
Positionering van de Evaluatiemethodologie van het Onderwijstoezicht

Uitgangspunten voor de inrichting van het toezicht op de
kwaliteit en deugdelijkheid van het onderwijs

Frans J.G. Janssens & Anne Bert Dijkstra

Universiteit Twente • Universiteit van Amsterdam • 1 februari 2013

‘Als je er geen verstand van hebt, ga dan ook niet kijken’

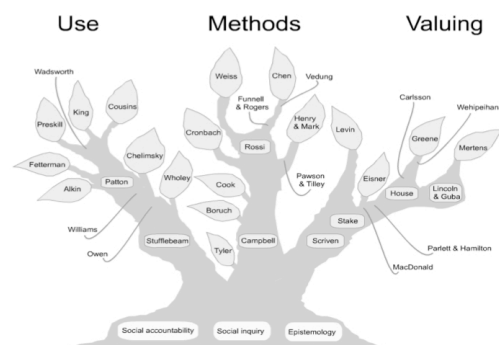
Ben Ale, hoogleraar Veiligheidskunde TU Delft



‘Telling the truth to people who may not want to hear it is, after all, the
chief purpose of evaluation’

Eleanor Chelimsky, US Government Accountability Office

Inhoudsopgave



INLEIDING	3
1 TOEZICHT EN METHODOLOGIE	5
2 METHODOLOGISCHE TYPERING ONDERWIJSTOEZICHT	8
2.1 EVALUATIE OF ONDERZOEK?	8
2.2 EVALUATIETHEORIEËN	11
2.3 WETENSCHAPSFILOSOFISCHE ACHTERGRONDEN	12
3 CATEGORISERING VAN EVALUATIETHEORIEËN	14
3.1 EVALUATIEBOOM	14
3.2 VERTAKKINGEN	16
3.3 INVLOED VAN TOEZICHT OP EVALUATIETHEORIEËN	17
4 METHODOLOGISCHE PARADOXEN	20
4.1 ONTWIKKELINGEN IN HET ONDERWIJSTOEZICHT	20
4.2 INSPECTIEPARADOXEN	21
4.3 ACHTERLIGGENDE REDENERING	21
5 DE EVALUATIEMETHODOLOGIE VAN DE INSPECTIE	25
5.1 RATIONALE	25
5.1.1 UITGANGSPUNTEN	25
5.1.2 CONCEPTUEEL KADER	26
5.2 INSTRUMENTARIUM	28
5.2.1 WAARDERINGSKADERS	28

5.2.2 NORMERING	30
5.2.3 DATA-VERZAMELING	31
6 DE INSPECTIEWERKWIJZE IN HET EVALUATIELANDSCHAP	33
6.1 METHODE-DIMENSIE	35
6.2 POSITIONERING INSPECTIE	38
6.3 WAARDEN-DIMENSIE	39
6.4 POSITIONERING INSPECTIE	41
6.5 GEBRUIKSDIMENSIE	44
6.6 POSITIONERING INSPECTIE	46
LITERATUUR	49

INLEIDING

Onderwijstoezicht bestaat voor een belangrijk deel uit beoordelingen van de kwaliteit en de deugdelijkheid van het onderwijs. In dat opzicht is toezicht een vorm van onderwijsevaluatie. Onderwijstoezichthouders zijn overigens niet de enigen die het onderwijs evalueren. Evaluaties vinden ook buiten het onderwijs plaats en worden ook uitgevoerd in (toegepaste) wetenschappelijke contexten. Uniek voor het onderwijstoezicht is wel dat evaluatie een wettelijk verankerde taak is binnen het Openbaar bestuur. Ten opzichte van andere partijen die evalueren heeft deze positie wel gevolgen voor de manier waarop de inspectie evaluaties uitvoert, al is het alleen maar omdat het werk van de inspectie de publieke zaak dient. Besturen, scholen, ouders, deelnemers, de Minister en de Staten-Generaal moeten nu eenmaal staat kunnen maken op de oordelen van de inspectie.

Evaluaties door de onderwijsinspectie zijn ook nog om een andere reden uniek te noemen. Ze worden niet uitgevoerd door een team van wetenschappers maar door materiedeskundigen die op basis van hun expertise tot oordelen komen over de kwaliteit en deugdelijkheid van het onderwijs. Dit betekent niet dat de uitvoering van evaluaties door de inspectie niet aan principes en regels is gebonden of subjectief van aard zijn, maar ook niet dat de kwaliteit van de evaluaties buiten kijf staat. Er zitten twee spanningsvelden in inspectie-evaluaties. Aan de ene kant is de gegevensverzameling en de oordeelsvorming in hoge mate een individuele aangelegenheid van een inspecteur, terwijl tegelijkertijd rechtsgelijkheid, fairness, betrouwbaarheid en validiteit wordt nagestreefd. Dat leidt tot het streven werkwijzen te standaardiseren. Maar aan de andere kant leidt standaardisering van werkwijzen tot verlies van het zicht op de uniciteit van scholen en instellingen.

Evalueren kent een lange traditie en is daarom ook onderwerp van wetenschapstheoretische en methodologische beschouwingen over hoe een evaluatie dient te worden opgezet en uitgevoerd. Sinds het begin van de vorige eeuw hebben vele wetenschappers bijdragen geleverd aan de wetenschappelijke onderbouwing van waarom en hoe er geëvalueerd moet worden. Dit heeft een evaluatielandschap opgeleverd van zienswijzen over de te volgen evaluatiemethode, het soort kennis dat opgeleverd moet worden, de waarden die een rol in de oordeelsvorming spelen en over het gebruik van evaluatie-uitkomsten.

Het voorliggende paper geeft een overzicht van de positie die de werkwijze van de onderwijsinspectie inneemt in het landschap van evaluatietheorieën voor wat de betreft drie dimensies: methode, waarden en gebruik. Deze positionering is niet alleen van belang om de wetenschappelijke kwaliteit van inspectie-evaluaties te beoordelen, maar ook om strategische keuzes te maken in de verdere fundering en ontwikkeling van de wijze waarop de inspectie de kwaliteit en deugdelijkheid van het onderwijs kan bepalen en beoordelen.

1 TOEZICHT EN METHODOLOGIE

In de literatuur over toezicht wordt doorgaans een definitie gehanteerd die steeds ongeveer als volgt luidt (zie bijv. Algemene Rekenkamer, 2002; Janssens, 2005; Robben, 2010; De Winter, 2010 en Mertens, 2011):

Toezicht betreft het verzamelen van informatie over de vraag of een zaak of handeling aan de daaraan gestelde eisen voldoet, het vormen van een oordeel daarover en het eventueel naar aanleiding daarvan interveniëren.

Ook internationaal is er overeenstemming over de definitie van toezicht. Wilcox (2000, p18) definieert in UNESCO-verband toezicht als *‘The systematic collection and interpretation of evidence leading (.....) to a judgment of value with a view to action.’* De World Bank (2010) omschrijft toezicht als *‘the regular/periodic oversight of individuals or entities, which uses the results of evaluation (and sometimes inspection) to inform and direct action of those supervised’.*

De kern van de werkwijze van toezicht is de drieslag: informatie verzamelen, oordelen en interveniëren. Om effectief te kunnen interveniëren moet de te verzamelen informatie antwoord kunnen geven op de vraag of en welke interventie nodig is en daarvoor zijn valide criteria en gefundeerde oordelen nodig. Het zal niemand verbazen dat er regels en principes aan dit proces verbonden zijn en dat het op een te verantwoorden wijze dient te verlopen. Er ligt dus een methodologie aan te grondslag.

De inhoud van het begrip methodologie is niet eenduidig en ook niet onomstreden. Oorspronkelijk was met name in de Angel-Saksische landen, waar men doorgaans spreekt van ‘research methodology’, het vak breed gedefinieerd en omvatte dit het hele domein van reflectie over, methoden voor en beoordeling van wetenschappelijk onderzoek. Dit laatste met inbegrip van een wetenschapsfilosofische reflectie op praktijken en resultaten van wetenschapsbeoefening. In Nederland wordt het vak al decennia lang ‘Methoden en Technieken van Onderzoek’ genoemd en heeft het, zoals de titel al aangeeft, een aanzienlijk smallere betekenis. Een van de belangrijkste vragen die men op dit gebied kan stellen, is of deze zich bekend tot de smalle of de brede opvatting van methodologie. Om deze vraag te beantwoorden moeten we terug willen gaan in de historie, naar de zeventiende eeuwse Franse

filosoof René Descartes (Verschuuren, 2002). Hij publiceerde in 1637 zijn meest beroemde werk ‘*Discours de la méthode, pour bien conduire sa raison, et chercher la vérité dans les sciences*’. Vertaald in het Nederlands; discussie over de methode, om het verstand in goede banen te geleiden, en de waarheid te zoeken in de wetenschap. Ook Nederlands meest bekende methodoloog, A.D de Groot, geeft zijn klassieke werk over methodologie als ondertitel mee ‘grondslagen van onderzoek *en denken* in de gedragswetenschappen’ (cursivering van de auteurs) (De Groot, 1961).

Als een der eersten was Descartes niet alleen bezig met het verkennen van de ons omringende werkelijkheid. Hij verdiepte zich bovendien in de vraag waar men bij deze zoektocht op moet letten, welke wegen moeten worden bewandeld en hoe de vele valkuilen die men op zijn zoektochten naar ware kennis kan tegenkomen, kunnen worden vermeden. Het is de kern van de methodologie in de brede zin: ‘weg waarlangs’ (men tot adequate en bruikbare kennis kan komen).

Afhankelijk van datgene wat men wil bereiken zijn er diverse soorten methodologie. De meest gangbare vorm is de onderzoeksmethodologie, waar de vraag wordt beantwoord hoe we kunnen komen tot geldige en bruikbare kennis via empirisch onderzoek. Daarnaast is er de interventiemethodologie, die zich bezighoudt met het formuleren van richtlijnen voor het handelend ingrijpen in de bestaande werkelijkheid en ontwerpmethodologie die regels en procedures opstelt die zijn gericht op het creëren van nieuwe materiële of immateriële zaken (Verschuuren, 2002).

Ofschoon toezicht geen wetenschap is, is de werkwijze wel gebaseerd op de principes van sociaal-wetenschappelijk onderzoek, nl. dat de gegevensverzameling en de oordeelsvorming objectief, betrouwbaar en valide moeten zijn. Daarmee is toezicht te beschouwen als een vorm van systematisch empirisch onderzoek dat zich onderscheidt van andere vormen van onderzoek, zoals bijvoorbeeld fundamenteel wetenschappelijk onderzoek dat vooral de theorievorming dient. Cronbach & Suppes (1969) spreken in dit verband van ‘*disciplined inquiry*’; een vorm van professioneel speurwerk waarbij sprake is van regelgeleide methoden & technieken om kennis te verzamelen, zoals bijvoorbeeld toegepast in de journalistiek, het reberchewerk, de rechtspraak en in het sociaal-wetenschappelijk onderzoek. De geestelijke vaders van het begrip *disciplined inquiry*, Cronbach & Suppes (1969, p. 15 e.v.), merken hierover op (zie ook Shulman, 1997) :

“Disciplined inquiry has a quality that distinguishes it from other sources of opinion and belief. The disciplined inquiry is conducted and reported in such a way that the argument can be painstakingly examined. The report does not depend for its appeal on the eloquence of the

writer or on any surface plausibility. The argument is not justified by anecdotes or casually assembled fragments of evidence. Scholars in each field have developed traditional questions that serve as touchstones to separate sound argument from incomplete or questionable argument.

Het kenmerk van ‘disciplined inquiry’ is dat er tijdens de gegevensverzameling en het daarop volgende redeneerproces controles zijn ingebouwd om geen foute of discutabele conclusies te trekken. Als er voor bepaalde fouten niet kan worden gecorrigeerd dan worden ze expliciet gemaakt door de gevolgen ervan voor de conclusies in beschouwing te nemen. “.....*But... fundamental to disciplined inquiry is its central attitude, which places a premium on objectivity and evidential test.*” (Cronbach & Suppes, 1969, p. 18).

Het moge duidelijk zijn dat toezicht geen wetenschappelijk onderzoek is maar wel een activiteit die volgens regels en principes onderzoek dient te verlopen. Deze vertonen in hoge mate overeenkomst met de regels en principes uit het sociaal-wetenschappelijk onderzoek.

2 METHODOLOGISCHE TYPERING ONDERWIJSTOEZICHT

2.1 Evaluatie of onderzoek?

De Engelse taal is verschillende woorden rijk waarmee onderscheid gemaakt kan worden tussen verschillende vormen van onderzoek, zoals research, evaluation, inquiry, testing, examination en investigation. In het Nederlands kennen we alleen het onderscheid tussen onderzoek en evaluatie. We hebben zojuist vastgesteld dat toezicht geen wetenschappelijk onderzoek in strikte zin is, maar wel een activiteit is waarvoor methodische regels en kwaliteitseisen gelden voor de gegevensverzameling en de oordeelsvorming. Maar hoe valt toezicht nu in methodologische zin te typeren?

De kern van de werkwijze van toezicht is de cyclus: informatie verzamelen, oordelen en interveniëren. Deze cyclus komt in hoge mate overeen met theorieën over onderwijsevaluatie (zie Janssens, 1995; Wilcox, 2000; Scheerens, Glas & Thomas, 2003; Mertens, 2011 en Dijkstra & Janssens, 2012), maar ook, en dat is niet zo verwonderlijk, met theorieën over accountability (zie bijv. Alkin & Christie, 2004; Wagner, 1998; Mark, Henry & Julnes, 2000; Knoops, 2010).

Fournier (2005, p. 139-140) definieert evaluatie als volgt¹:

“Evaluation is an applied inquiry process for collecting and synthesizing evidence that culminates in conclusions about the state of affairs, value, merit, worth, significance, or quality of a program, product, person, policy, proposal, or plan. Conclusions made in evaluations encompass both an empirical aspect (that something is the case) and a normative aspect (judgment about the value of something). It is the value feature that distinguishes evaluation from other types of inquiry, such as basic science research, clinical epidemiology, investigative journalism, or public polling.”

¹ Er zijn ook andere definities mogelijk maar centraal daarin staat in alle gevallen dat er gegevens worden verzameld worden om te kwalificeren en ergens over te oordelen (zie Mertens & Wilson, 2012, p. 9-11).

Toezicht is dus een vorm van onderwijsevaluatie omdat de kern ervan is nagaan en beoordelen of de onderwijspraktijk in overeenstemming is met:

- wet- en regelgeving;
- de eisen die aan de inrichting, organisatie en kwaliteit van het onderwijs worden gesteld;
- het onderwijsbeleid van de regering;
- het eigen beleid van de scholen en instellingen.

Er zijn duidelijke verschillen maar ook overeenkomsten tussen onderwijsevaluatie en onderwijsonderzoek en dat is niet alleen maar een semantische kwestie (Mertens & Wilson, 2012). Daarom wagen we een poging deze twee activiteiten te definiëren.

Onderwijsevaluatie is gericht op het bepalen van de waarde of ‘hoe-goedheid’ of waarde van ‘iets’, hetzij intrinsiek hetzij in relatie tot een bepaald criterium (Scriven, 1976). Bij onderwijsevaluatie gaat het dan om de waardering van een curriculum, scholingsconcept, leerprestaties of om onderwijsprocessen. Om te kunnen evalueren is gegevensverzameling nodig alsmede methoden om te oordelen, waaronder 1) criteria, standaarden of normen om de kwaliteit te kunnen beoordelen en de beslissing of deze relatief of absoluut moeten zijn, 2) verzamelen van relevante gegevens en 3) toepassen van de criteria, standaarden of normen om de kwaliteit te kunnen vaststellen. Het object van evaluatie kan de onderwijspraktijk zijn of het beleid om de onderwijspraktijk te beïnvloeden.

Onderwijsonderzoek is gericht op het verkrijgen van generaliseerbare kennis door enerzijds de veronderstelde relatie tussen variabelen (hypothesen of modellen) te toetsen of door generaliseerbare verschijnselen (casus) te beschrijven. De uitkomst van onderwijsonderzoek, zoals theorieën, functionele relaties of beschrijvingen, wordt verkregen door toepassing van empirische of andere systematische methoden en kan wel of niet direct toepasbaar zijn in de onderwijspraktijk.

Onderwijsonderzoek is dus een complexere methodologische activiteit die boven evaluatie uitstijgt, omdat het bij evalueren enkel gaat om het waarderen van een object, terwijl onderzoek gericht is op het vaststellen van causale relaties binnen het object van onderzoek.

Ook komt het begrip ‘evaluatie-onderzoek’ voor waardoor de verwarring alleen maar groter wordt.

Evaluatie en onderzoek hebben vele raakvlakken, maar verschillen van elkaar op één belangrijk onderdeel: de beoordeling van kwaliteit (Scriven, 2013). Beide activiteiten trekken een zware wissel op empirische methoden en technieken. Zo kunnen bijvoorbeeld zowel in

onderzoek als evaluaties de meting van leerprestaties een belangrijke effectmaat zijn. In beide gevallen is de productie van nieuwe kennis aan de orde. Maar ook onderwijsonderzoek kent zijn verschillende vormen. Een belangrijk onderscheid is bijvoorbeeld fundamenteel versus toegepast onderzoek, ofschoon hier misschien eerder sprake is van een continuüm. Maar in beide gevallen is het onderzoek gericht op generaliseerbare kennis. Bij toegepast onderzoek gaat het dan in het bijzonder om het genereren van een oplossing voor een algemeen probleem. Echter, onderwijsevaluatie is wat anders dan onderzoek van onderwijs (Stake & Denny, 1969; Worthen & Sanders, 1987).

Toezicht, voorzover het betrekking heeft op het beoordelen of de onderwijspraktijk in overeenstemming is met hetgeen is bedoeld, is dus een vorm van onderwijsevaluatie. Die evaluatieve taak van de inspectie kan, met een parafrase op de eerder gebruikte definitie van toezicht, gedefinieerd worden als ‘het systematisch beoordelen van de kwaliteit van het onderwijs op basis van systematisch vergaarde informatie waarvan de oordelen bruikbaar moeten zijn in en voor het onderwijs’ (Janssens, 1997).

Niet alleen de uitkomst in de vorm van een waardering of beoordeling is kritisch voor de vraag of er wel of geen sprake is van een evaluatie, maar ook het startpunt is van groot belang: de evaluatievraag. Niet alle vragen die leiden tot gegevensverzameling of onderzoek zijn evaluatief van aard. Een vraag als ‘Wat is de kwaliteit van....’ is duidelijk een evaluatievraag omdat het uiteindelijk om een oordeel of kwalificatie vraagt. Ook de vraag of A beter is dan B, is een zuivere evaluatievraag. Een vraag als ‘Wat zijn de oorzaken van schooluitval’, is wel een onderzoeksvraag maar vraagt niet om een beoordeling en is daarom geen evaluatievraag. Dat geldt ook voor de vragen als ‘Hoe wordt ICT gebruikt in het onderwijs’, ‘Hoeveel thuiszitters kent Nederland?’ en ‘Welke rol speelt het basisschooladvies bij de plaatsing van leerlingen in het VO?’².

Een veel gemaakte ‘fout’ bij evaluatie is het verschaffen van empirische data alsof dit het antwoord op een evaluatievraag is (Scriven, 2013, p. 169). De constatering dat de opbrengsten

² Interessant bij dit soort vragen is de kwestie of deze door de toezichthouder moeten worden beantwoord, dus of hier sprake is van toezicht. Van toezicht is sprake als de gegevensverzameling gericht is op nagaan of de praktijk in overeenstemming is met een gewenste of vereiste situatie en dat, waar nodig, acties worden ondernomen om de situatie te veranderen. Een vraag als ‘Hoeveel scholen gebruiken de code Goed Bestuur?’ is dus in principe een inventariserende vraag die niet perse en ook niet uitsluitend door de toezichthouder hoeft te worden beantwoord, tenzij tevens wordt beoordeeld of de scholen zich er ook aan houden en er eventuele acties aan zijn verbonden gericht op de handhaving.

van een school drie jaar volgens verwachting zijn is ten principale geen antwoord op de evaluatievraag ‘Wat is de kwaliteit van het onderwijs op die school’³.

2.2 Evaluatietheorieën

Het zal niemand verbazen dat er van onderwijsevaluatie uiteenlopende definities bestaan en dat het ook voorwerp is van ideologische discussies. Verschillende gezaghebbende auteurs op dit terrein verschillen diepgaand van mening over wat onderwijsevaluatie is en hoe dat moet worden uitgevoerd. Deze evaluatietheorieën bevatten concepten, definities en assumpties waarmee de relaties binnen een evaluatieobject kunnen worden beschreven of verklaard (Alkin, 2013, p. 4). Gedurende de laatste decennia zijn er op basis van verschillende evaluatietheorieën meer dan 50 evaluatiemodellen ontstaan (Scriven, 1991; Worthen & Sanders, 1987; Alkin & Christie, 2004; Christie & Alkin, 2008; Christie & Alkin, 2013; Mertens & Wilson, 2012). Onder een evaluatietheorie wordt verstaan een set concepten, definities en assumpties die betrekking hebben op de wijze waarop de aard van het object van evaluatie te begrijpen is, hoe daaraan een bepaalde waarde gegeven kan worden, hoe daarover gegevens kunnen worden verzameld en hoe deze gebruikt kunnen worden (Donaldson & Lipsey, 2006)

Sommige auteurs zien evaluatie alleen maar als een instrument om relevante gegevens te verzamelen, opdat anderen (stakeholders) erover kunnen oordelen of gefundeerde beslissingen kunnen nemen. Patton (2008) en Weiss (1977) bijvoorbeeld stellen de evaluatie in dienst van de opdrachtgever en vinden dat deze ook de oordelen velt en de beslissingen neemt en niet de evaluatoren. Anderen zijn van mening dat de evaluatie pas compleet is als er ook oordeelsvorming aan te pas komt en dat de opdrachtgever de beslissingen neemt (zie bijv. Scriven, 1980; Worthen & Sanders, 1987). Deze verschillen van inzicht hebben uiteraard praktische implicaties voor de uitvoering van evaluaties.

Anderen leggen vooral de nadruk op het soort gegevens dat nodig (de kenniscomponent) is voor de oordeelsvorming en op de wijze van gegevens verzamelen. Dit betreft de discussie met de enigszins misleidende titel ‘kwantitatieve versus kwalitatieve evaluatie’. Aan de ene kant van het spectrum bevinden zich empirisch-analytisch georiënteerde theoretici die voorstander zijn van (quasi-)experimentele evaluaties, zoals bijvoorbeeld Campbell en Stanley (1966) en Cronbach (1982). Aan de andere kant van het spectrum zitten naturalistisch-interpretatief georiënteerde theoretici die voorstander zijn van een etnografische of

³ Scriven (2013, p. 178) wijst in dit verband op ‘missing link errors’ omdat voor de beantwoording van de onderliggende evaluatievraag componenten in de causale keten ontbreken, zoals de vraag of de opbrengsten met betrouwbare toetsen of examens zijn gemeten, of er niets onrechtmatig is gebeurd tijdens afname en scoring, hoe het onderwijs is geweest dat tot de uitkomsten heeft geleid, welk aanname- en doorstroombeleid de school voert, etc., etc..

narratieve aanpak, zoals bijvoorbeeld Stake (2010) en Guba en Lincoln (1989). De kern van deze discussie betreft niet het gebruik van kwantitatieve of kwalitatieve gegevens, maar heeft meer te maken met wetenschapsfilosofische verschillen en dus vooral met de interpretatie van gegevens (zie Janssens, 1985). Daarom zijn binnen deze discussie ook andere typeringen van evaluatie in omloop zoals bijvoorbeeld naturalistische, responsieve of narratieve evaluatie.

In feite is deze discussie niet meer aan de orde omdat men allengs is gaan inzien dat hier geen sprake is van een tegenstelling maar van een continuüm en dat ze complementair aan elkaar kunnen worden ingezet (Mertens & Wilson, 2012).

Ondanks de soms grote verschillen in theoretische noties over onderwijsevaluatie is over het algemeen geaccepteerd dat evalueren meer is dan beschrijven maar dat het uiteindelijke doel is het beoordelen van de waarde van iets.

2.3 Wetenschapsfilosofische achtergronden

Met enige verbazing kan men zich afvragen hoe het komt dat er zoveel uiteenlopende visies zijn op onderwijsevaluatie. Dat komt omdat de theoretici op dit gebied verschillende achtergronden hebben en ook verschillende visies op onderwijs en op wat relevante gegevens zijn op basis waarvan geëvalueerd kan worden. Deze filosofisch en ideologische verschillen vinden we natuurlijk ook op andere terreinen, zoals leren en onderwijzen, onderwijskundig leiderschap en onderwijsverbetering (Kelly, 2006).

Voor het domein onderwijsonderzoek en onderwijsevaluatie geldt ook dat wetenschapsfilosofische verschillen leiden tot verschillen in opvattingen over wat kennis is, hoe kennis tot stand komt en wat de waarde van iets is (zie voor een wetenschapsfilosofisch overzicht op het terrein van onderwijsonderzoek Bredo, 2006 en voor evaluatie Mertens & Wilson, 2012, p. 33 e.v.). Ernest House (1978; 1980 en 1983) die belangrijke bijdragen heeft geleverd aan theoretische noties omtrent onderwijsevaluatie, groepeerde evaluatiemodellen in twee groepen: *objectivisten* en *subjectivisten*⁴. Objectivisten gaan ervan uit dat een evaluatie wordt uitgevoerd met methoden en technieken die resultaten opleveren die onder dezelfde omstandigheden door anderen herhaald kunnen worden en tot dezelfde resultaten leiden. Deze manier van denken komt voort uit de wetenschapsfilosofie van het empirisme (zie bijv. Koningsveld, 2006). Subjectivisten daarentegen vinden dat de kwaliteit van een evaluatie niet

⁴ Dit is een grove indeling in wetenschapsfilosofische paradigma's. Een meer gedetailleerde indeling is te vinden bij Mertens en Wilson (2012, p. 35) die een postpositivistisch, constructivistisch, transformatief en pragmatisch paradigma onderscheiden.

bepaald wordt door de ‘wetenschappelijkheid’ van de methoden en technieken, maar door de expertise van de evaluator en dat reproduceerbaarheid geen kritisch kenmerk is.

House (1972, 1983) maakt ook nog een ander onderscheid dat meer samenhangt met beoordelen en waarderen. Hij spreekt in dat verband van utilitarian- en intuitionist-pluralist-evaluatiemodellen. Bij utilitarian-modellen wordt de waarde van het object van evaluatie bepaald door de algehele impact die het object heeft op de doelgroep, terwijl het andere model uitgaat van de impact op de leden van de doelgroep of van een subgroep. Het eerste model hangt samen met de objectivistische manier van denken en het tweede past beter bij het subjectivisme.

Natuurlijk is het zo dat de door House onderscheiden categorieën feitelijk dimensies zijn en dat in de evaluatiepraktijk ook ‘mixed-methods’ voorkomen (zie bijv. Alkin, 2013). Dat neemt niet weg dat beide manieren van wetenschapsfilosofisch denken hebben geleid tot een langdurige fundamentele discussie, nl de discussie ‘kwantitatieve versus kwalitatieve evaluatie’.

Bij de ontwikkeling van evaluatiemodellen hebben de theoretici zich niet alleen laten leiden door wetenschapsfilosofische opvattingen of hun kijk op wat kennis is en hoe je de ‘ze kunt verwerven. Men werd ook geconfronteerd met nieuwe behoefte aan evaluatieve kennis. Denk daarbij aan vragen die in jaren tachtig opkwamen zoals hoe schoolboeken beoordeeld moesten worden, of een bepaald curriculum wel zijn vruchten afwierp, hoe goed leerlingen feitelijk presteren, of scholen wel aan de maat waren, of onderwijs zijn geld waard is, etc. Ook deze beleids- en praktijkvragen hebben bijgedragen aan de verdere ontwikkeling van evaluatiemodellen.

3 CATEGORISERING VAN EVALUATIETHEORIEËN

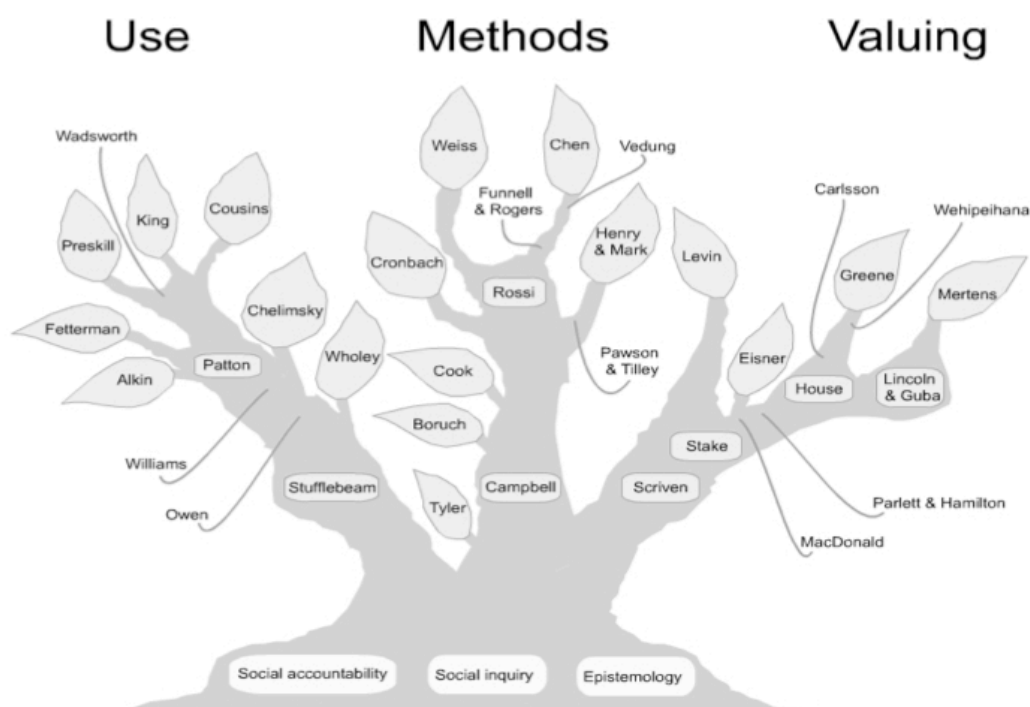
3.1 Evaluatieboom

Alkin en Christie (Alkin & Christie, 2004; Christie & Alkin, 2008 en Christie & Alkin, 2013) hebben gedurende 10 jaar op verschillende plaatsen een poging gewaagd de benaderingen van verschillende invloedrijke theoretici te classificeren en hebben dit gevisualiseerd in de vorm van een boomstam met drie hoofdtakken. De stam van deze boom heeft drie wortels (zie figuur 1): *Social Accountability* (= Rekenschap en Toezicht), *Social Inquiry* (= sociaal-wetenschappelijk onderzoek van het gedrag van groepen en individuen in verschillende sociale situaties) en *Epistemology* (= onderdeel van de filosofie dat zich richt op de aard en de validiteit van kennis). Gedragswetenschappen zijn voor evaluaties van belang omdat ze theoretische kennis opleveren over de principes die ten grondslag liggen aan menselijk gedrag. Deze theorieën zijn op hun beurt van belang om bijvoorbeeld gewenste en ongewenste effecten te evalueren die worden opgeroepen door programma's die bedoeld zijn om menselijk gedrag te beïnvloeden, zoals bijvoorbeeld een bepaalde didactische aanpak, interventie- of preventieprogramma's.

Volgens Alkin en Christie vormen deze drie domeinen de basis van alle ontwikkelingen op het terrein van onderwijsevaluatie. Uit de aard van het werk van toezichthouders volgt enerzijds de behoefte na te gaan of het object van toezicht in overeenstemming handelt met wetgeving, regels, voorschriften en verwachtingen. Anderzijds moeten toezichthouders nagaan of dat ook daadwerkelijk in voldoende mate het geval is. Daarvoor zijn criteria, evaluatieve gegevens en beoordelingen nodig. Het werk van vroege toezichthouders en accountants is te beschouwen als systematische evaluatie 'avant la lettre'.

Het is met name het domein van de *Social Inquiry* waarbinnen methoden en technieken zijn ontwikkeld om ook daadwerkelijk gegevens te kunnen verzamelen om verantwoording over te kunnen afleggen. Hieruit zijn volgens de auteurs verschillende evaluatiebenaderingen ontstaan. Historisch gezien zijn bijdragen van wetenschappers als Montesquieu, Rousseau, Marx, Durkheim en Weber die hebben geleid tot het verzamelen van gegevens en het toepassen van statistiek. Psychologen introduceerden experimenteel onderzoek en antropologen ontwikkelden etnografisch onderzoek waaruit zich later narrative inquiry

ontwikkelde. Deze ontwikkelingen, die in hoge mate parallel lopen met het eerder gemaakte onderscheid tussen objectivisme en subjectivisme, leidde tot fundamentele discussies over het onderscheid tussen enerzijds voorspellen en verklaren en anderzijds interpreteren en begrijpen⁵. Deze ontwikkelingen kregen ook hun weerslag in de verschillende evaluatietheorieën. Zo zijn invloeden vanuit de narratief georiënteerde antropologie te vinden in het werk van Robert Stake en Elliot Eisner.



Figuur 1: Evaluatietheorie-boom (bron: Alkin, 2013).

Epistemologie speelt een belangrijke rol in het denken over evaluatie omdat hieraan de argumenten worden ontleend voor de legitimiteit van waardeoordelen en van kennis van de waarheid of van wat ‘feiten’ zijn. Het spreekt voor zich dat dit ook consequenties heeft voor de wijze waarop evaluaties worden uitgevoerd. Discussies rond waarden en de aard van kennis vinden hun grondslag in grofweg drie manieren van denken over kennis: 1) postpositivisme, 2) constructivisme en 3) pragmatisme. Vanuit de postpositivistische hoek wordt beaamd dat het weliswaar in onderzoek en evaluatie gaat om het streven naar het begrijpen van de waarheid, maar dat er altijd een mate van onzekerheid blijft bestaan bijvoorbeeld vanwege meet- en observatie-onnauwkeurigheden. Deze manier van denken is terug te vinden bij

⁵ Clifford Geertz’s klassieke essay “Thick Description: Toward an Interpretive Theory of Culture” in *The Interpretation of Cultures* (1973) is een exponent van de interpretatieve of narratieve stroming waarbij het niet gaat om voorspellen maar om het begrijpen van gedrag.

verschillende wetenschappers die op de ‘methode-tak’ van de evaluatieboom zijn geplaatst, zoals Campbell en zijn navolgers.

Constructivisten daarentegen gaan er vanuit dat er niet één maar verschillende werkelijkheden zijn, afhankelijk van het subjectieve waardesysteem van zowel de evaluator als van degenen die worden geëvalueerd. Representanten van deze manier van denken vinden we op de ‘waarde-tak’ van de evaluatieboom, zoals Stake, Guba en Lincoln.

Pragmatisten zien objectiviteit en subjectiviteit als twee posities op hetzelfde continuüm en zijn daarom van mening van deductieve en inductieve logica hand in hand gaan. Aan de ene kant leunen pragmatisten tegen postpositivisten aan als het gaat om de aanname dat er geen absolute waarheid is. Aan de andere kant vinden pragmatisten, net zoals constructivisten, dat er meerdere verklaringen van de werkelijkheid zijn en dat op een bepaald moment de ene verklaring meer ‘waar’ is dan een andere. Het pragmatisme speelt een rol bij theoretici die zich vooral bezig houden met het gebruik van evaluatie-uitkomsten, zoals bijvoorbeeld Patton.

3.2 Vertakkingen

De boom heeft vervolgens drie vertakkingen waarvan de middelste betrekking heeft op de *methode-component* c.q. de kennisproductie van evaluatie. De theoretici die op deze tak zijn geplaatst staan bekend om het feit dat hun evaluatie-aanpakken primair worden geleid door methodologische principes om kennis te produceren op - rekening houdend met de omstandigheden van de evaluatie- de meeste robuuste manier (Shadish, Cook & Leviton, 1991). De grondlegger van deze manier van denken is Donald Campbell (Campbell & Stanley, 1966), die zich richtte op de condities waaronder experimenten en quasi-experimenten dienen te worden uitgevoerd.

De tweede hoofdtak heeft betrekking op de *waardecomponent* van evaluatie waarin het werk van Michael Scriven (1980) een centrale plaats inneemt. Daarbij gaat het om manier waarop evaluatiegegevens van een waarde-oordeel worden voorzien. Volgens Shadish et al. (1991, p. 94) is Scriven de eerste en belangrijkste theoreticus die een expliciete theorie heeft over waarde-oordelen. Van hem is ook de uitspraak ‘*Bad is bad and good is good and it is the job of evaluators to decide which is which*’ (Scriven, 1983, p. 19). De wetenschappers die op zijn schouders zijn gaan staan, waaronder Robert Stake (1975), onderstrepen dat het meest typische kenmerk van evalueren is het toekennen van waarden aan gegevens (oordeelsvorming). Deze tak splitst zich in twee richtingen: objectivisten en subjectivisten.

De objectivisten hebben zich het meest laten beïnvloeden door Scriven die van mening is dat de evaluator degene is die het waarde-oordeel moet vellen. De subjectivistische tak daarentegen is van mening dat de realiteit een doorlopend dynamisch proces is en dat de ‘waarheid’ altijd relatief is ten opzichte van een bepaalde referentiekader. Waarde-oordelen worden dus geveld binnen de subjectieve betekenis van de evaluatiegegevens. In tegenstelling tot de objectivisten worden oordelen dan ook niet uitsluitend geveld door de evaluator.

De derde hoofdtak wordt gevormd door de *gebruikscomponent* van evaluatie op basis van het pionierswerk van Daniel Stufflebeam die zich vooral richtte op het beslissingsproces naar aanleiding van de gegevensverzameling en de oordeelsvorming en dus op de gebruikers en de toepassingen van evaluatie. De vroege theoretici van deze tak waren van mening dat de opdrachtgever tot evaluatie de enige was die beslissingen nam. Latere theoretici hebben ook aandacht gevraagd voor andere betrokkenen waarvoor de evaluatie interessant was en ook voor het beter toerusten van deze doelgroepen om het doel, het proces en de uitkomsten van evaluaties beter te kunnen begrijpen. Dat heeft zelfs geleid tot evaluatiemodellen die vooral werden toegepast om doelgroepen te emanciperen, ook wel ‘Social Justice Evaluation’ genoemd (Mertens & Wilson, 2012, p. 161 e.v.).

De drie takken moeten niet gezien worden als onafhankelijke stromingen, maar de theoretici op de verschillende takken staan ook in relatie tot elkaar. Als de boom in 3D-weergave wordt geprojecteerd dan zou te zien zijn dat meest rechtste vertakking van de waarden-tak dicht tegen de meest linkse vertakking van de Gebruikstak aan liggen (en deze zelfs een beetje raakt)⁶.

3.3 Invloed van toezicht op evaluatietheorieën

Dat evaluatietheorieën invloed hebben gehad op de ontwikkeling van het onderwijstoezicht zoals we dat nu kennen, zal niemand verbazen, mogelijk wel dat het werk van toezichthouders uit het begin van de vorige eeuw het fundament vormt van de ontwikkeling van diezelfde evaluatietheorieën. Interessant is het feit dat Alkin en Christie (2004, 2008, 2013) Rekenschap en Toezicht (Social Accountability) zien als bron van de ontwikkeling van evaluatietheorieën. Historisch gezien hebben volgens de auteurs Europese

⁶ Sommige critici van Christie en Alkin vinden dat de boom-metafoor beter vervangen had kunnen worden door een delta-rivier met verschillende vertakkingen met zoet, zout en brak water waardoor een beter zicht verkregen kan worden op de inter-relaties tussen theoretici en hun evaluatiemodellen. Ook is de indeling van Christie & Alkin gekritiseerd vanwege het ontbreken van niet-westerse evaluatietheoretici en niet-academische instellingen die evaluaties uitvoeren zoals de Worldbank, het Brookings Institute en de UNESCO (Mertens & Wilson, 2012, p. 41). Ook het ontbreken van evaluatie-aanpakken van met name Europese inspecties is een ommissie.

onderwijsinspecties hieraan in het begin van de 20ste eeuw een belangrijke bijdrage geleverd (Alkin & Christie, 2013, p. 14):

(.....), the European tradition of school inspectors (.....) has been depicted in countless movies showing dreaded visits of the school inspector, whose presence is designed to ensure that teachers are engaged in performing prescribed lessons at prescribed times. Perhaps a contemporary evaluation procedure designed for obtaining process accountability is the system of school accreditation employed in secondary schools, private schools, universities, and professions. In these instances, evaluation procedures are designed whereby teams with presumed expertise, frequently guided by established process standards, visit a site to observe, account, and make a report. The result is a judgment about whether institutions are accountable and should be accredited.

Opmerkelijk in dit verband is ook dat in de eerste versie van The International Encyclopedia of Educational Evaluation (Walberg & Haertel, 1990) een compleet hoofdstuk aan onderwijstoezicht is gewijd. Dit hoofdstuk beschrijft de werkwijze van onderwijsinspecties van enkele Europese- en van de Britse Gemeenebest-landen in de periode 1970 -1990 (Bolam, 1990). Dat aan onderwijstoezicht in zo'n standaardwerk een hoofdstuk is gewijd, laat dan ook duidelijk zien dat inspectiewerk aan het al einde van de vorige eeuw beschouwd werd als een serieuze evaluatie-aangelegenheid.

Vele auteurs erkennen dat rekenschap en toezicht een belangrijke rol hebben gespeeld in de ontwikkeling van evaluatiemodellen. Mark, Henry en Julnes (2000) zien toezicht en rekenschap zelfs als een van de vier hoofdfuncties van evaluatie omdat ze een belangrijke rol spelen in de politiek en de beleidsontwikkeling.

Dat ook Rekenschap tot de wortels van evaluatietheorieën wordt gerekend is niet verwonderlijk omdat zogenaamde 'financial accounting' theorieën gaan over de vraag of bepaalde transacties of gebeurtenissen al dan niet moeten worden verwerkt in de jaarrekening, en zo ja op welke wijze, welke waarderingsgrondslagen moeten worden gehanteerd, hoe omgegaan moet worden met eventuele waardemutaties, welke additionele gegevens verstrekt moeten worden in de toelichting, op welke wijze deze informatie gepresenteerd moet worden, en welke informatie verder nog verstrekt moet worden, bijvoorbeeld in een operationele en financiële analyse van de financiële positie en de resultaten in het bestuursverslag. Kortom: het gaat over de wijze waarop verwerking, waardering en resultaatbepaling, presentatie en gegevensverstrekking zou moeten plaatsvinden (Knoops, 2010). Theoretici houden zich onder andere bezig met het ontwikkelen van nieuwe verslaggevingsmodellen, nieuwe meetmethoden voor niet-financiële

prestatie-indicatoren en er worden voorstellen gedaan voor de verwerking en presentatie van immateriële activa. De overeenkomst tussen onderwijsinspecteurs en accountants is treffend. Zij verklaren in de accountantsverklaring dat de jaarrekening een getrouw beeld geeft van de omvang en samenstelling van het vermogen en resultaat van een onderneming. Aangezien er (enige) flexibiliteit is in verslaggevingskeuzen zijn er in ieder geval verschillende getrouwe beelden mogelijk (Knoops, 2010).

4 METHODOLOGISCHE PARADOXEN

4.1 Ontwikkelingen in het onderwijstoezicht

Volgens Mertens (2011, p. 133-134) is bij inspectie-organisaties sprake van een toenemende institutionalisering. Het toezicht door inspecties was van oorsprong toezicht door inspecteurs en dus een individuele ambtsuitoefening. Dat geldt ook voor de onderwijsinspectie. Tot aan het begin van de jaren tachtig van de vorige eeuw bepaalde de inspecteur in hoge mate zelf zijn eigen agenda en besliste wat op welke wijze aan de orde moest komen op de scholen uit zijn ambtsgebied (zie Elte, 1988; Janssens, 1997 en Dodde, 2001). Tot 1980 werd er door de inspectie bijzonder weinig op landelijk niveau gerapporteerd. Aan de schoolbezoekverslagen en aan het Onderwijsverslag lagen toen geen gestructureerde aanpak voor systematische gegevensverzameling ten grondslag. De rapporten kwamen tot stand op basis van informatie die inspecteurs over het onderwerp paraat hadden. Deze werkwijze was volgens Janssens (1997) idiosyncratisch-impressionistisch van aard en past daarom binnen de traditie van het subjectivisme.

Allengs werden werkwijzen ontwikkeld voor een systematische planning van onderwerpen die voor onderzoek en wetshandhaving op scholen in aanmerking komen en voor een gestandaardiseerde wijze waarop gegevens over de kwaliteit van het onderwijs verzameld, gewaardeerd en gerapporteerd moeten worden (Janssens, 1997; Elte & Scholtes, 2001; Smeets & Verkroost, 2011). Deze ontwikkelingen kwamen mede tot stand onder invloed van de wens om expliciet te oordelen over de kwaliteit van scholen, de scholen daarover te informeren, deze oordelen openbaar te maken en transparant te zijn over de werkwijze van de inspectie en de criteria en normen waarlangs de oordeelsvorming plaatsvindt. Deze manier van werken is beduidend minder impressionistisch van aard en past in de traditie van het objectivisme.

De ontwikkeling die de inspectie heeft door gemaakt naar een meer gestandaardiseerde en transparante werkwijze op het terrein van gegevensverzameling en oordeelsvorming heeft wel geleid, en leidt nog steeds, tot discussies over aard en kwaliteit ervan. Daarin spelen twee paradoxen een rol.

4.2 Inspectieparadoxen

Paradox 1

De gegevensverzameling is in hoge mate een individuele activiteit is van een inspecteur die (steekproefsgewijs) naar bepaalde onderdelen van de onderwijspraktijk in of van een school kijkt, zich vervolgens een oordeel vormt over de kwaliteit van het onderwijs van die school, terwijl tegelijkertijd rechtsgelijkheid, fairness, betrouwbaarheid en validiteit wordt nagestreefd.

Het oordeel van de inspectie over (onderdelen van) de naleving van de wet- en regelgeving en de kwaliteit van het onderwijs, dient onafhankelijk van de persoon van de inspecteur plaats te vinden.

Paradox 2

Naarmate de inspectie haar werkwijze standaardiseert raakt het zicht op de uniciteit van scholen verloren.

Het risico van de huidige werkwijze is dat individuele scholen worden beoordeeld binnen een rigide systeem van genormeerde waarderingskaders en gestandaardiseerde instrumenten.

Omdat dergelijke normen, kaders en instrumenten zijn gebaseerd op statistische regelmatigheden, leidt een rigide toepassing voorspelbaar tot onjuiste beoordelingen in veel individuele gevallen. Hoe zwakker de statistische regelmatigheden, des te meer individuele gevallen onjuist zullen worden beoordeeld. Het streven naar betrouwbare en geldige evaluatie-instrumenten kan dus gevolgen hebben voor ongeldige conclusies in de evaluatie.

4.3 Achterliggende redenering

‘De inspecteur: enquêteur of connaisseur’ zo luidde de titel van een bijdrage van Wijnand Hoebe in een bundel over de ontwikkeling van de methodologie van de inspectie (Janssens, 1997). Deze titel is ook representatief voor discussies die de inspectie in de jaren negentig intern en extern hield over taak en werkwijze en die de kern raken van de twee paradoxen. Deze kritiek kende de volgende achterliggende gedachtengang.

Niet zelden kwam het voor dat in de jaren negentig de inspectie voor de voeten geworpen kreeg dat inspectierapporten over de toestand van het onderwijs – net als rapporten van visitatiecommissies – minder zijn gebaseerd op een systematische verzameling van vergelijkbare gegevens en meer op – eventueel intuïtieve – expertbeoordelingen (zie bijv.

Hoeben, 1993, 1995). Deze beoordeling van de inspectiemethodologie vraagt eerder vertrouwen in de deskundigheid van inspecteurs in plaats van controleerbaarheid van hun conclusies. Inspecteren zou dan ook uitstekend passen in het door Eisner (1977) geïntroduceerde evaluatiemodel van ‘educational connoisseurship’ (zie ook Barone, 1991). Dat model kapitaliseert niet op een controleerbare systematische verzameling van gegevens, maar op evaluatoren die als fijnproevers of kunstkenners optreden en hun oordeel beargumenteren. Eisner (1979, p. 14) ziet evalueren dan ook als ‘the art of appreciation’ en vindt dat evaluatoren tot taak hebben ‘*to enable others to experience what they may have missed*’ (ibid, p. 15). Ook Mertens (2011, p. 150) wijst nog eens in zijn boek ‘Inspecteren, toezicht door inspecties’ op de mogelijkheden van deze aanpak in bepaalde situaties.

Nog steeds is de kritiek niet van de lucht dat het onderwijstoezicht zich teveel uitkristalliseert in de richting van een enquêteurs- of accountantsrol van de inspecteurs en te weinig gebruik maakt van hun potentieel van kritische kenners van het onderwijs (zie bijv. Boonstra, 2002; Obbink, 2011; Beukenholdt & Dietvorst, 2011). Aan de ene kant leidt een gestandaardiseerde inspectiemethodologie tot grotere vergelijkbaarheid van oordelen tussen inspecteurs en dat is zonder meer een winstpunt. Maar aan de andere kant maakt een grotere vergelijkbaarheid tussen inspecteurs hun werk echter niet per definitie empirisch-analytisch van aard (Hoeben, 1997). Integendeel: in de kern blijft het onderzoek van scholen door inspecteurs toch vooral kwalitatief van aard. De inspecteur beschrijft en beoordeelt in beginsel de kwaliteit van afzonderlijke gevallen. Het oordeel van de inspecteur over de kwaliteit van het onderwijs op een school is eerst en vooral gebaseerd op haar/zijn expertise en op ‘connoisseurship’ van onderwijs, zo luidt de redenering.

Als schoolonderzoeken niet meer leiden tot evaluaties van afzonderlijke scholen waarin relevante gegevens ook worden verzameld als zij niet in de structuur van vooraf gemaakte instrumenten passen en waarin kwaliteitsoordelen worden beargumenteerd vanuit – soms unieke, maar vaak uitzonderlijke – contexten, dreigt de balans in het werk van de inspectie door te slaan ten nadele van kwaliteitsimpulsen naar afzonderlijke scholen en ten gunste van een verantwoording van macro-evaluaties, zo luidt de redenering. Kritisch beargumenteerde en afgewogen oordelen over afzonderlijke scholen dreigen het af te leggen tegen een sociaal-wetenschappelijk geïnspireerde methodologische verantwoording vanuit een empirisch-analytisch, maar bovenal kwantitatief kader (Hoeben, 1997).

Het grootste gevaar van deze ontwikkeling zit niet zozeer in de kwantitatieve methodologie op zichzelf, maar in het risico dat het zicht op de individualiteit van scholen verloren raakt. Het grootste gevaar dat de huidige werkwijze in zich bergt, is dat individuele gevallen worden beoordeeld binnen een rigide systeem van genormeerde waarderingskaders en ontwikkelde

instrumenten; omdat dergelijke normen, kaders en instrumenten zijn gebaseerd op statistische regelmatigheden, leidt een rigide toepassing voorspelbaar tot onjuiste beoordelingen in veel individuele gevallen. Hoe zwakker de statistische regelmatigheden, des te meer individuele gevallen onjuist zullen worden beoordeeld. Het streven naar betrouwbare en geldige evaluatie-instrumenten kan dus gevolgen hebben voor ongeldige conclusies in de evaluatie.

Als we enkele inspectierapporten uit het begin van de jaren negentig (Inspectie van het Onderwijs, 1991; 1992) in ogenschouw nemen, dan blijkt dat de basiswerkwijze en -argumentatie berust op de ervaring van de inspecteurs en dat de inspecteurs wordt gevraagd hun 'educational connoisseurship' toe te passen op gedetailleerde en van tevoren gespecificeerde aspecten van het onderwijs. Hoeveel begrippen van tevoren ook worden gespecificeerd en in onderdelen nader worden toegelicht, hoeveel de gegevensverzameling ook wordt gesystematiseerd en op een rij gezet, inspecteurs hebben grote vrijheid in het selecteren van gegevens die zij relevant vinden, en de operationalisering in termen van variabelen blijft formeel: 'het oordeel van de inspecteur over ...'. We hebben met andere woorden te maken met beoordelingen door experts. Hun oordeel is enerzijds evenveel waard als hun expertise waard is. Anderzijds zijn zij in staat om beargumenteerd af te wijken van de vooraf vastgestelde standaarden; zij behoren dit ook te doen als zij daarvoor aanleiding vinden.

'Educational connoisseurship' en de er bij behorende presentatie in de vorm van 'educational criticism' zijn termen die door Eisner (1977) zijn geïntroduceerd. Connoisseur is een Engelse term met een betekenis die overeenkomt met de pregnante betekenis van het Franse *connaissanceur*, namelijk: (kunst)kenner; fijnproever. In de discussie over evaluatiemethodologie gedurende de jaren zeventig zijn de termen door Eisner gebruikt als benchmarking van een kwalitatieve evaluatiemethodologie. Een connoisseur is een (kunst)kenner of een fijnproever die op basis van persoonlijke ervaringen een oordeel uitspreekt over een kunstwerk, wijn of gerechten, of - in het onderwijs - over de moeilijkheid van examens, de kwaliteit van een school, de doeltreffendheid van een curriculum of de kwaliteit van het beleid. Het oordeel wordt gepresenteerd in de vorm van een kritisch essay waarin het oordeel tevens wordt beargumenteerd, zoals in literatuur- of filmkritiek. Het gaat dus om fijnproevers van het onderwijs, die zich concentreren op waarde-oordelen en argumenten. Zij baseren deze argumenten niet per se op systematisch en controleerbaar verzamelde en geanalyseerde gegevens - ook al hebben ze deze gegevens tot hun beschikking - maar op hun inzicht in individuele gevallen (Mertens, 2011, p. 151). De controleerbaarheid van hun waarde-oordelen is gebaseerd op de argumenten die zij gebruiken.

In de huidige inspectiewerkwijze, zo luidt de redenering, wordt teveel gekapitaliseerd op expliciete, regelgeleide, systematische en controleerbare gegevensverzameling, waarbij inspecteurs moeten doen wat in de afgesproken regels staat en waarbij zij de normen moeten toepassen, zelfs als het onzin zou zijn. De rol van de inspecteurs helt teveel over naar de rol van enquêteur of accountant. Het schoolonderzoek stelt kwalitatieve eisen die behoren bij gevalsstudies, formatieve evaluatie, beoordeling van individuele gevallen, feeling voor uitzonderingen op algemene wetmatigheden, en dergelijke. Evaluatie van macro-aspecten ten behoeve van bijvoorbeeld het Onderwijsverslag, stelt eisen van vergelijkbaarheid, betrouwbaarheid en geldigheid van instrumenten, althans van waarnemingen, generaliseerbaarheid, kwantitatieve statistische analyses, en dergelijke. Als de nadruk op macro-evaluaties tot gevolg zou hebben dat schoolonderzoeken niet meer voldoen aan de genoemde kwalitatieve eisen, dan gaat dit ten koste van de rapportage aan de scholen en dus aan hun kwaliteitsontwikkeling. Als de nadruk op schoolonderzoeken tot gevolg zou hebben dat macro-evaluaties niet meer tot generaliseerbare resultaten leiden, dan gaat dit ten koste van de rapportages aan het onderwijsbeleid en aan de kwaliteit van dit beleid.

Als een evenwicht tussen meso- en macroniveau niet gevonden kan worden, dan verdient het aanbeveling de eisen aan het schoolonderzoek zwaarder te laten wegen. Voor het schoolonderzoek van de inspecteurs is immers geen alternatief. Voor macro-evaluaties kan de inspectie terugvallen op opdrachten aan onderzoeksinstituten, die dergelijk onderzoek met plezier zullen uitvoeren. De positie van inspecteurs in onderwijsevaluatie wordt diensgevolge primair bepaald door de eisen die het schoolonderzoek aan hen stelt.

5 DE EVALUATIEMETHODOLOGIE VAN DE INSPECTIE⁷

5.1 Rationale

5.1.1 Uitgangspunten

De wettelijke opdracht van de inspectie behelst, kort gezegd, toezicht houden op en informatieverschaffing over de kwaliteit van het onderwijs. Het perspectief van waaruit de inspectie dient te evalueren is bepalend voor de definitie van het begrip ‘kwaliteit van het onderwijs’ en voor de evaluatiecriteria die zij hanteert, of anders gezegd, voor de eisen die zij aan het onderwijs stelt. Maar namens wie of op wiens gezag stelt de inspectie deze eisen?

Gezien taak en positie van de inspectie ligt het voor de hand dat haar kwaliteitseisen niet dezelfde behoeven te zijn als die van de scholen, de leerlingen, de ouders, de vakbonden, etc.. Aangezien de inspectie een overheidsorgaan is dat namens de minister belast is met wetshandhaving en informatieverschaffing over de kwaliteit van het onderwijs aan de bewindslieden en aan de Staten-Generaal, vertegenwoordigt de inspectie ook het Algemeen Belang (zie Groenendijk, 1981)⁸. Om die reden betreft de inspectie naast het perspectief van de wetgever, neergeslagen in de wet- en regelgeving en het daaraan gekoppelde onderwijsbeleid, nog twee andere in haar evaluatie:

1. *Het perspectief van de wetenschap en van deskundigen.*

Dit perspectief is vastgelegd in wetenschappelijke publicaties en in vakpublicaties. Vanuit dit perspectief gaat het om criteria die vooral gericht zijn op de kwaliteit van de (vak)didactische aanpak en organisatie van het onderwijs.

⁷ Bewerking van Janssens, F.J.G. (1997). *Zo zijn onze manieren: een twintigjarige zoektocht naar een inspectiemethodologie voor onderwijsevaluatie*. In: Janssens, F.J.G. (red). *Van onderzoek naar evaluatie: de methodologie van de onderwijsinspectie*. Den Haag: Delwel Uitgevers.

⁸ Het algemeen belang (*Public Interest*) is een juridische term en duidt op datgene dat voor het welzijn van het volk in het algemeen nuttig, gewenst of nodig is. De term is - hoewel niet eenvoudig te definiëren - toch in de maatschappelijke en politieke discussie binnen democratieën onmisbaar. Ze duidt erop wat het volk - als het bijvoorbeeld over allerlei zaken zou stemmen - kennelijk wil.

2. *Het perspectief van de direct betrokkenen (stakeholders) bij het onderwijs.*

De opvattingen van de direct betrokkenen (leraren, directieleden, leerlingen/studenten, ouders, het georganiseerd onderwijs/overleg) over goed onderwijs spelen evenzeer een rol in de bepaling van evaluatiecriteria. Deze opvattingen worden verzameld via raadplegingen.

Naast het evaluatieperspectief speelden ook andere principiële uitgangspunten een rol bij de evaluatie-aanpak:

1. Uit het oogpunt van rechtsgelijkheid en zorgvuldigheid ten aanzien van het scholenveld dient de evaluatie-aanpak systematisch en zoveel mogelijk gestandaardiseerd te zijn;
2. Het instrumentarium moet betrouwbare en valide gegevens en billijke en rechtvaardige oordelen over de kwaliteit van het onderwijs (aan individuele scholen) opleveren.

Deze uitgangspunten lijken met elkaar in strijd. Systematisering en standaardisering vragen, vanuit methodologisch oogpunt, in principe om een sterk voorgestructureerd en geprecodeerd instrumentarium (denk daarbij aan de vragenlijsten en interview/observatieschema's) opdat (onbedoelde en ongewenste) ruis (vanwege de persoon die gegevens verzamelt) wordt uitgesloten. Daarentegen worden de expertise en de rol van de inspecteur optimaal benut wanneer hij of zij juist eigen kennis en ervaringen, maar ook de specifieke omstandigheden waarin de individuele school verkeert, in de beoordeling kan betrekken, zoals het ontwikkelingsstadium en de verandercapaciteit van de school. Daarmee komt de onderlinge vergelijkbaarheid van bevindingen en oordelen van inspecteurs, en daarmee de kwaliteit van landelijke rapportages, in het geding.

De werkwijze moet wetenschappelijk te verantwoorden zijn, niet exclusief gericht zijn op beschrijven of inventariseren, maar vooral op het beoordelen van de onderwijspraktijk en moet kunnen worden geïntegreerd in het alledaagse schoolonderzoek.

5.1.2 Conceptueel kader

Wetenschappelijk onderzoek en evaluatie hebben veel gemeenschappelijk, met name de systematische gegevensverzameling, maar verschillen op één belangrijk punt van elkaar: onderzoek is vooral gericht op uitbreiding van generaliseerbare kennis met het oog op theorievorming. Bij evaluatie gaat het om een oordeel over de waarde van 'iets'. Daarvoor zijn enerzijds criteria nodig, waarlangs kan worden vastgesteld of het object van evaluatie is gerealiseerd. Anderzijds zijn er normen nodig om vast te kunnen stellen of en in welke mate

aan die criteria wordt voldaan. Criteria en normen zijn een eerste noodzakelijke voorwaarde voor een goede evaluatie (De Groot, 1986; Scheerens, 1983; Worthen & Sanders, 1987).

Voor het typeren van de criteria van en kwalificaties voor te evalueren objecten zijn in wetenschappelijke publicaties verschillende begrippen in omloop: criteria, standaarden, maatstaven en normen (zie bijv. Van den Berg, Hoeben & Van der Werf, 1993; Janssens & Zwarts, 1993). Het onderscheid tussen criteria en kwalificaties wordt vaak impliciet gelaten. Zo is een zogenaamde ISO-NEN-norm een ‘criterium inclusief kwalificatie’. Bovendien worden de bovengenoemde termen vaak door elkaar gebruikt. Dezelfde termen worden gebruikt om zowel de lat waarmee gemeten wordt als de onderscheidingen die op die lat worden gemaakt, aan te duiden. Omdat het gebruik van de bedoelde termen, noch taalkundig noch wetenschappelijk gezien, eenduidig is, heeft de inspectie er destijds twee gekozen: ‘Standaard⁹’ om de kritische kwaliteitskenmerken van het object van evaluatie te definiëren en ‘norm’ om de mate waarin de kwaliteitscriteria in de onderwijspraktijk worden gerealiseerd van een kwalificatie te voorzien (Janssens & Zwarts, 1993).

De inspectie heeft een evaluatie-aanpak ontwikkeld waarin op een effectieve wijze bovengenoemde uitgangspunten zijn gecombineerd. Deze aanpak laat zich als volgt typeren:

- systematisering en standaardisering worden bereikt door per evaluatie een gemeenschappelijk instrumentarium te ontwikkelen.
- de ontwikkeling van deze instrumentaria gebeurt door inspecteurs die deskundig zijn op het onderwerp dat wordt geëvalueerd. Zij dragen ook zorg voor de scholing van alle collega’s die met het instrumentarium moeten werken.
- de instrumenten waarmee de kwaliteit van het onderwijs wordt beoordeeld doen in hoge mate een beroep op de materiedeskundigheid van inspecteurs.

⁹ Later vervangen door de neutralere term ‘kwaliteitsaspect’.

5.2 Instrumentarium

5.2.1 Waarderingskaders

Voor de beoordeling van de toestand van het onderwijs maakt de Inspectie gebruik van zgn. waarderingskaders¹⁰ (zie Janssens, 1997). Deze betreffen (aspecten van) de inhoud en/of de organisatie van het onderwijs en bestaan uit een set standaarden, indicatoren en normen.

De standaarden hebben betrekking op de belangrijkste kwaliteitsaspecten van het onderwerp dat wordt geëvalueerd. Normen geven de beslissingsregels en de kwalificaties voor de mate waarin - naar de mening van de inspectie - de standaard in het onderwijs gerealiseerd dient te zijn. Omdat het zelden mogelijk is direct de realisering van een standaard vast te stellen, worden standaarden door middel van het definiëren van de essentiële kenmerken ervan, binnen de inspectie 'indicatoren' genoemd, nader geoperationaliseerd.

Er zitten - in ieder geval bij de inspectie - materieel en principieel geen methodologische verschillen tussen een standaard en een indicator¹¹. In beide gevallen gaat het om een observeerbaar fenomeen dat een signalerende functie heeft voor aspecten van de kwaliteit van een instituut of een systeem (Oakes, 1986; Goldstein & Spiegelhalter, 1996). Zowel een standaard als een indicator zijn een factor die betrekking heeft op de kwaliteit van een aspect van het onderwijs, zij het dat een standaard van een hogere orde is dan een indicator.

Standaarden zijn kenmerken van een - conceptueel gezien - hoger domein, bijvoorbeeld het pedagogisch-didactisch handelen, terwijl indicatoren kenmerken zijn van de afzonderlijke standaarden.

¹⁰ Ofschoon de begrippen waarderingskader en toezichtskader in de praktijk door elkaar worden gebruikt zijn het toch twee verschillende zaken. Een toezichtskader beschrijft de wijze waarop de Inspectie van het Onderwijs haar toezicht vorm en inhoud geeft. Het is dus een set werkwijzen, procedures en instrumenten waarmee het toezicht wordt uitgevoerd. Het waarderingskader vormt een onderdeel van het toezichtskader. Een waarderingskader bevat de criteria en de normen om de kenmerken van aspecten van het object van toezicht te beschrijven, te beoordelen en van een kwalificatie te voorzien. Echter de wijze waarop en de instrumenten waarmee de gegevens worden verzameld om te kunnen beschrijven en te oordelen, zijn onderdeel van het toezichtskader.

¹¹ Opgemerkt dient te worden dat anderen wel een principieel onderscheid maken tussen een standaard en een indicator. Zie bijvoorbeeld de Comon Core Standards die te vergelijken zijn met eindtermen (<http://www.corestandards.org>) In de Angelsaksische literatuur is een standaard vaak een norm (standaarden voor goede onderwijsprestaties, voor goed onderwijs, voor goede evaluaties), terwijl bij de inspectie de mate waarin in de praktijk indicatoren die onder een standaard ressorteren zijn gerealiseerd, de norm voor een standaard zijn.

Indicators may serve three basic functions. Their first function is the measurement function, in order to undertake analysis, assessment, monitoring or evaluation of the quality of individual schools and education in general. The second purpose of indicators is communication. Indicators help to inform relevant stakeholders or society about the state or development of a particular educational system. Finally, indicators can be used for normative or standardization purposes. Following this function, indicators show to what extent the actual situation deviates from an established bench, predefined standards or a set of goals. In other words, performance indicators allow value judgements about key aspects of the functioning of educational systems in a quantified way (Van den Berghe, 1997).

Standaarden en indicatoren kunnen kwantitatief of kwalitatief, enkelvoudig of meervoudig van aard zijn. Kwantitatieve standaarden of indicatoren zijn numeriek, bijvoorbeeld de gemiddelde uitslag op een eindexamen van een school. Dat is tevens een voorbeeld van een enkelvoudige indicator omdat slechts één variabele gemeten is. Kwalitatieve standaarden en indicatoren vereisen een vorm van expert-beoordeling. Een voorbeeld hiervan is de beoordeling of de opbrengsten van een school overeenkomen met hetgeen van deze school verwacht mag worden. Dat is tevens een voorbeeld van een meervoudige standaard of indicator omdat meerdere variabele een rol spelen bij het oordeel, zoals rendement, lesuitval en de samenstelling van de schoolbevolking.

De standaarden en de indicatoren moeten aan een aantal inhoudelijke en technische voorwaarden voldoen. Enkele belangrijke zijn:

- Elke standaard of indicator moet een geoperationaliseerd element van respectievelijk het domein en de standaard zijn.
- Een standaard of indicator moet enkelvoudig en observeerbaar zijn.
- Een standaard of indicator doet appel op de expertise van inspecteurs.
- Een standaard of indicator moet hanteerbaar zijn in het contact met de instellingen.
- De standaarden- of indicatorenset moet volledig, zuinig, evenwichtig, wederzijds uitsluitend en consistent zijn.

De bronnen waaraan de standaarden en indicatoren uit de eerste waarderingskaders (1991-1997) van de inspectie zijn ontleend is het zogenaamde CIPO-model en de ideeën van de effectieve school (Scheerens, 1990). In het CIPO-model wordt onderscheid gemaakt tussen Context, Input, Proces en Output. Dit model is door Janssens (1991) bij de inspectie

geïntroduceerd om een overzicht te krijgen van zoveel mogelijk bij de evaluatie te betrekken aspecten van het onderwijs van individuele scholen en instellingen, om daaruit vervolgens een beargumenteerde keuze te kunnen maken ten behoeve van de feitelijk uit te voeren evaluatie. Schoolevaluatie in deze optiek staat zowel de evaluatie van een gehele instelling toe, alsook de evaluatie van facetten ervan, zoals: het curriculum (bijv. een vak), van (een onderdeel van) de organisatie van de school (bijv. de leerlingbegeleiding), of van een segment van de school (bijv. de brugklas). Steeds echter staat het CIPO-model centraal, dat wil zeggen dat van het object van evaluatie wordt nagegaan in welke mate de context, de input en/of het proces bijdragen aan de opbrengsten (van het vak, het aspect van de schoolorganisatie of van het segment van de school). De evaluatie leidt tot een overzicht van de sterke en de zwakke kanten van de school of instelling met betrekking tot het evaluatie-object (zie ook Corporaal, Janssens, Sandbergen & Zwarts, 1992).

5.2.2 Normering

Er zijn verschillende soorten normen, die op verschillende manieren gebruikt kunnen worden. Het gebruik hangt vooral af van de evaluatievraag. Bij internationaal evaluatie-onderzoek, zoals bijvoorbeeld Pisa, worden landen onderling met elkaar vergeleken. In dat geval worden relatieve normen gehanteerd, waarbij een kwalificatie wordt toegekend aan de positie in een verdeling.

De evaluaties van de inspectie zijn echter gericht op het beoordelen van de intrinsieke kwaliteit van het onderwijs. In dat geval ligt het hanteren van absolute normen voor de hand. Dit houdt in dat kwalificaties worden toegekend op basis van een inhoudelijke redenering over de mate waarin bepaalde standaarden of indicatoren gerealiseerd zouden moeten zijn. De sterke kant van een dergelijke wijze van normeren is het expliciete karakter van de basis waarop is genormeerd. De zwakke kant van absoluut normeren is dat het tot op zekere hoogte een arbitraire kwestie blijft. Vooral het bepalen van de cesuur is en blijft ‘mensenwerk’ (Janssens, 1997). Bij het bepalen van de cesuur (= grens tussen onvoldoende en onvoldoende realisering van een standaard) gaat de inspectie als volgt te werk. De normen zijn gebaseerd op een beredeneerde keuze uit de (gewogen) geoperationaliseerde indicatoren van een standaard. Daarbij wordt rekening gehouden met relevante informatie uit andere bronnen, zoals empirische gegevens en informatie uit raadplegingen van externe deskundigen.

Bij het opstellen van normen gaat de inspectie uit van de minimale eisen die aan de realisering van de standaarden of indicatoren gesteld kunnen worden, omdat het - voor de overheid - van

belang is in ieder geval te kunnen vaststellen of het onderwijs op een acceptabel niveau wordt gerealiseerd en georganiseerd.

Om tot een valide beoordeling te komen is het nodig om de scores op de indicatoren te koppelen aan het oordeel. Dat oordeel is een feite een kwalificatie van de toestand waarin de standaard of de indicator verkeert. Er zijn verschillende kwalificaties mogelijk, bijvoorbeeld een vier-puntsschaal oplopend van zeer zwak tot en met zeer goed, of een twee-puntsschaal: onvoldoende of voldoende. Die koppeling van een kwalificatie aan de mate waarin een standaard of indicator naar het oordeel van een inspecteur is gerealiseerd vindt plaats in de vorm van beslisregels. In de praktijk varieert de striktheid van de beslisregels. Er zijn evaluaties waarbij de inspecteur vrij zijn oordeel kan formuleren en er zijn er waarbij uit de scores op de indicatoren het oordeel rechtstreeks volgt. Een voorbeeld van de laatste is 'de standaard is voldoende als vier van de zes indicatoren positief zijn'. De voorkeur gaat uit naar het formuleren van één of meer duidelijke beslisregels. Deze regels behoeven echter niet mechanisch toegepast te worden. De inspecteur kan, als daar gronden voor zijn, afwijken van het uit de beslisregel voortvloeiende oordeel, mits deze afwijking wordt beargumenteerd.

Het aantal beslisregels is maximaal gelijk aan het aantal overgangen tussen kwalificaties, respectievelijk het aantal schaalpunten. In de praktijk blijkt het moeilijk om meer dan één beslisregel te formuleren.

Tot nu toe is niet gesproken over de wijze waarop de normering tot stand komt. Vaak is dat een inhoudelijke afweging van de inspectie, waarbij overwogen wordt hoe ernstig het is als bepaalde indicatoren niet worden aangetroffen of welke indicatoren minimaal aangetroffen moeten worden wil van een redelijk realisering van de standaarden sprake zijn. Het is van belang dat dit proces zorgvuldig wordt doorlopen, dat de argumenten goed worden vastgelegd en dat resultaten de instemming hebben van de relevante groepen. De ervaring leert dat bij onwelgevallige uitkomsten de normen het eerst onder de loep worden genomen (zie Treffers, 1996, Wijnstra, 1995 en Van der Linden & Zwarts, 1995; 1996; Van der Linden, Meijer & Vos, 1997).

5.2.3 Data-verzameling

De waarderingskaders vormen het uitgangspunt voor de ontwikkeling en het gebruik van instrumenten die inspecteurs hanteren om van en in individuele scholen gegevens te verzamelen. De inspectie gebruikt bij het verzamelen van gegevens verschillende soorten instrumenten. Bij het lesbezoek wordt gebruik gemaakt van observaties. Interviews worden

gebruikt ten behoeve van gesprekken met directie, leraren, niet-onderwijzend personeel en leerlingen. Documentanalyse vindt plaats bij schoolplannen en andere schooldocumenten.

Bij de constructie van de benodigde instrumenten wordt gebruik gemaakt van de kennis die is opgedaan binnen sociaal-wetenschappelijk onderzoek. De instrumenten behoren te voldoen aan de eisen die aan de validiteit en betrouwbaarheid worden gesteld.

6 DE INSPECTIEWERKWIJZE IN HET EVALUATIELANDSCHAP

Methode, waarden en gebruik

Eerder stelden we dat het vroege werk van toezichthouders aan de wieg stond van verschillende evaluatietheorieën (Alkin & Christie, 2013). Maar deze theorieën hebben op hun beurt ook weer invloed gehad op de verdere ontwikkeling van de evaluatieve component van het onderwijstoezicht. De belangstelling voor evaluatietheorieën binnen de inspectie ontstond begin jaren tachtig in discussies over het evaluatieve karakter van een schoolbezoek en de ontwikkeling van zogenaamde schoolportretten (zie bijv. Mertens, 1982). Er is destijds zelfs een voorstel gedaan om het evaluatiemodel van Stake (1967), *The countenance of Educational Evaluation*, op het schoolbezoek toe te passen. (Wieringa, et al., 1984). Deze vroege pogingen om het schoolbezoek een evaluatief karakter te geven, strandden echter vanwege gebrek aan draagvlak (zie Janssens, 1997, p. 10-20).

De indeling in Methode, Waarden en Gebruik is gebaseerd op de categorie-indeling van Alkin en House (1992, p. 463) en zijn te beschouwen als drie dimensies waarop de benaderingen van invloedrijke evaluatietheoretici zijn te plaatsen (zie figuur 2).

DIMENSIE	CONTINUÛM
METHODE	KWANTITATIEF-----KWALITATIEF
WAARDEN	ENKELVOUDIG-----MEERVOUDIG
GEBRUIK	INSTRUMENTEEL-----CONCEPTUEEL

Figuur 2: Taxonomie van Alkin en House (1992)

De methode-dimensie varieert van kwantitatieve tot en kwalitatieve aanpakken, de waarden-dimensie van een enkelvoudig tot en met een meervoudig waardenperspectief en de gebruiksdimensie van instrumenteel tot en met conceptueel.

Aanvankelijk werden evaluaties in methodisch opzicht hoofdzakelijke kwantitatieve methoden gebruikt, zoals de analyse van toetsprestaties en gerandomiseerde quasi-experimenten, maar tegenwoordig is sprake van een ‘mixed-methode aanpak’, waarbij meer kwalitatieve methoden

worden gebruikt om ook andere gegevens te verzamelen, zoals interviews, observaties en documentanalyse (Mertens & Wilson, 2012).

Om oordelen te kunnen vellen werden aanvankelijk enkelvoudige criteria gebruikt, zoals het evaluatiecriterium of de doelen van een programma wel of niet zijn bereikt. Hiermee werd doorgaans de opdrachtgever van de evaluatie of de ‘eigenaar’ van een programma bediend. Maar om ook andere doelgroepen te informeren over de waarde van het object van evaluatie is steeds meer de nadruk komen te liggen op meervoudige evaluatiecriteria zodat ook de waarden van andere stakeholders met betrekking tot het object in de evaluatie kunnen worden betrokken. Bij de evaluatie van een curriculum kunnen dat zijn: de ontwikkelaars van het curriculum, leraren, ouders en leerlingen. De standpunten over de betrokkenheid van stakeholders bij een evaluatie varieert van ‘representeren’ van hun perspectief (House) tot en met participatie in doel, opzet en uitvoering van de evaluatie (Fetterman, 1996).

Oorspronkelijk waren evaluaties bedoeld om het object van evaluatie te ‘begrijpen’ - kennis van hoe het programma in de praktijk werkt. Dat wordt de enlightenment-functie van gebruik of ‘conceptueel gebruik’ genoemd die gericht is op het verzamelen van generaliseerbare kennis. Daar staat tegenover het verzamelen van gegevens om te kunnen oordelen over de impact van het programma of om beslissingen te kunnen nemen over de vraag of verbetering wel of niet nodig is en op welke onderdelen. Dat betreft de instrumentele kant van het gebruik van evaluaties, vooral bedoeld om direct betrokkenen te informeren.

Aan alle evaluatietheorieën en uitwerkingen daarvan is een methode-, een waarde- en een gebruiksdimensie te onderscheiden. Concrete evaluatie-aanpakken zijn dan ook op deze drie dimensies te positioneren. Evaluatoren houden rekening met methoden, waarden en gebruik wanneer ze evaluaties ontwerpen en implementeren. Via kwantitatieve, kwalitatieve methoden of een combinatie daarvan worden gegevens verzameld. Zij kunnen enkelvoudige of meervoudige criteria gebruiken om de waarde en het nut van een programma te beoordelen. Deze criteria kunnen afhankelijk zijn van de methoden die ze gebruiken en in dienst staan van wel of niet instrumenteel gebruik van de uitkomsten. De manier waarop de evaluatie zijn doelgroepen adresseert laat meestal zien welk type gebruik wordt nagestreefd.

De vraag die we nu gaan beantwoorden is de kwestie op welke positie de inspectie-aanpak op de drie dimensies inneemt.

6.1 Methode-dimensie

Wanneer het evaluatielandschap wordt overzien wat betreft de methodisch-technische aanpak, dan varieert dat aan de ene kant van quasi-experimentele aanpakken, waarbij individuen of groepen ‘at random’ worden toegewezen aan een of meerdere ‘treatments’ en controlegroepen om tot oordelen over de effecten van het object van evaluatie te komen (zie bijv. Boruch et al., 2000), tot en met meer holistische, interpretatieve of narratieve evaluatie-aanpakken aan de andere kant. Elliott Eisner is de meest representatieve vertegenwoordiger van de laatste aanpak waarbij de persoon die evaluatie uitvoert ‘het instrument’ is waarmee zowel de gegevens worden verzameld als de oordelen worden gegeven¹². Zie voor een recente toepassing van deze aanpak Trotman (2006).

Het verschil tussen kwantitatieve en kwalitatieve methoden zit niet in het gebruik van ‘cijfers’ omdat ze in beide ‘kampen’ als data worden gebruikt en ook niet in dataverzamelingstechnieken. In kwalitatief onderzoek wordt ook ‘geteld’ en spreekt men van ‘quasi-statistics’. Daarentegen wordt in kwantitatief onderzoek ook geobserveerd, geïnterviewd en worden documenten geanalyseerd (Janssens, 1985). Het verschil zit hoofdzakelijk in de onderliggende wetenschapstheoretische assumpties over wat ‘kennis van de wereld’ is en in het doel van de evaluatie. In de kwalitatieve traditie probeert men (variabelen) te controleren en gedrag te voorspellen, terwijl de andere traditie vooral ‘de werkelijkheid’ probeert te begrijpen.

What qualitative researchers hold in common is the study of stories or narratives or descriptions of a series of events. These researchers usually embrace the assumption that the story is one if not the fundamental unit that accounts for human experience. But what counts as stories, the kinds of stories they choose to study, or the methods they use for study vary. Within the framework of qualitative research, researchers use a number of research approaches, strategies, and methods (Lieblich, Mashiach-Tuval, & Zilber, 1998).

De kwantitatieve benadering volgt het empirisch-analytische model (Van Koningsveld, 2006) met een nadruk op standaardisering en statistische methoden en -analyse. Het gaat in deze

¹² De connoisseurship-benadering is in de praktijk hoofdzakelijk door Eisner zelf en door zijn studenten toegepast. Een van de problemen van deze benadering is dat het metaforische taalgebruik niet goed aansluit bij de stakeholders en dat de interpretatie van de gegevens persoonlijk van aard zijn en niet altijd gedeeld worden door de direct betrokkenen (Trotman, 2006, p. 259). Zo worden bijvoorbeeld door filmcritici ook bij de beoordeling van films begrippen gebruikt die nogal verhullend zijn: Absurdistisch (= zonder pointe), Caleidoscopisch (= chaotisch), Experimenteel (= onbegrijpelijk), Poëtisch (= traag), Rustige composities (= saai), etc. (vgl. Coen van Zwol in de NRC van 23 januari 2013)

benadering dus niet zozeer om het produceren van cijfers maar om data die geschikt is voor statistische toetsing. In het verleden gebruikten evaluatoren uitsluitend kwantitatieve methoden om de impact van een object van evaluatie te meten, zoals de leerprestaties. Maar in de jaren zestig en zeventig toen met name in de VS grootschalige evaluaties (bijv. van de Title I- en II-onderwijsbeleidsprogramma's¹³) een enorme vlucht namen, bleek dat de kwantitatieve uitkomsten van deze evaluaties onvoldoende aanknopingspunten boden om 'lukken en mislukken' van onderwijsvernieuwingen te verklaren. Evaluatoren moesten ook data achter de cijfers verzamelen om te achterhalen waarin in het ene geval een programma wel werkte en in het andere geval niet (Cronbach et al., 1981). Met name methoden om kwalitatieve gegevens te verzamelen, zoals observaties en interviews, boden uitkomst om data uit de 'black box' te verzamelen (zie bijv. Hall & Loucks, 1977; Astbury & Leeuw, 2010). Daardoor werden kwalitatieve methoden steeds meer gemeengoed in het evaluatielandschap (Worthen & Sanders, 1987; Alkin & House, 1992).

Programmatheorie

Een van de best uitgewerkte en meest toegepaste methodes om te evalueren is de zogenaamde 'theory-driven-evaluation approach' of theory-based evaluation. Deze aanpak is ontwikkeld door Peter Rossi, Carol Weiss en Huey-Tsych Chen (Mertens & Wilson, 2012, p. 60 e.v.).

Evaluatietheorie en programmatheorie zijn verschillende zaken. Evaluatietheorieën houden zich bezig met de wijze waarop een evaluatie moet worden uitgevoerd, terwijl een programmatheorie zich richt op de aard van het object van evaluatie en de aannames die eraan te grondslag liggen over hoe het object van evaluatie zou moeten functioneren, zoals een curriculum of een interventieprogramma. Het gaat dan om de (re)constructie van:

- Een plausibel model van de wijze waarop een programma geacht wordt te werken (Bickman, 1987);
- Een serie veronderstellingen over wat er in een 'black box' gebeurt tijdens de transitie van input naar output om van een ongewenste naar een gewenste situatie te komen (Lipsey, 1993), bijvoorbeeld de werking van een wet of beleidsprogramma;

¹³ Title I: School Improvement, Reading First, Even Start Family Literacy, Migrant Children, Children who are Neglected or Delinquent, and Comprehensive School Reform.
Title II: Teacher and principal training, and Enhancing Education through Technology.

- Het proces waarlangs programma-onderdelen verondersteld worden de opbrengsten te bewerkstelligen en de omstandigheden waaronder het proces geacht wordt te verlopen (Donaldson, 2007).

De basis voor deze constructie wordt gevormd door de beoogde werkwijze van het programma zelf of van de gedragswetenschappelijke theorieën waarop het programma is gebaseerd of waaraan het appelleert (Donaldson, 2007). Een programmatheorie heeft als functie het evaluatieproces te sturen omdat het de criteria levert waaraan het object van evaluatie zou moeten voldoen en derhalve ook de gegevensverzameling en de oordeelsvorming bepaalt (Donaldson, 2007).

Voorbeelden van een programmatheorie in een evaluatiecontext zijn:

- de beoogde werking van een curriculum - het curriculum-zoals-bedoeld - zoals bijvoorbeeld Appelhof (1979) over een bepaalde aanpak van gedifferentieerd lesonderwijs of Streumer (1988) over een curriculum voor natuuronderwijs;
- de beoogde werking van didactische werkvormen of van het didactisch handelen, zoals bijvoorbeeld Janssens (1986) over de praktijk van het evalueren in de klas.
- de beoogde werking van onderwijsbeleid, zoals bijv. Cramer, Scheerens en Zegveld (1988) over het Informatica Stimuleringsplan, of Janssens en De Wolf (2009) over Educational Governance;
- de beleidstheorie achter een wet, zoals Ehren, De Leeuw en Scheerens (2005) over de wet op het onderwijstoezicht (WOT);
- de waarderingskaders van onderwijsinspectie waaraan een serie assumpties ten grondslag liggen waaraan het onderwijs zou moeten voldoen.

Het apriori beschrijven van een object van evaluatie is sterk verbonden met theory-based evaluaties omdat juist via deze benadering de discrepantie wordt opgespoord tussen hoe een programma wordt geacht te werken en hoe het in de praktijk is gerealiseerd. In de evaluatiepraktijk wordt daarbij in veel gevallen gebruik gemaakt van twee modellen om degelijke beschrijvingen te ontwikkelen (Mertens & Wilson, 2012, p. 244 e.v.). De eerste is de 'Logic Model Development Guide' van de W.K. Kellogg Foundation (WKKF, 2004) en de tweede is het CIPO-model van Stufflebeam (zie bijv. Sidani & Sechrest, 1999). In beide aanpakken wordt het te evalueren programma ge(re)construeerd in termen van input, processen, output en context.

6.2 Positionering inspectie

Tussen de experimentele evaluatie-aanpak van Boruch en de hoogstpersoonlijke subjectieve aanpak van Eisner zitten vele varianten waarin de gegevensverzameling en de oordeelsvorming min of meer worden geobjectiveerd door toepassing van regels en instrumenten om tot uitkomsten te komen die onafhankelijk zijn van de technische uitvoering van een evaluatie. Dit is in hoge mate ook van toepassing op evaluaties zoals uitgevoerd door de inspectie.

Inspecteurs worden getraind in de toepassing van waarderingskaders, waarderingskaders worden extern gevalideerd en de gegevensverzameling en de oordeelsvorming worden in hoge mate gestructureerd door regels en intersubjectieve instrumentatie. Op de methoden-dimensie zit het inspectiewerk ergens in het midden. Er is geen sprake van een zuiver kwantitatieve methode omdat de gegevensverzameling niet primair gericht is op het verzamelen van numerieke gegevens of op inferente statistische analyses om causale relaties vast te leggen. Er is ook geen sprake van een zuiver kwantitatieve benadering omdat de gegevensverzameling niet bepaald wordt door de wens om de schoolwerkelijkheid te beschrijven zoals deze feitelijk is maar af te zetten tegen een wenselijk kader van deugdelijkheid- en kwaliteitseisen.

Aan de ene kant is dus sprake van een voorgestructureerd perspectief op naleving van wettelijke voorschriften en uitvoering van onderwijs dat aan bepaalde eisen voldoet en aan de andere kant is de gegevensverzameling en de oordeelsvorming niet maximaal in regels, voorschriften en instrumenten geprecodeerd. De persoon van de inspecteur heeft een zekere mate van vrijheid om te handelen en te oordelen 'naar bevind van zaken' en rekening te houden met de feitelijke schoolwerkelijkheid en haar omstandigheden. Dat heeft niet alleen te maken met de discretionaire bevoegdheid van de inspecteur maar ook met de aard van de standaarden en indicatoren in het waarderingskader die in hoge mate een beroep doen op de expertise van de inspecteur.

Een belangrijk punt bij de hantering van waarderingskaders en technieken en instrumenten om gegevens te verzamelen, is dat er rekening is gehouden met de benutting van de expertise van inspecteurs (in zin van Eisner's Educational connoisseurship). Door hun ervaring en veldkennis zijn zij in principe in staat om op hoog inferentieniveau betrouwbare en valide informatie te verzamelen en daarover te oordelen (Janssens, 1997). Dat leidt ertoe dat de inspectie zoveel mogelijk gebruik maakt van instrumenten met een hoog-inferentiële karakter, dat wil zeggen dat de waarnemingen en de daarop gebaseerde oordeelsvorming in

geringe mate worden voorgestructureerd. Oordelen vragen dus relatief veel interpretatie van de zijde van de beoordelaar (Rosenshine, 1970¹⁴; Muys & Reynolds, 2005).

Voorbeelden van indicatoren die een hoog-inferentieel oordeel vragen zijn (Inspectie van het Onderwijs, 2012):

- *De leraren bevorderen door hun handelen de verwerving van sociale en maatschappelijke competenties van leerlingen.*
- *De leraar past de onderwijskundige ambities/doelen, die op schoolniveau zijn geformuleerd, toe in zijn les.*

De beoordeling van deze indicatoren vragen dat vanuit meerdere databronnen (lesobservaties, interviews en documentanalyse) gegevens worden verzameld en door de inspecteur onderling gewogen worden om tot een oordeel te komen.

Voorbeelden van indicatoren die een lager inferentieel oordeel vragen omdat volstaan kan worden met minder gegevens die soms slechts uit één bron komen, bijvoorbeeld uit een schooldocument, zijn. Voor een juiste beoordeling van deze indicatoren is geen specifieke expertise nodig.:

- *De schoolleiding verantwoordt zich intern over de gerealiseerde onderwijskwaliteit.*
- *De school heeft doelen gesteld die erop gericht zijn om achterstanden te bestrijden.*

6.3 Waarden-dimensie

De waarden-dimensie van Alkin c.s. kent aan de ene kant theoretici die uitgaan van een enkelvoudig waardenperspectief en aan de andere kant theoretici die voor meervoudige waarden in de oordeelsvorming pleiten. Belangrijke representanten van deze tak zijn Ernest House, David Fetterman en Huey Chen. De overeenkomst tussen House (bijv. 1991) en Fetterman (bijv. 1996) is opmerkelijk; beiden zijn begaan met ‘empowering’ van de

¹⁴ Rosenshine (1970) schreef een hoofdstuk in de Review of Educational Research over “available instruments for the observation of classroom instruction and to suggest modifications for local evaluation of instruction” (p. 279). Rosenshine groepeerde klasse-observatie-instrumenten in twee groepen: category systems en rating systems. Hij stelde dat ‘category systems are low-inference measures because they focus on specific, observable, objective behaviors and, also, because the events “are recorded as frequency counts’ (p. 281). Hij classificeerde ‘rating systems as high-inference because they lack such specificity and because the observer must infer the constructs to be rated, such as enthusiasm, clarity of presentation, or supportive of students. Further, rating systems require the observer to infer the frequency of such behaviors to arrive at ratings such as consistently, sometimes, or always (p. 281).

stakeholders en zijn derhalve voorstanders van meervoudige beoordelingscriteria om tegemoet te komen aan de behoeftes van de verschillende stakeholders. Chen vertegenwoordigt de andere kant van het continuüm (enkelvoudig waardenperspectief). Zijn ‘theory-driven evaluation approach’ (Chen, 1990) is vooral gericht op het toepassen van een programmatheorie om een evaluatie-design te ontwerpen.

Enkelvoudige waarde-oordelen

Een van de ontwikkelaars van de theory-based-benadering, Peter Rossi (Rossi, Freeman & Wright, 1979; Rossi, Freeman & Lipsey, 1999), hechtte bijzonder veel waarde aan een hoge validiteit van evaluaties. De validiteit kon in zijn ogen het best gewaarborgd worden door uit te gaan van een programmamodel van het object van evaluatie en dit model te vergelijken met de uitvoering van het programma in de praktijk waardoor ook een breder publiek bij de uitkomsten van de evaluatie betrokken kon worden. Bij evaluaties waar het gaat om de vraag of iets wel of niet overeenkomstig de bedoelingen werkt, bijvoorbeeld de vraag of een bepaalde didactische aanpak het beoogde effect heeft, is juist een goede apriori-theorie nodig om te kunnen verklaren welke factoren nu precies wel of geen bijdrage aan het geobserveerde effect heeft gehad.

‘We have argued for a paradigm that accepts experiments and quasi-experiments as dominant research designs, but that emphasizes that these devices should be used in conjunction with a priori knowledge and theory to build models of the treatment process and implementation system to produce evaluations that are more efficient and that yield more information about how to achieve desired effects.’ (Chen & Rossi, 1983, p. 300)

Bij de ‘theory-based approach’ is sprake van een enkelvoudig waardenperspectief omdat het oordeel over de mate waarin de programmatheorie overeenkomt met de praktische uitvoering het enige criterium wat voor het oordeel telt.

Meervoudige waarde-oordelen

Om uiteindelijk een waarde-oordeel te kunnen uitspreken over het object van evaluatie zijn criteria nodig. Volgens Alkin and House (1992), varieert de waarde-dimensie van het gebruik van een enkelvoudig tot en met een meervoudig perspectief. Van oudsher ontleen evaluatoren deze criteria aan de geformuleerde of aan de beoogde doelen van een programma. Als bijvoorbeeld het doel van het programma is dat leraren meer gedifferentieerd gaan

lesgeven dan is dat ook tevens het beoordelingscriterium. Als de doelen niet expliciet zijn geformuleerd dan dienen de evaluatoren deze apriori te reconstrueren, bijvoorbeeld door de ontwikkelaars erop te bevragen.

In veel gevallen spelen ook andere waarden-perspectieven dan die van de ontwikkelaars een rol in de beoordeling van een programma. Door meervoudige perspectieven in de beoordeling te betrekken worden de uitkomsten van een evaluatie ook interessanter voor andere stakeholders. Het is dan de taak van de evaluatie op welke wijze een finaal oordeel tot stand moet komen en wat het gewicht is dat aan de verschillende perspectieven moet worden toegekend. Sommige theoretici, zoals Fetterman en Cousins (Cousins & Earl, 1999) gaan zelfs zo ver dat ze stakeholders in alle fasen van het evaluatieproces betrekken om maximaal gebruik van de uitkomsten te kunnen waarborgen.

(Empowerment Evaluation). Anderen, zoals House, Stufflebeam en Patton (2008), gaan minder ver in de betrokkenheid van stakeholders en zien hun betrokkenheid vooral als een manier om eventuele 'evaluator bias' te reduceren door *'inviting extensive communication with stakeholders and by "promoting extensive deliberation about the study's conclusions,"* (House, 2004, p. 220). House pleit er vooral voor de input van alle stakeholders bij de evaluatie te betrekken en niet alleen van degenen die het programma hebben ontwikkeld of uitvoeren.

6.4 Positionering inspectie

De waardendimensie van inspectie-evaluaties is te vinden de waarderingskaders, met name in de kwaliteitsaspecten, de daarbij behorende indicatoren en in de normering. Deze kaders bevatten de eisen die vanuit verschillende perspectieven gesteld worden aan het object van toezicht. Deze kaders hebben de kenmerken van een 'programma' zoals hierboven bedoeld. Daarin zijn naast het perspectief van de inspectie als onderdeel van het openbaar bestuur, de wetenschap en onderwijsdeskundigen ook de perspectieven betrokken van verschillende stakeholders, zoals het georganiseerd onderwijsoverleg, bestuurders, schoolleiders, leraren, ouders en deelnemers.

Wat betreft de positionering van de inspectie-aanpak op de waardendimensie van evaluatietheoretici is sprake van een theorie- of programma-gedreven evaluatiebenadering in de traditie van Peter Rossi, Carol Weiss en Huey-Tsych Chen, waarbij sprake is van een meervoudig waardenperspectief.

Het meervoudige waardenperspectief zit dichterbij de opvattingen van Ernest House dan die van bijvoorbeeld David Fetterman. House is een voorstander van het betrekken van het perspectief van anderen bij de evaluatie, terwijl Letterman een voorstander is van 'samen evalueren'. Zover gaat de werkwijze van de inspectie niet en daar heeft de inspectie ook goede redenen voor.

In de eerste plaats heeft de inspectie namens de overheid de wettelijke opdracht oordelen over de kwaliteit van het onderwijs uit te spreken en ook openbaar te maken¹⁵. Vanuit deze eigenstandige verantwoordelijkheid dient voor de scholen en ook voor het publiek duidelijk te zijn dat deze oordelen uiteindelijk het perspectief van de overheid vertegenwoordigen. Omdat de overheid ook het algemeen belang vertegenwoordigt dient de inspectie in haar oordelen wel rekening te houden met de opvattingen van derden over de eisen die aan het object van toezicht gesteld worden. Daarom is in de Wet op het onderwijstoezicht (WOT) ook een bepaling opgenomen over overleg over het toezichtskader¹⁶.

In de tweede plaats dient de evaluatie-aanpak van de inspectie uit het oogpunt van transparantie, rechtsgelijkheid en zorgvuldigheid systematisch en zoveel mogelijk gestandaardiseerd te zijn. Dat houdt in dat de inspectie-aanpak betrouwbare en valide gegevens en billijke en rechtvaardige oordelen over de kwaliteit van het onderwijs (aan individuele scholen) dient op te leveren. De gebruiksmogelijkheden van evaluatie-uitkomsten worden in hoge mate bepaald door de kwaliteit van het oordeel dat uit de evaluatie voortvloeit. Aan dit oordeel kan een aantal eisen worden gesteld: dat moet betrouwbaar, zorgvuldig, rechtvaardig en relevant zijn (Janssens, 1997).

- Een oordeel is betrouwbaar als de juiste kwalificatie van de stand van zaken wordt gegeven.
- Een oordeel is zorgvuldig als alle relevante informatie daarbij betrokken is.
- Een oordeel is rechtvaardig als in gelijke situaties een gelijk oordeel wordt gegeven.
- Een oordeel is relevant als het voor de geëvalueerde en/of het geëvalueerde object van betekenis is.

Carol Weiss (1991) heeft in haar werk gewezen op de politieke context waarin de meeste evaluaties plaatsvinden. Dat geldt ook voor evaluaties vanwege de Inspectie van het

¹⁵ Artikel 3, lid 2 van de Wet op het onderwijstoezicht

¹⁶ Artikel 13, lid 2: Alvorens een toezichtskader vast te stellen of te wijzigen voert de inspectie overleg met vertegenwoordigers van het onderwijsveld en andere betrokkenen, terwijl bij onderwerpen betrekking hebbend op de vrijheid van inrichting in ieder geval overleg wordt gevoerd met de erkende richtingen.

Onderwijs. Evaluaties worden vaak beïnvloed door de politieke context, zowel wat betreft het proces als de uitkomsten ervan.

“Politics intrudes on program evaluation in three ways: (1) programs are created and maintained by political forces; (2) higher echelons of government, which make decisions about programs, are embedded in politics; and (3) the very act of evaluation has political connotations.” (Weiss, 1991, p. 213)

Om die reden dienen dergelijke evaluaties volgens Weiss ook te voldoen aan de zwaarste kwaliteitscriteria: *“Research quality is of value not only because of adherents to the norms of science, but perhaps more importantly because it increases the power of research as ammunition in intra-organizational argument”* (Weiss & Bucuvalas, 1980, p. 256).

Ook aan inspectie-evaluaties kunnen politieke consequenties worden verbonden die een onomkeerbaar karakter hebben, zoals het toekennen van sancties aan individuele scholen of de beïnvloeding van het onderwijsbeleid. Ook vanuit dat oogpunt dienen inspectie-evaluaties aan de hoogste kwaliteitseisen te voldoen en, om met Cronbach en Suppes (1969, p. 15) te spreken, *‘conducted and reported in such a way that the argument can be painstakingly examined. The report does not depend for its appeal on the eloquence of the writer or on any surface plausibility. The argument is not justified by anecdotes or casually assembled fragments of evidence.’*

Uit de positie en de verantwoordelijkheid die de inspectie heeft om de hoogste kwaliteitseisen voor haar evaluatiewerk in acht te nemen. Bij deze eisen passen methodisch-technische systematisering en standaardisering en op zijn minst intersubjectiviteit¹⁷. Deze zouden onder druk komen te staan als oordelen subjectief zijn, er onduidelijkheid ontstaat over het perspectief van waaruit geoordeeld en over de positie die stakeholders hebben in het evaluatiewerk van de inspectie. Chen en Rossi (1980) waarschuwen voor een te nauwe betrokkenheid van stakeholders in de waardendimensie van evaluaties omdat hun opvattingen vertekend en subjectief kunnen zijn.

¹⁷ Objectiviteit in strikte zin is ook in zuiver wetenschappelijk onderzoek niet mogelijk (De Groot, 1961); waarnemingen en oordelen zijn gebaseerd op menselijke selecties uit de mogelijke waarnemingen, menselijke zintuigen en menselijk taal- (en ander teken-)gebruik. Daarom wordt de eis van objectiviteit meestal afgezwakt tot de eis van intersubjectiviteit: we eisen dat waarnemings- en analysemethoden zodanig zijn dat verschillende onderzoekers, die hetzelfde object bestuderen, tot dezelfde conclusies komen. Een belangrijk hulpmiddel daarbij is het gebruik van gestandaardiseerde methoden. In het (wetenschappelijk en wetenschappelijk gefundeerd) onderzoek spreekt men van intersubjectiviteit om aan te duiden, dat men uitspraken doet over een werkelijkheid vanuit de waarneming en interpretatie van de betrokken wetenschapper. Dat geldt ook voor het inspectiewerk. Dit dient op een zodanige manier te gebeuren dat ook inspectie-oordelen door anderen (virtueel) moeten kunnen worden herhaald, geverifieerd dan wel gefalsificeerd.

House (1991) wijst in dit verband op de volgende valkuilen in het beoordelingsperspectief van evaluatoren:

1. Cliëntisme: het belang van de klant, bijvoorbeeld de ouder, voorop stellen;
2. Manageralisme: het belang van de manager, bijvoorbeeld het bestuur of de schoolleiding, boven alles stellen;
3. Methodologicalisme: er vanuit gaan dat een degelijke opzet en uitvoering van een evaluatie alle ethische problemen oplost;
4. Elitisme: alleen rekening houden met de perspectieven van de machtigste stakeholders;
5. Relativisme: evenveel waarde hechten aan alle perspectieven.

6.5 Gebruiksdimensie

Evaluaties worden uitgevoerd om uiteindelijk ergens een beslissing over te nemen om hetzij tot een summatief waarde-oordeel te komen (werkt het wel of werkt het niet), hetzij het object van evaluatie te verbeteren. Alkin en House (1992) plaatsen het gebruik van evaluaties op een continuüm dat loopt van instrumenteel tot en met conceptueel. Van oudsher werden evaluatieresultaten gebruikt om het functioneren van een programma te begrijpen. Het diende vooral een wetenschappelijk doel. Dit wordt of conceptueel gebruik of ‘enlightenment’ (Weiss, 2005) genoemd. Aan de andere kant van het continuüm zit het instrumentele gebruik van evaluaties waarbij vooral het verantwoordings-, verbeterperspectief of beleidsontwikkeling centraal staan.

Evaluations conducted for enlightenment purposes are similar to those of research studies and are intended for similar academic audiences. Enlightenment evaluations produce knowledge on the program under evaluation. Instrumental evaluations directly inform stakeholders and provide the information needed to make decisions (Alkin & House, 1992).

Vertegenwoordigers van theoretici die zich vooral bezig hielden met het gebruik zijn J. Bradley Cousins, Michael Patton en Daniel Stufflebeam, die allen aan de instrumentele kant van de dimensie zijn te plaatsen.

Stufflebeam is een van de meeste bekende evaluatietheoretici en hij bepleitte dat het centrale doel van een programma-evaluatie zou moeten zijn om het programma te verbeteren (Stufflebeam, 1983). Zijn theorie staat ook wel bekend als “decision-oriented” evaluatie. Hij is ook de geestelijke vader van The Program Evaluation Standards (Joint Committee on

Standards for Educational Evaluation, 1994) die betrekking hebben op: Utility, Feasibility, Propriety, and Accuracy. Deze standaarden zijn ontwikkeld om te borgen dat evaluatie voldoen aan de behoeften van bedoelde gebruikers, realistisch en ethisch worden uitgevoerd en adequate en accurate informatie opleveren over de waarde en het nut van te evalueren programma's. Zie voor een Nederlands equivalent 'Het oog der natie: scholen op rapport. Standaarden voor de publicatie van schoolprestaties' (Dijkstra et al., 2001).

Sommige evaluaties worden uitgevoerd met het vooropgezette doel om aanknopingspunten te vinden om een programma te verbeteren en zijn dus in optima forma gericht op de instrumentele kant van het gebruik. Dit is bijvoorbeeld het geval bij zelfevaluaties door scholen of bij de evaluatie van beleidsprogramma's, zoals Weer Samen Naar School (Peschar & Meijer, 1997). Andere programma-evaluaties zijn eerder opgezet vanuit een accountability-perspectief om verantwoording af te leggen over de toestand waarin een programma verkeert. Dat is bijvoorbeeld het geval bij jaarverslagen of bij rapporten over de kwaliteit van het onderwijs, zoals de onderwijsinspectie die per school of over de staat van het onderwijs publiceert. Deze rapporten zijn enerzijds bedoeld om stakeholders te informeren maar kunnen anderzijds ook leiden tot acties door individuele scholen, ouders en leerlingen, de politiek, het onderwijsbeleid of door de inspectie zelf, bijvoorbeeld om het toezicht op een school te intensiveren. Ook deze vormen van gebruik zijn instrumenteel te noemen. Er zit wel enige spanning tussen evaluaties primair gericht op verantwoording en evaluaties die vooral verbetering op het oog hebben omdat er verschillen zijn in de informatie die nodig is voor het afleggen van verantwoording en gegevens die nodig zijn om tot verbetering te komen (Janssens & Van Amelsvoort, 2008; Patton, 2008).

Gebruik en niet-gebruik

Ook al is het gebruik van evaluatie-uitkomsten instrumenteel bedoeld wil dat niet zeggen dat de uitkomsten ook zodanig worden gebruikt. Zorgen om het gebruik van evaluaties is niet nieuw: het houdt de gemoederen al jarenlang bezig. Verschillende onderzoekers zijn het gebruik van evaluatie-uitkomsten nagegaan en hebben ook onderzocht welke factoren het gebruik beïnvloeden (bijv. Chelimsky, 1977). Johnson et al. (2009) onderzocht over een periode van 25 jaar studies naar het gebruik van evaluaties. Factoren die een gunstige invloed hebben op het gebruik zijn de betrokkenheid van stakeholders, maar nog meer invloed hebben, de competentie van de evaluator en een goede interactie tussen en communicatie met de evaluatoren en de personen die object van evaluatie zijn (Johnson et al., 2009, p. 398).

De internationale literatuur over benutting van onderzoek en evaluatie onderscheidt een groot aantal factoren die kunnen bijdragen tot gebruik van evaluatieresultaten (Leviton & Hughes, 1981, Nutley, Walter & Davies, 2003, Leeuw, 2009):

Onderzoek en evaluatie staat of valt bij de geloofwaardigheid van wat de evaluatie ons leert en hoe het ons wordt geleerd (Chelimsky, 1995). De waarde daarvan baseert zich onder andere op het gebruik van een wetenschappelijk verantwoord onderzoeksontwerp. Andere belangrijke indicatoren voor wetenschappelijkheid en daarmee de geloofwaardigheid van evaluaties zijn validiteit, betrouwbaarheid, objectiviteit en onafhankelijkheid (Chelimsky, 1995).

Evaluaties produceren verschillende typen kennis. Zo maken Nutley, Walter en Davies (2003) een onderscheid tussen vijf typen evaluatieve kennis, te weten *'knowledge about problems, what works, how to put it into practice, who to involve, en why'*. Het maken van onderscheid in de aard van de geproduceerde kennis is vooral belangrijk omdat er een mismatch kan zijn tussen wat de onderwijspraktijk, beleidsmakers, bestuurders en politici verwachten, hopen of wensen dat aan kennis geproduceerd wordt en wat de evaluator daadwerkelijk biedt. Uiteenlopende verwachtingen over wat aan inzichten over de impact van een interventie geproduceerd wordt, kunnen een belangrijke factor zijn bij het niet gebruiken van de resultaten. Als gehoopt wordt dat uitspraken over impact gedaan worden, en over het soort instrumenten waarmee effecten al of niet te bereiken zijn en vervolgens informatie geproduceerd wordt over de vraag of het in de praktijk is uitgevoerd conform de procedurele en organisatorische voorwaarden in bijvoorbeeld het beleidsontwerp, dan kan dit leiden tot het niet gebruiken van evaluatie-uitkomsten.

6.6 Positionering inspectie

Evaluaties van de inspectie dienen primair een instrumenteel doel. Ze zijn niet bedoeld om tot wetenschappelijke theorievorming te komen of de praktijk van het onderwijs te begrijpen, zoals bijvoorbeeld in John Goodlad's (1984) *'A Place Called School: Prospects for the Future'*¹⁸. Ze zijn bedoeld om over de stand van zaken van het onderwijs te oordelen en daarover verantwoording af te leggen met het oog op het nemen van beslissingen op school- of landelijk niveau. Bij het gebruik van inspectie-evaluaties staat dus het accountability- en verbeterperspectief centraal. Ofschoon het beoogde gebruik van inspectie-evaluaties instrumenteel van aard is, wil dat niet zeggen dat ze ook als zodanig worden gebruikt. Voor

¹⁸ Beschrijving en analyse van de schoolpraktijk in de VS gebaseerd op gegevens van meer dan 27000 leerlingen, leraren en ouders en op meer dan 1000 klassenobservaties.

het Onderwijsverslag geldt mogelijk dat het eerder fungeert om politiek en beleid inzicht te verlenen in thema's, om problemen te verhelderen, oude waarheden open te breken, enzovoorts, dit ter verbreding van de oordeelsvorming.

Naast de begrippen instrumentele en conceptuele benutting zijn er nog legio andere aanduidingen en typering van het gebruik en negeren van evaluatie-uitkomsten. Zo wordt er ook gesproken over 'symbolisch' gebruik. Dit betekent bijvoorbeeld dat een onderzoeksbevinding slechts ter illustratie dient dat een bepaald onderwerp aandacht heeft gehad (Zie voor enkele voorbeelden World Bank, 2004).

Tegenover deze vormen van gebruik kan niet-gebruik dan wel misbruik van evaluatie-uitkomsten geplaatst worden. Daarnaast wordt nog gesproken in de literatuur over evaluaties die politieke standpunten bekrachtigen of wel weerleggen en evaluaties die nuttig zijn vanwege het proces dat daarbij doorlopen wordt. Fabian (2009) verwijst in navolging van Stamm (2003) ook nog op niet-gebruik in de vorm van het opwerpen van blokkades gericht op het voorkomen van enig gebruik, omdat er een aanzienlijke weerstand tegen de onderzoeksbevindingen is.

De gebruikscomponent van het toezicht wordt bepaald door de doelgroepen die de inspectie adresseert. Deze doelgroepen zijn enerzijds de verantwoordelijken voor wetgeving en beleid en de uitwerking daarvan, in het bijzonder de minister, het parlement en het georganiseerde onderwijsveld, en anderzijds de verantwoordelijken voor de realisering van goed onderwijs: het bevoegd gezag en hun scholen of instellingen. De afgelopen jaren zijn daarbij in toenemende mate ook ouders, leerlingen, deelnemers en de publieke opinie een rol gaan spelen (Dijkstra, Karsten, Veenstra & Visser, 2001). In beginsel houdt dit in dat de inspectie in haar rapportages over het onderwijs in het algemeen en over de scholen en instellingen in het bijzonder, met de informatiebehoeften van al deze partijen rekening zou moeten houden. Afhankelijk van de groep die moet worden bediend, stelt dat uiteenlopende eisen (zie bijv. Karsten, Visscher, Dijkstra & Veenstra, 2010).

De gebruikscomponent bij de beoordeling van individuele scholen betreft in de eerste plaats het bestuur en de daaronder ressorterende school, en in het verlengde daarvan de ouders, leerlingen, deelnemers en het publiek in het algemeen. Vanuit instrumenteel oogpunt zouden de bevindingen die het toezicht oplevert, deze partijen feedback moeten geven over verschillende aspecten van de inrichting en organisatie van het onderwijs. Deze informatie kunnen de besturen en de scholen vervolgens aanwenden voor kwaliteitsverbetering en andere stakeholders kunnen deze gebruiken om zich een beeld te vormen van de kwaliteit van scholen.

Voor wetgeving en beleid speelt onder meer de informatie uit het jaarlijkse Onderwijsverslag een rol. Op basis van jaarlijkse representatieve steekproeven worden scholen systematisch onderzocht op de onderdelen van het waarderingskader van de inspectie. Hierdoor ontstaat een dwarsdoorsnede van de kwaliteit van het onderwijs met voldoende aanknopingspunten voor instrumenteel gebruik.

- Algemene Rekenkamer (2002). *Systemen van checks and balances bij rechtspersonen met een wettelijke taak*. Den Haag; Algemene Rekenkamer.
- Alkin, M. C., ed. (2013). *Evaluation Roots: a wider perspective of theorists' views and influences*. Thousand Oaks, Cal. Sage.
- Alkin, M. & Christie, C. (2004). An evaluation theory tree. In: Marvin C. Alkin (ed.), *Evaluation Roots, tracing theorists' views and influences*, San Francisco: Sage publications Inc.
- Alkin, M. C., & House, E. R. (1992). Evaluation of programs. In M. C. Alkin (Ed.), *Encyclopedia of Educational Research* (pp.462-467). New York: MacMillan.
- Appelhof, P. (1979). *Begeleide onderwijsvernieuwing: evaluatie van een curriculum-innovatie gericht op differentiatie van het aanvankelijk lesonderwijs*. Tilburg: Zwijsen BV (Diss).
- Astbury, B. & F.L. Leeuw (2010). Unpacking Black Boxes: Mechanisms and Theory Building in Evaluation. *American Journal of Evaluation*. 31, 3. 363-381 doi: 10.1177/1098214010371972.
- Barone, T.E. (1991). Educational connoisseurship and criticism. In A. Lewy (Ed.) *The International Encyclopedia of Curriculum* (429 - 433). Oxford: Pergamon Press.
- Berg, G. van den, Hoeven, W.Th.J.G. & Van der Werf, G. (1993). De bijdrage van evaluatie-onderzoek aan de kwaliteitsverbetering van het onderwijs. In B.P.M. Creemers, W.Th.J.G. Hoeven, J.L. Peschar & J. Snippe (1993). *Wat is onderwijsonderzoek waard?* De Lier: Academisch Boekencentrum.
- Berghe, W. van den (1997). *Indicators in perspective: The use of quality indicators in vocational education and training*. Thessaloniki: CEDEFOP.
- Bickman, L. (1987). The functions of program theory. *New Directions for Program Evaluation*. 33,5-18.
- Beukenholdt, T. & Dietvorst, C., Red. (2011). *Onderwijsinspectie onder druk*. Den Haag: Boom | Lemma uitgevers.
- Bolam, R. (1990). School Inspectors and Advisors. In: Walberg, H.J. & Haertel, G.D. (1990). *The International Encyclopedia of Educational Evaluation*. Oxford: Pergamon Press Inc.
- Boonstra, H. (2002). *De beoordelaar beoordeeld: analyse van de inspectie basisonderwijs*. Groningen: Chopper Publishers.

- Boruch, R., Snyder, B., & DeMoya, D. (2000). The importance of randomized field trials. *Crime and Delinquency*, 46 (2), 156-180.
- Bredo, E. (2006). Philosophies of Educational research. In: Green, J.L., Camilli, G., Elmore, P.B. et al.: *Handbook of Complementary Methods in Education Research*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.
- Bruggen, J.C. van (2010). *Inspectorates of Education in Europe; some comparative remarks about their tasks and work*. <http://www.sici-inspectorates.eu/en/87754-inspectorates-of-education-in-Europe>.
- Campbell, D.T. & Stanley, J.C. (1966). *Experimental and Quasi-Experimental Designs for Research*. Chicago, Illinois: Rand McNally.
- Chelimsky, E. (1977). *Proceedings of a symposium on the use of evaluation by federal agencies*. Washington, DC.: National Institute for Law Enforcement and Criminal Justice.
- Chelimsky, E. (1995). Where we stand today in the practice of evaluation. *Knowledge and Policy*, 8: 8 - 19.
- Chen, H. (1990). *Theory-driven evaluations*. Newbury, CA: Sage.
- Chen, H., & Rossi, P. (1987). The theory-driven approach to validity. *Evaluation and Program Planning*, 10, 95-103.
- Cramer, J.S., J. Scheerens, and W. Zegveld. (1988). *De evalueerbaarheid van het Informatica Stimuleringsplan (INSP)*. Groningen: Commissie Programma Evaluatie.
- Christie, A.C. & Alkin, M.C. (2013). An Evaluation Theory Tree. In: Alkin, M.C., ed. *Evaluation Roots: a wider perspective of theorists' views and influences*. Thousand Oaks, Cal: Sage.
- Christie, A. C., & Alkin, M. C. (2008). Evaluation theory tree re-examined. *Studies in Educational Evaluation*, 34, 131-135.
- Corporaal, A.H., Janssens, F.J.G., Sandbergen, S. & Zwarts, M. (1992). *Evaluatie door de inspectie*. Zoetermeer: Inspectie van het Onderwijs.
- Cousins, J. B., & Earl, L. M. (1999). When the boat gets missed: Response to M. F. Smith. *American Journal of Evaluation*, 20(2), 309-317.
- Cronbach, L.J. (1982). *Designing Evaluations of Educational and Social Programs*. San Francisco: Jossey- Bass.
- Cronbach, L.J and Associates (1981). *Toward reform of program evaluation*. San Francisco: Jossey-Bass Publishers.
- Cronbach, L.J. & Suppes, P. (1969) *Research for tomorrow's schools. Disciplined Inquiry for Education*. New York: MacMillan.

- Dijkstra, A.B. & Janssens, F.J.G., red., (2012). *Om de kwaliteit van het onderwijs: kwaliteitsbepaling en kwaliteitsbevordering*. Den Haag: Boom | Lemma uitgevers.
- Dijkstra, A.B., Karsten, S., Veenstra, R., & Visscher, A.J. (Eds.), *Het oog der natie: scholen op rapport. Standaarden voor de publicatie van schoolprestaties*. Assen: Van Gorkum.
- Donaldson, S. I. (2007). *Program theory-driven evaluation science*. New York, NY: Lawrence Erlbaum.
- Donaldson, S.L. & Lipsey, M. W. (2006). Roles for theory in contemporary evaluation practice: developing practical knowledge. In: Shaw, I.F., Greene, J.C. & Mark. M.M. (2006). *The Sage Handbook of Evaluation*. London: Sage.
- Dodde, N.L. (2001). *Een speurtocht naar samenhang: het rijksschooltoezicht van 1801-2001*. Den Haag: Sdu Grafisch Bedrijf bv.
- Ehren, M.C.M, Leeuw, F.L. & Scheerens, J. (2005). On the Impact of the Dutch Educational Supervision Act: Analyzing Assumptions Concerning the Inspection of Primary Education. *American Journal of Evaluation*. (26) 1, 60-76.
- Eisner, E.W. (1979). The use of qualitative forms of evaluation for improving educational practice. *Education Evaluation and Policy Analysis*, 1, 11-19.
- Eisner, E.W. (1977). On the uses of educational connoisseurship and criticism for evaluating classroom life. *Teachers College Record*. 78, 345 - 358.
- Elte, R. (1988). *Van het toezicht op het onderwijs: over de inrichting en het functioneren van de inspectie in het algemeen voortgezet onderwijs en het lager beroepsonderwijs*. De Lier: Academisch Boeken Centrum. (diss.)
- Elte, R. & Scholtes, E. (2001). *Uit de luwte: over strategische veranderingen in en rond de onderwijsinspectie*. Den Haag: SDU Grafisch bedrijf bv.
- Fabian, C. (2009). Nutzung der Evaluationsfunktion im Vergleich. In: T. Widmer et al (Hrsg.) *Evaluation; ein systematisches Handbuch*: 501 - 506. Wiesbaden: VS Verlag fuer Sozialwissenschaften.
- Fetterman, D. M. (1996). Empowerment evaluation: An introduction to theory and practice. In D. M. Fetterman, S. J. Kaflarian, & A. Wandersman (Eds.). *Empowerment Evaluation*. Thousand Oaks, CA: Sage.
- Fournier, D.M. (2005). Evaluation. Pp. 139-40 In: S. Mathison, ed. *Encyclopedia of Evaluation*. Thousand Oaks, CA: Sage.
- Geertz, C. (1973). Thick description: Toward an interpretive theory of culture. In C. Geertz (Ed.), *The interpretation of cultures* (pp. 3-30). New York: Basic Books.

- Goldstein, H. & Spiegelhalter, D. J. (1996). League Tables and their Limitations: Statistical Issues in Comparisons of Institutional Performance. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, Vol. 159, No. 3. (1996), pp. 385-443.
- Goodlad, J. I. (1984). *A Place Called School: Prospects for the Future*. New York: McGraw-Hill.
- Groenendijk, C.A. (1981). *Bundeling van belangen bij de burgerlijke rechter*. Zwolle: Tjeenk Willink (Diss.)
- Groot, A.D. de (1986). *Begrip van evalueren*. Den Haag: VUGA.
- Groot A.D. de (1961). *Methodologie. Grondslagen van onderzoek en denken in de gedragswetenschappen*. Van Gorcum, Assen.
- Guba, E. & Lincoln, Y. (1989), *Fourth Generation Evaluation*. Beverly Hills: SAGE Publications Inc.
- Hall, G.E. & S.F. Loucks (1977). A developmental model for determining whether the treatment is actually implemented. *American Educational Research Journal*, 13, 263-276.
- Hoeben, W.Th.J.G. (1993). Evaluatie van onderwijsbeleid. In Nijhof et al. *Handboek Curriculum. Modellen, Theorieën, Technologieën*. (p.165-190). Amsterdam/Lisse: Swets & Zeitlinger.
- Hoeben, W.Th.J.G. (1995). Integrale kwaliteitszorg. In J. Lowyck en N. Verloop (Red.). *Onderwijskunde; een kennisbasis voor professionals*. (hoofdstuk 9, pp. 286-314). Groningen: Wolters-Noordhoff.
- Hoeben, W.Th.J.G (1997). De inspecteur: enquêteur of connoisseur? Een kritische beschouwing van de zich uitkristalliserende methodologie van de Inspectie van het Onderwijs. In Janssens, F.J.G., red. (1997). *Van onderzoek naar evaluatie: de methodologie van de onderwijsinspectie*. Den Haag: Delwel Uitgeverij BV.
- House, E. R. (1972). The conscience of educational evaluation. *Teachers College Record*. 73(3), 405-414.
- House, E. R. (1978). Assumptions underlying evaluation models. *Education Research*, 7(3), 4-12.
- House, E. R. (1983). Assumptions underlying evaluation models. In G. F. Madaus, M. Scriven, & D. L. Stufflebeam (Eds.). *Evaluation models: Viewpoints on educational and human services evaluation* (pp. 45-64). Boston, MA: Kluwer.
- House, E. R. (1980). *Evaluating with validity*. Beverly Hills, CA: Sage Publications.
- House, E. (1991). Evaluation and social Justice: where are we? In: McLaughlin M.W. & Phillips, D.C. (eds.). *Evaluation and Education: at quarter century (90th Yearbook of the National Society of the*

- Study of Education, part II*. Chicago, Ill.: University of Chicago Press.
- House, E. R. (2004). Intellectual history in evaluation. In M. C. Alkin (Ed.), *Evaluation Roots* (pp. 218-224). Thousand Oaks, CA: Sage.
- Inspectie van het Onderwijs (1991). *Muziek de maat genomen. Rapport over de kwaliteit van het muziekonderwijs op de Nederlandse basisscholen*. Zoetermeer: Inspectie van het Onderwijs.
- Inspectie van het Onderwijs (1992). *Examens op punten getoetst. Onderzoek naar de ontwikkeling van de normen bij de centrale examens in het Voortgezet Onderwijs*. Inspectierapport, november 1992. Zoetermeer: Ministerie van O&W.
- Inspectie van het Onderwijs (2012). *Toezichtkader Voortgezet Onderwijs 2013*. Utrecht: Inspectie van het Onderwijs.
- Janssens, F. J. G. (1985). Betrouwbaarheid en validiteit in interpretatief onderzoek. *Pedagogisch tijdschrift*, 10(3), 149-161.
- Janssens, F.J.G. (1986). *De evaluatiepraktijken van leerkrachten*. Groningen: Rijks Universiteit Groningen (Diss.).
- Janssens, F.J.G. (1991). *Schoolevaluatie: naar een inspectie-aanpak voor het evalueren van scholen*. Zoetermeer: Inspectie van het Onderwijs.
- Janssens, F.J.G. (2005). *Toezicht in discussie: over onderwijstoezicht en educational governance*. Enschede: Universiteit Twente (oratie).
- Janssens, F.J.G. & Zwarts, M.A. (1993). Het gebruik van 'standaarden en normen' binnen de inspectie. *Introspectie*, 14 (1) p. 7-8.
- Janssens, F.J.G., red. (1997). *Van onderzoek naar evaluatie. De methodologie van de onderwijsinspectie*. Den Haag: Delwel uitgeverij BV.
- Janssens, F.J.G. & Van Amelsvoort, H.W.C.H. (2008). School self-evaluation and school improvement in Europe: an exploratory study. *Studies in Educational Evaluation*. 34, 15-23.
- Janssens, F.J.G. & De Wolf, I.F. (2009). Analyzing the assumptions of a policy program: An ex-ante evaluation of Educational Governance in the Netherlands. *American Journal of Evaluation*. 30(3), 411-425.
- Johnson, K. et al. (2009). Research on evaluation use: a review of the empirical literature from 1986 to 2005. *American Journal of Evaluation*, 30, 377-410.
- Karsten, S., Veenstra, A.J., Dijkstra, A.B., & Veenstra, R. (2010). Towards standards for the publication of performance indicators in the public sector: the case of schools. *Public Administration*, 88, 90-112.

- Kelly, G.J. (2006). Epistemology and Educational Research. In: Green, J.L., Camilli, G., Elmore, P.B. et al.: *Handbook of Complementary Methods in Education Research*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.
- Knoops, C.D. (2010). *Verslaggevingstheorieën: een wetenschapsfilosofische analyse*. Rotterdam: Erasmus Universiteit Rotterdam (Diss.)
- Koningsveld, H. (2006). *Het verschijnsel wetenschap*. Amsterdam: Boom (12e druk).
- Leviton, L. & E. Hughes (1981). Research on the utilizations of evaluations: A review and synthesis. *Evaluation*, 5, No. 4, 525 - 548.
- Leeuw, F. L. (2009). Evaluation policy in the Netherlands. In W.M.K. Trochim, M. M. Mark, & L. J. Cooksy, ed. *Evaluation policy and evaluation practice*. New Directions for Evaluation, 123, 87 - 102.
- Lieblich, A., Mashiach-Tuval, R., & Zilber, T. (1998). *Narrative research: Reading, analysis and interpretation*. Thousand Oaks, CA: Sage.
- Linden, W.J. van der & Zwarts, M.A. (1995). De opbrengsten van het basisonderwijs: een replek. *Tijdschrift voor Onderwijsresearch*. 20,1, 34-41.
- Linden, W.J van der & Zwarts, M.A. (1996). Treffers schiet mis. Maar raakt wel zijn eigen voet. *Tijdschrift voor Onderwijsresearch*, 21,1, 106-114.
- Linden, W.J. van der, Meijer, R.R. & Vos, H.J. (1997). *Normeringsmethoden voor inspectie-evaluaties*. Enschede: Vakgroep Onderwijskundige Meetmethoden en Data-analyse, Universiteit Twente.
- Lipsey, M.W. (1993). Theory as Method: Small theories of treatment. *New Directions for Program Evaluation*. 57, 5-38.
- Mark, M., Henry, G., & Julnes, G. (2000). *Evaluation: An integrated framework for understanding, guiding, and improving public and nonprofit policies and programs*. San Francisco: Jossey-Bass.
- Madaus, G.F., Scriven, M. & Stufflebeam, D.L. (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (pp. 229-260). Boston: Kluwer-Nijhoff.
- Mertens, F.J.H. (1982). *Evaluatietheorie en schoolbezoek*. Den Haag: Inspectie van het Onderwijs.
- Mertens, F. (2011). *Inspecteren: toezicht door inspecties*. Den Haag: SDU Uitgevers bv.
- Mertens, D.M. & Wilson, A.T. (2012). *Program Evaluation: Theory and Practice*. New York: The Guilford Press.

- Muijs D & Reynolds D (2005). *Effective Teaching - Introduction & Conclusion*. (2nd ed.) London: Sage Publications.
- Nutley, S., Walter, I. & H. T.O. Davies. (2003). From knowing to doing: a framework for understanding the evidence-into-practice agenda. *Evaluation*, 9 (2): 125 - 149.
- Oakes, J. (1986). *Educational indicators: A guide for policymakers*. New Brunswick: Rutgers University, Center for Policy Research in Education.
- Obbink, H. (2011). De onderwijsinspectie denkt niet langer begripvol mee. *Trouw*, 23 februari 2011 http://www.trouw.nl/tr/nl/5009/Archief/archief/article/detail/1851091/2011/02/23/De-onderwijsinspectie-denkt-niet-langer-begripvol-mee.dhtml?utm_source=scherm1&utm_medium=button&utm_campaign
- Patton, M. Q. (2008). *Utilization Focused Evaluation* (4th ed.). Thousand Oaks: Sage Publications.
- Peschar, J. en C. Meijer (1997). *WSNS op weg. De evaluatie van het 'Weer Samen Naar School'-beleid*. Groningen: Wolters-Noordhoff.
- Robben, P.B.M. (2010). *Toezicht in een glazen huis*. Rotterdam: Erasmus Universiteit (oratie).
- Rosenshine, B. (1970). Evaluation of classroom instruction. *Review of Educational Research*, 40(2), 279-300.
- Rossi, P., Freeman, H., & Wright, S. (1979). *Evaluation: A systematic approach*. Beverly Hills, CA: Sage.
- Rossi, P., Freeman, H., & Lipsey, M. (1999). *Evaluation: A systematic approach* (6th ed.). Thousand Oaks, CA: Sage.
- Sidani, S. & Sechrest, L (1999). Putting program theory into operation. *American Journal of Evaluation*, 20, 227-238.
- Scheerens, J. (1983). *Evaluatie-onderzoek en beleid: methodologische en organisatorische aspecten*. Den Haag: SVO.
- Scheerens, J. (1990). School effectiveness research and the development of process indicators of school functioning. *School effectiveness and school improvement*. 1 (1), p. 61-80.
- Scheerens, J., Glas, C. & Thomas, S.M. (2003). *Educational Evaluation, Assessment and Monitoring. A systematic approach*. Lisse: Swets & Zeitlinger Publishers.
- Scriven, M. J. (1980). *The logic of evaluation*. Inverness, CA: Edgepress.
- Scriven, M. (1983). Evaluation ideologies. In G. F. Madaus, M. Scriven, & D. L. Stufflebeam (Eds.), *Evaluation models: Viewpoints on educational and human*

- services evaluation* (pp. 117-141). Boston: Kluwer-Nijhoff.
- Scriven, M. J. (1991). *Evaluation thesaurus* (4th ed.). Newbury Park, CA: Sage.
- Scriven, M. (2013). Conceptual revolutions in evaluation. In: Alkin, M.C., ed. *Evaluation Roots: a wider perspective of theorists' views and influences*. Thousand Oaks, Cal.: Sage Publications Inc.
- Shadish, W. R., Jr., Cook, T. D., & Leviton, L.C.L. (1991). *Foundations of program evaluation*. Thousand Oaks, CA: Sage.
- Shulman, L.S. (1997). Disciplines of Inquiry in Education: a new overview. In R.M. Jaeger (Ed.). *Complementary methods for research in education*. Washington, D.C: AERA.
- Smeets, G. & Verkroost, J. (2011). *Selectief en slagvaardig: werken met de WOT (2000-2010)*. Utrecht: Inspectie van het Onderwijs.
- Stamm, M. (2003). *Evaluation und Ihre Folgen für die Bildung – eine unterschätzte pädagogische Herausforderung*. Münster: Waxmann.
- Stake, R.E. (1967). The Countenance of Educational Evaluation. *Teachers College Record*. 68, 523-540.
- Stake, R.E. & Denny, T. (1969). Needed concepts and techniques for utilizing more fully the potential of evaluation. In: R.W. Tyler (ed). *Educational Evaluation*. Chicago: National Society for the Study of Education.
- Stake, R. (1975). *Program evaluation, particularly responsive evaluation* (Occasional Paper No. 5). Kalamazoo: Western Michigan University Evaluation Center.
- Stake, R.E. (2010). *Qualitative Research: Studying How Things Work*. New York: The Guildford Press.
- Streumer, J.N. (1988). *Evalueren van techniek*. De Lier: Academisch Boeken Centrum (Diss.).
- Stufflebeam, D. (1983). The CIPP model for program evaluation. In G. F. Madaus, M. S. Scriven, & D. L. Stufflebeam (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (pp. 117-141). Boston: Kluwer-Nijhoff.
- Treffers, A. (1996). Rekeninhoudelijke kritiek op CEB-replik. *Tijdschrift voor Onderwijsresearch*. 21,1, 103-105
- Trotman, D. (2006). Interpreting imaginative lifeworlds: Phenomenological approaches in imagination and the evaluation of educational practice. *Qualitative Research*, 6, 245-265.
- Verschuuren, P.J.M. (2002). *Dogma's en ontwikkelingen in wetenschap en methodologie: bedreigingen en kansen*. Nijmegen: Katholieke Universiteit (oratie).

- Wagner, R. B. (1989). *Accountability in education: A philosophical inquiry*. New York: Routledge.
- Walberg, H.J. & Haertel, G.D. (1990). *The International Encyclopedia of Educational Evaluation*. Oxford: Pergamon Press Inc.
- Weiss, C. (1991). Evaluation research in the political context: Sixteen years and four administrations later. In M. W. McLaughlin & D. C. Phillips (Eds.), *Evaluation and education: At quarter century (90th yearbook of the National Society for the Study of Education, Part II)* (pp. 211-231). Chicago: University of Chicago Press.
- Weiss, C.H. (1997). *Evaluation*, (2nd ed.) New Jersey: Prentice Hall.
- Weiss, C., & Bucuvalas, M. (1980). *Social science research and decision-making*. New York: Russell Sage.
- Wijnstra, J.M (1995). De opbrengsten van het basisonderwijs volgens de CEB: enkele kanttekeningen bij de gevolgde normeringsprocedure. *Tijdschrift voor Onderwijsresearch*. 20,1, 28-33.
- Wilcox, B. (2000). *Making school inspection visits more effective: The English Experience*. Paris: Unesco. <http://unesdoc.unesco.org/images/0012/001202/12028ge.pdf>
- WKKF (2004). *Logic Model Development Guide*. Zie: <http://www.wkkf.org/knowledge-center/resources/2006/02/WK-Kellogg-Foundation-Logic-Model-Development-Guide.aspx>
- Wieringa, J.H. et al. (1984). *Op ander spoor; discussienota algemeen schoolbezoek binnen de inspectie van het Onderwijs*. Den Haag: Inspectie van het Onderwijs.
- Winter, H.B. (2010). *Zicht op Toezicht? Over de meerwaarde van toezicht in de risicosamenleving*. Groningen: Rijksuniversiteit Groningen (oratie).
- World Bank (2004). *Influential evaluations. Evaluations that Improved Performance and Impacts of Development Programs*. Washington: the World Bank.
- World Bank (2010). Supervision of Primary and Secondary Education: A Five-Country Comparison. England, Finland, New Zealand, the Netherlands, and the Republic of Korea. *ECA Knowledge Brief*. Vol. 33. 2010. Zie http://siteresources.worldbank.org/INTECALEA/Resources/ECA_KB_33_School_Supervision_comparative_study.pdf
- Worthen, B.R. & Sanders, J.R. (1987). *Educational evaluation: alternative approaches and practical guidelines*. New York: Longman.