarchies and typologies: Horses for Courses. *Journal of Epidemiology ans Community Health*, 57.

■ Petticrew, M., Roberts, H. 2006. *Systematic Reviews in the Social Sciences*. Oxford UK: Blackwell Publishing.

■ Pfeffer, J. and Sutton, R. 2006. Evidence-Based Management. *Harvard Business Review*, 84(1): pp. 63-74.

■ Phillips, B., C. Ball, D. Sackett, D. Badenoch, S. Straus, B. Haynes, M. Dawes. 2001. Levels of Evidence: Oxford Centre for Evidence-based Medicine (http://www.cebm.net/index.aspx?o=1025).

■ Poolman RW, Kerkhoffs GM, Struijs PA, Bhandari M, International Evidence-Based Orthopedic Surgery Working Group. 2007. Don't be misled by the orthopedic literature : tips for critical appraisal. *Acta Orthopaedica*, 78(2): 162-71.

■ Rosenthal, R. 1979. The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3): pp. 638-41.

■ Rosenzweig, P. 2007. *The Halo Effect and the Eight Other Business Delusions That Deceive Managers*. New York: Free Press.

■ Rousseau, D.M. 2006. Is there such a thing as evidence-based management? *Academy of Management Review*, 31(2): pp. 256-69.

■ Sterling, T. 1959. Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance - Or Vice Versa. *Journal of the American Statistical Association* 54(285): pp. 30-34.

■ Stroup, D.F., Berlin, J.A., Morton, S.C., Olkin, I., Williamson, D., Rennie, D., Moher, D., Becker, B.J., Sipe, T., Thacker, S.B. for the Meta-analysis Of Observational Studies in Epidemiology (MOOSE) Group. 2000. Meta-analysis of Observational Studies in Epidemiology. *JAMA*, 283(15): pp 2008-12.

■ Ten Have, S. 2007. De quintessens van de high performance organisation: het kennen van de eigen organisatie en context als basis voor goed organiseren. *M&O, tijdschrift voor Management & Organisatie*, 51(7): pp. 5-20.

■ Ubbink, D. T., Legemate, D. A. 2004. Evidence-based surgery. *British Journal of Surgery*, 91(9): 1091-2.

Noten

1 Wie een groot aantal mogelijke verbanden onderzoekt, vindt bij een significantie van 0,05 namelijk altijd wel een of meer significante relaties. Dit fenomeen wordt data-dredging of kanskapitalisatie genoemd.

2 Bij vergelijkend onderzoek worden twee of meer groepen met elkaar vergeleken, meestal een groep waarbij een interventie wordt gepleegd (interventiegroep) en een groep waarbij geen of alternatieve interventie wordt gepleegd (controlegroep). Bij randomisatie worden de groepen die met elkaar worden vergeleken volledig willekeurig (random) samengesteld, bijvoorbeeld door middel van loting. Hierdoor heeft iedere deelnemer (of andere eenheid zoals team, afdeling of bedrijf) evenveel kans om in de interventie- of de controlegroep te komen. Op deze wijze wordt de invloed van eventuele verstorende factoren over beide groepen verdeeld zodat deze groepen, afgezien van de interventie, zo goed mogelijk vergelijkbaar zijn.

3 Voor een lijst met publicaties en een overzicht van de managers die de uitkomsten van de review promoten wordt verwezen naar www.hpocenter.nl.