De kwaliteit van (juridische) oordelen

Verslag van een experiment over verantwoording, bias en de kwaliteit van beoordelingen bij toezichthouders en in de rechtspraak

Thomas Schillemans & Ivo Giesen[1]

Er is al veel bekend over toezicht, verantwoording en beoordeling in de publieke sector enerzijds en over menselijk oordeelsvermogen, biases en beoordelingsfouten anderzijds. Maar die kennis is nog maar in zeer beperkte mate onderling verbonden. In dit artikel wordt verslag gedaan van drie verkennende experimenten naar de kwaliteit van beoordelingen door professionele beoordelaars in de publieke sector. De inzet daarvan is om te leren van de gedragswetenschappen in onderzoek naar kwaliteit en effectiviteit van professioneel oordelen. In het onderzoek is een klassiek sociaalpsychologisch experiment over verantwoording, biases en oordeelsvermogen (Tetlock, 1983) vertaald naar de publieke sector en uitgevoerd met achtereenvolgens toezichthouders, studenten en rechters en gerechtssecretarissen.

1. Inleiding en context

1.1. Steeds meer toezicht, steeds meer oordelen

De oude verzorgingsstaat heeft zich de afgelopen decennia getransformeerd tot een *reguleringsstaat* (Majone, 1994). Een relatief steeds groter deel van het werk van overheden bestaat uit het normeren, reguleren, toetsen en beoordelen van activiteiten die door andere actoren in private, publieke en semipublieke contexten worden uitgevoerd. In de publieke sector bestaat een keur aan instanties wier taak het is om toe te zien op, en oordelen te vellen over, het werk en handelen van organisaties of individuen. Zo zijn er bijvoorbeeld negen inspecties aangesloten bij de Inspectieraad en er zijn zeven markttoezichthouders aangesloten bij het Markttoezichthoudersberaad (MTB). In zorg, onderwijs en wonen zijn er certificaties en visitaties als ook certificerende instellingen en visitatiebureaus. Er is ook weer een Raad voor accreditatie die certificaties beoordeelt en er zijn instanties die visitatiesystemen beoordelen. Ook binnen semipublieke organisaties zijn toezicht en beoordeling cruciaal en prevalent. Er zijn volgens schatting van de WRR (2014) zo'n 4000 tot 7000 raden van toezicht en raden van commissarissen in de semipublieke sector waarin plaats is voor 24.000 tot 47.000 toezichthouders. Als het over (be)oordelen van anderen gaat, komt natuurlijk ook nog de rechterlijke macht in beeld. Volgens telling van de Raad voor de rechtspraak (2018) waren er eind 2017 bijna 2500 rechters.

De verscheidenheid tussen deze instanties en professionals is enorm. Het toezicht op het beleid van een woningcorporatie verloopt volgens andere procedures, naar andere normen en door andersoortige professionals dan het toezien op het College van Bestuur van een universiteit of het oordelen over geschillen door een rechter. Het werk van de individuen die uiteenlopende beoordelende rollen vervullen is gereguleerd binnen verschillende organisaties en naar uiteenlopende professionele codes en mores. Desalniettemin, bij alle relevante verscheidenheid, hebben al die instanties en personen gemeen dat ze goed moeten oordelen over het werk of handelen van anderen en daarmee dat de kwaliteit van hun individueel oordeelsvermogen cruciaal is.

1.2. Het probleem: een feilbaar brein

En wat is nu het probleem? Het probleem is niet – of hier niet, althans – dat het er zo veel zijn. Het probleem is,

Onderzoek laat zien dat mensen de meeste beslissingen nemen met zo min mogelijk cognitieve inspanning

schertsend gesteld, dat die professionals, inclusief de rechters, ook allemaal *mensen* zijn. Want het oordeelsvermogen van mensen is, zo leren de gedragswetenschappen, feilbaar en kwetsbaar (o.a. Kahneman e.a., 1982; Baron, 1998; Pohl (ed.), 2004; Baron, 2008). Het menselijk brein is, in de woorden van Kahneman (2011) *a lazy machine jumping to conclusions*. Onderzoek laat zien dat mensen de meeste beslissingen nemen met zo min mogelijk cognitieve inspanning. Dat heeft als voordeel dat we relatief veel dingen kunnen doen, dat we kunnen multi-tasken of op zijn minst schakel-serieel-tasken. Maar het heeft wel als nadeel dat we geneigd zijn om olifantenpaadjes te bewandelen in onze oordeelsvorming en veel informatie overslaan of beperkt waarnemen. Niet voor niets was 'expect error' een van de principes die Thaler & Sunstein (2008) formuleerden over hoe mensen beslissingen nemen. Vertaald naar de context van toezicht en beoordeling is dat potentieel zorgwekkend: komt de beoordelaar die ook maar mens is, wel tot een correct oordeel?

De gedragswetenschappen leren dat wij ons laten leiden door allerhande biases. Zo is er de confirmation bias (Giesen, 2015, nr. 253-255): nieuwe informatie bevestigt vaak wat we toch al vermoedden. Denk aan een nieuwe toezichthouder die met groot zelfvertrouwen stelt: 'Ik ken de problematiek bij dit soort organisaties wel zo'n beetje' (Schillemans, 2007). Er zijn daarnaast nog veel andere biases die zijn besproken in relatie tot bijvoorbeeld toezicht (Ottow, 2015), het recht (o.a. Van Boom et al., 2013; Giesen, 2015, nr. 482 e.v.; Van Koppen, (ed.) 2017), en auditing (Cushing & Alawat, 1996; Peecher & Piercey, 2008)

We weten veel over toezicht, verantwoording en beoordeling in de publieke sector enerzijds en veel over menselijk oordeelsvermogen, biases en beoordelingsfouten anderzijds. Maar die kennis is nog maar in zeer beperkte mate onderling verbonden. Dat komt mede doordat heel veel van het belangwekkende onderzoek naar oordelen en biases in de Verenigde Staten is uitgevoerd met studenten die onder strenge laboratoriumcondities opdrachten uitvoeren met soms een heel laag realistisch gehalte (Aleksovska et al., 2019). Dat levert prikkelende en soms ook zorgwekkende conclusies op over ons oordeelsvermogen die relevant lijken voor de professionele praktijk. Alleen is de vraag wel, doen dergelijke effecten zich ook voor buiten het lab, bij echte professionals en bij echte beoordelingsvragen? Specifieker: gelden dergelijke effecten ook voor bijvoorbeeld toezichthouders en rechters? Het onderzoek daarnaar is vooralsnog schaarser (zie echter Ten Velden & De Dreu 2012; Rachlinski, 2012). Het is voorstelbaar dat beoordelaars met een specifieke en gerichte opleiding en soms vele jaren van ervaring anders (hopelijk beter!) zullen oordelen en anders omgaan met beoordelingsbiases dan studenten.

1.3. Drie verkennende experimenten

Tegen die achtergrond doen wij hier verslag van drie verkennende experimenten naar de kwaliteit van beoordelingen door professionele beoordelaars in de publieke sector. De inzet daarvan is om te leren van de gedragswetenschappen in onderzoek naar kwaliteit en effectiviteit van professioneel oordelen. Het betreft een experiment in drie ronden waaraan 171 professionele 'beoordelaars' en 77 studenten hebben deelgenomen. We hebben een klassiek sociaalpsychologisch experiment over verantwoording, biases en oordeelsvermogen (Tetlock, 1983) vertaald naar de publieke sector en uitgevoerd met achtereenvolgens toezichthouders (86), studenten (77) en rechters en gerechtssecretarissen (totaal 85). Het originele onderzoek van Tetlock liet zien dat mensen (studenten) snel biases in hun oordelen hebben hetgeen met verantwoording kan worden gecorrigeerd, waardoor de kwaliteit van hun beoordelingen omhoog gaat. In ons onderzoek hebben we twee vragen gesteld. In de eerste plaats de vraag of dezelfde twee effecten (biases en kwaliteit van oordelen respectievelijk het positieve effect van verantwoording) ook te vinden zijn bij professionele beoordelaars in meer realistische settings. En in de tweede plaats de vraag hoe de oordelen van professionele beoordelaars zich verhouden tot die van studenten. Antwoord op deze vragen draagt bij aan onze inzichten in kwaliteiten en risico's van professioneel oordelen en biedt handvatten voor de toezichts- en gerechtelijke praktijk.

2. Het experiment

2.1. Een klassiek experiment driemaal herhaald

In het experiment van Tetlock (1983) stond een moordzaak centraal, waarin ene 'mister Grey' de verdachte was en er achttien verschillende stukjes bewijs waren, waarvan de helft belastend en de andere helft ontlastend. Deelnemende studenten werd verteld dat zij de jury waren die moest oordelen over de schuld van mister Grey en zij deden dat onder verschillende condities van verantwoording. Sommige deelnemers waren volkomen anoniem terwijl anderen werd verteld dat hun oordeel zou worden bekeken door een expert, waardoor zij onder een conditie van verantwoording oordeelden.

Auteurs

1. Prof. dr. T. Schillemans is als hoogleraar bestuur en beleid (met een focus op verantwoording, gedrag en instituties) verbonden aan het Departement Bestuurs- en Organisatiewetenschap van de Universiteit Utrecht. Prof. dr. I. Giesen is als hoogleraar Privaatrecht verbonden aan Ucall en aan het Montaigne Centrum van het Departement Rechtsgeleerdheid van de Universiteit Utrecht. Beiden zijn tevens verbonden aan de stream 'Institutions & Behaviour' van het Utrechtse Strategisch thema 'Instituties voor Open Samenlevingen'. Dit onderzoek is uitgevoerd als onderdeel van het NWO Vidi project 'Calibrating Public Accountability'.

Zie daarvoor https://accountablegovernance.sites.uu.nl/. Dit onderzoek is mogelijk gemaakt doordat zes verschillende organisaties bereid waren mee te werken en 248 mensen bereid zijn geweest om tijd te nemen om serieus aan dit experiment deel te nemen. Onze dank aan de organisaties, de contactpersonen en individuele deelnemers is dan ook groot. Dank gaat ook uit naar Maj Jeppesen en Marija Aleksovska voor assistentie met Qualtrics, naar prof. mr. Maarten Feteris die de eerste auteur op het spoor zette om ook eens bij rechters te gaan neuzen, waarna de tweede auteur betrokken werd, en aan Eddy Bauw voor het kritisch meelezen.

In ons experiment vertaalden we deze opzet naar een realistische casus voor de publieke sector. Nu was er geen persoon overleden maar was een semipublieke instelling in Noordrijn-Westfalen failliet. Er waren negen brokken informatie die suggereerden dat de bestuurder zich zou hebben schuldig gemaakt aan wanbestuur en er waren negen brokken informatie die suggereerden dat hij juist een heel goede bestuurder was. De deelnemers werd verteld dat de regering van Noordrijn-Westfalen hen als onafhankelijke maar ervaren Nederlandse beoordelaars had gevraagd om de beschikbare informatie te lezen en daarover een oordeel te vellen. De informatie was kort beschikbaar en deelnemers konden geen aantekeningen maken, zodat zij op hun geheugen moesten vertrouwen en de aandacht erbij moesten houden. Dat wijkt weliswaar af van de gebruikelijke gang van zaken, maar modelleert wel in geconcentreerde vorm de praktijk waarin een overdaad van informatie onder tijdsdruk moet worden verwerkt en waarin beoordelaars het geheugen gebruiken in hun beoordelingsproces.

Na lezing van de informatie werd deelnemers eerst gevraagd om een cijfermatig oordeel over de waarschijnlijkheid van wanbestuur te geven, op een schaal van 0-100. Daarna werd hen gevraagd dit te onderbouwen met alle informatie die zij zich konden herinneren.

Het experiment is in drie onderling verbonden maar aparte deel-experimenten uitgevoerd: eerst met toezichthouders, vervolgens met studenten en uiteindelijk in een versimpelde (en 'scherpere') vorm met rechters en gerechtelijk secretarissen (werkzaam op alle rechtsgebieden). Steeds was het experiment de start van een workshop of college over professioneel oordelen. Zo was het derde experiment een onderdeel van een tweemaal door ons verzorgde workshop voor Nederlandse rechters en gerechtssecretarissen welke beoogde om de rechterlijke macht meer inzicht te bieden in de werking van heuristieken en biases op de oordeelsvorming. De informatieoverdracht daarover werd gekoppeld aan het voormelde experiment. We konden zo de werking van enkele biases ter plekke demonstreren; de workshop leverde daarmee tevens extra data op, gerelateerd aan een nog niet onderzochte doelgroep, die we konden benutten om de eerder verkregen resultaten te verifiëren.[2]

Het derde experiment was, als gezegd, vereenvoudigd om de manipulatie van verantwoording te versterken. In de eerste twee experimenten was de verantwoordingsinstructie opgenomen bij alle andere informatie die de deelnemers ontvingen. Ook na mondelinge evaluatie werd duidelijk dat mensen die (terecht!) niet serieus namen en dus niet het gevoel hadden dat ze hun oordeel moesten verantwoorden. Ook bleek bij de debriefing dat deelnemers regelmatig hadden gemist of ze wel of niet in een verantwoordingssetting moesten opereren. Anders dan bij Tetlock (1983) had de manipulatie dan ook geen effect. In het derde experiment met rechters werd de verantwoording daarom realistischer gemanipuleerd. De helft van de deelnemers aan de workshops werd gemeld dat hun antwoorden input voor de rest van de workshop vormden en dat zij tijdens de workshop door hun collega's maar ook door de twee aanwezige hoogleraren zouden kunnen worden gevraagd om hun antwoord toe te lichten. Dit had effect op hoe zij tot hun oordeel kwamen, zo zal straks (par. 5) duidelijk worden. Hierbij merken wij op dat replicatie van experimenten, zoals hier gebeurd is, van belang is voor de betrouwbaarheid van de bevindingen. De meeste relaties tussen de gemeten variabelen blijken, zo wordt hierna duidelijk, vergelijkbaar tussen deze drie

Deel-experiment 1: toezichthouders		**Deel-experiment 2: studenten**		**Deel-experiment 3: rechters en gerechtssecretarissen**	
N	**Datum**	**N**	**Datum**	**N**	**Datum**
18	8 juni 2017	46	7 maart 2018 (4 groepen)	21	8 november 2018
11	18 mei 2017				
13	20 april 2017	31	6 maart 2018 (2 groepen)	64	8 oktober 2018
33	21 maart 2017				
11	7 maart 2017				
86		77		85	
totaal aantal unieke deelnemers: 248					

Tabel 1: Data en omvang groepen deel-experimenten

deel-experimenten en ook met het originele onderzoek. Dat duidt op stabiele relaties tussen verantwoording, biases en oordeelskwaliteit.

De tabel (1) hiervoor geeft inzicht in de deelnemers aan de drie deel-experimenten.

2.2. De casus

De deelnemers bogen zich over een casus over een failliete semipublieke dienstverlener. Enerzijds ontvingen zij negen statements over de organisatie die suggereerden dat er sprake zou zijn van wanbestuur door de bestuurder. Zo lazen ze dat de administratie rommelig was, dat de bestuurder slecht tegen tegenspraak kon en dat een waarschuwing van de externe accountant in de wind was geslagen. Aan de andere kant was er ook informatie die juist heel positief was over de bestuurder. Zo liepen de clienten van de organisatie met hem weg, werden grote besluiten intern aan anderen voorgelegd en was hij genomineerd als 'publieke manager van het jaar'. Numeriek hield de belastende en ontlastende informatie elkaar in evenwicht. De 248 deelnemers vonden echter gemiddeld dat de balans meer naar 'schuldig' dan naar 'onschuldig' uitsloeg, waarbij de groepen niet significant van elkaar verschilden met gemiddelden van 60,6 tot 68,6. De tabel (2) hieronder geeft de gemiddelden en de spreiding weer.

	Exp 1	Exp 2	Exp 3
N	86	77	85
Gemiddelde	67,88	66,27	60,59
Std. Dev.	17,02	13,80	19,40
Range	97,00	72,00	92,00
Minimum	3,00	26,00	8,00
Maximum	100,00	98,00	100,00

Tabel 2: Gemiddelde oordelen en spreiding bij de deelexperimenten

Klaarblijkelijk was de belastende informatie meer overtuigend, al kan daar ook een element van hindsight bias inzitten: de deelnemers wisten immers al dat de organisatie failliet is.

2.3. Beoordelingscontext: wel of geen verantwoording

In het experiment beoordeelden de deelnemers de casus onder verschillende condities van verantwoording. Steeds werd een random deel van de deelnemers verteld dat ze volkomen anoniem zou kunnen oordelen. En een ander deel van de deelnemers werd verteld dat hun oordeel zou worden beoordeeld door een ander (eerste twee deelexperimenten) of dat hen gevraagd zou kunnen worden hun oordeel toe te lichten aan de aanwezige peers en hoogleraren binnen de workshop.

De gedachte vooraf was dat de (verantwoordings)context invloed zou hebben op de kwaliteit van oordeelsvorming. In psychologisch onderzoek naar besluitvorming komt immers naar voren dat mensen snel voor *low effort strategies* kiezen (Aleksosovska et al., 2019) en taken als het ware 'op de automatische piloot' (volgens systeem 1, zou Kahneman zeggen), uitvoeren. De condities waarin iemand beslist of oordeelt, hebben effect op dat oordeel; zo suggereert ook de grote hype die naar aanleiding van het boek 'Nudge' is ontstaan (Thaler & Sunstein, 2008; Feitsma, 2016). Een conditie van anonimiteit brengt normaliter niet het beste in mensen naar boven en zou dus kunnen leiden tot relatief minder goede oordelen. Omgekeerd zou mogen worden verwacht dat beoordelaars die wel verwachten dat ze zich moeten verantwoorden, een grotere inspanning plegen en tot beter doordachte en meer precieze oordelen komen.

Een minimale maat om beoordelaars mee te beoordelen is of ze eigenlijk hun best doen

Kortom, de beoordelingscondities kunnen effect hebben op de kwaliteit van het oordeel. Dit roept echter nog wel de vraag op: wat is eigenlijk beoordelingskwaliteit?

3. Beoordelingskwaliteit

In deze casus is het niet mogelijk te bepalen wat het juiste oordeel is over de mate van wanbestuur van de bestuurder. Antwoorden varieerden letterlijk van bijna 0 (totaal onschuldig) naar 100 (volkomen schuldig). Vanuit het onderzoek is er geen objectieve maat voor het juiste antwoord. Tegelijk zijn er wel – in elk geval los van de rechterlijke context – procesmaatstaven bekend waarmee procedureel 'betere' besluiten kunnen worden onderscheiden van mindere besluiten. In dit onderzoek hanteren we drie van die procedurele maatstaven voor de kwaliteit van beoordelingen, welke vervolgens vaak ook sterk onderling bleken samen te hangen.

Beoordelingsinzet: Een minimale maat om beoordelaars mee te beoordelen is of ze eigenlijk hun best doen. Spannen ze zich in om een casus te beoordelen? In veel psychologisch beoordelingsonderzoek wordt de tijd die iemand aan een taak besteedt gebruikt als betrouwbare maat voor de inzet die iemand pleegt, want op groepsniveau is het waarschijnlijk dat langzamere experimentele groepen meer aandacht besteden aan de opdracht dan snellere groepen. Het is dan ook waarschijnlijker dat zij zich meer inspannen en een groter beroep doen op het reflectieve, systeem II-denken (van Kahneman). In psychologisch onderzoek komt vaak een sterk, positief verband naar voren tussen de tijd die personen besteden aan een taak en de conditie van verantwoording (Aleksovka et al., 2019). In de eerste twee deel-experimenten konden we meten hoe lang deelnemers over hun beoordeling deden. In het derde deel-experiment was dat niet mogelijk, maar was wel vast te leggen hoeveel woorden zij gebruikten om hun oordeel toe te lichten. Meer woorden gebruiken, staat daarbij gemiddeld genomen gelijk aan meer inzet want dat kost extra tijd.

Noten

2. De deelnemende rechters en secretarissen selecteerden zichzelf door hun opgave voor de workshops, de onderzoekers hebben daar niet de hand in gehad.

Zorgvuldigheid. De deelnemers was gevraagd om het cijfermatige oordeel ook inhoudelijk te onderbouwen aan de hand van de verkregen informatie. Op het moment van rapporteren was de informatie zelf niet meer beschikbaar, ze moesten dus putten uit hun geheugen. De antwoorden zijn vervolgens gecodeerd door twee codeurs. Zij telden hoeveel van de achttien stukjes informatie correct werden herhaald maar ook hoeveel kleine slordigheden en echte fouten de deelnemers maakten. In dit geval is de lat hoog gelegd: een goed oordeel moest helemaal foutloos zijn. Zeggen dat de bestuurder de prijs van overheidsmanager kreeg was bijvoorbeeld gecodeerd als onzorgvuldig omdat hij alleen genomineerd was. Hiermee konden we vaststellen hoe zorgvuldig en volledig de deelnemers de informatie hadden verwerkt en meegenomen in hun oordeel. Hieronder volgen de bevindingen over het maken van fouten door deelnemers in de drie deelexperimenten. De tabel (3) laat zien dat de professionele deelnemers gemiddeld nog geen halve fout in totaal maakten terwijl de studenten gemiddeld bijna een hele fout per persoon maakten. De rechters deden het net iets beter dan de toezichthouders.

	Exp 1	Exp 2	Exp 3
N	85	77	83
Gemiddelde	0,4235	0,8182	0,3900
Std. Deviatie	0,777	0,899	0,678
Range	4	4	3
Minimum	0	0	0
Maximum	4	4	3

Tabel 3: Gemiddeld aantal fouten per persoon per deelexperiment

Cognitieve complexiteit. Een belangrijkere, vaak gebruikte, maar ook ingewikkelder, maat voor de kwaliteit van beoordelingen is cognitieve complexiteit (Scott, 1962; Schillemans, 2016). Een cognitief complex besluit wordt gekenmerkt doordat veel informatie wordt verwerkt, geanalyseerd (differentiatie) en vervolgens weer onderling gerelateerd (integratie). De gedachte is dat een hoger geïnformeerd en beter doordacht besluit als zodanig 'beter' is dan een matig geïnformeerd en slecht doordacht besluit, nog los van de eventuele uitkomst of gevolgen. De mate van cognitieve complexiteit van besluiten kan worden onderzocht aan de hand van vragen over informatieverwerking en -afweging (Suedfeld & Tetlock, 1977; Doney & Armstrong, 1996). Ten aanzien van *differentiatie* is gekeken naar het aantal argumenten dat deelnemers correct memoreren in hun toelichting. Hoeveel van de achttien stukjes informatie konden deelnemers correct resumeren? Vervolgens is voor wat betreft de integratie ook gekeken naar de redeneerstijl. De deelnemers aan de eerste twee experimenten bleken hun toelichting op twee heel verschillende manieren te schrijven. Sommige deelnemers maakten een ongesorteerd lijstje met argumenten in bullets, zonder onderlinge verbinding. Anderen schreven een meer samenhangend betoog waarin ze de argumenten ook samenbrachten en onderling relateerden. Die tweede manier van werken getuigt van integratie van informatie en duidt op een hogere mate van cognitieve complexiteit van beoordelingen.[3]

De drie verschillende maten voor beoordelingskwaliteit liggen theoretisch in elkaars verlengde. Als je beter je best doet, is de kans groter dat je je meer feiten kunt herinneren, dat je minder fouten maakt en dat je de informatie beter onderling afweegt. Dat zou impliceren dat deze maten ook onderling verbonden zouden moeten zijn waardoor een robuuste maat voor beoordelingskwaliteit ontstaat. Analyse laat zien dat dat inderdaad in de verschillende deel-experimenten het geval is, hoewel dat niet op alle punten significant is.

Uit de eerste twee deelexperimenten komt het onderstaande beeld naar voren. De tijd die mensen besteedden aan het experiment bleek sterk samen te hangen met het aantal correct gememoreerde feiten over de casus en een meer argumentatieve redeneerstijl. Daarmee ontstaat een robuuste maat voor beoordelingskwaliteit waarmee kan worden gesteld dat sommige respondenten de beoordelingsopdracht nadrukkelijk 'beter' deden dan andere. Onze analyse (voor de statistiek, zie tabel (4) hieronder) laat een sterke correlatie zien tussen beoordelingsinzet, aantallen correct gememoreerde argumenten en redeneerstijl. Dat betekent dat in de antwoorden van respondenten drie van de vier gebruikte maten van beoor-

De respondenten die een grotere beoordelingsinzet pleegden differentieerden ook meer relevante argumenten én hadden een meer integrerende redeneerstijl

			1. Beoordelings-inzet (duur)	2. Zorgvuldigheid: vermijden fouten	3. Cognitieve complexiteit	
					a. Differentiatie (aantallen argumenten)	b. Integratie (redeneer stijl)
Spearman's rho	**Beoordelings-inzet (duur)**	**Correlatie Coëfficiënt**	1,000	-,123	**,299****	**,336****
		Sig. (2-tailed)	.	,120	,000	,000
		N	161	161	161	161

Tabel 4: De samenhang tussen aspecten van oordeelskwaliteit

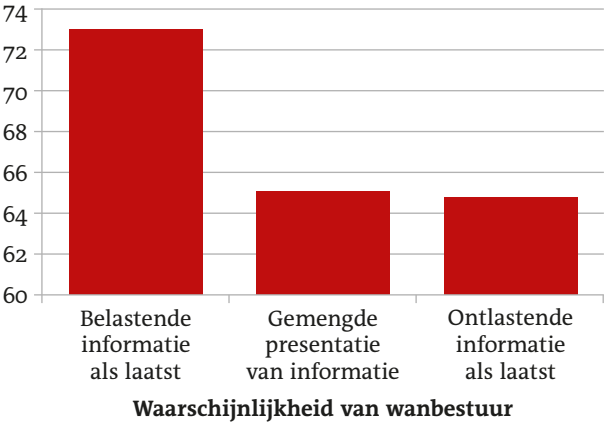
delingskwaliteit significant samenhingen. Ofwel, in woorden: de respondenten die een grotere beoordelingsinzet pleegden (criterium 1), differentieerden ook meer relevante argumenten (criterium 3a) én hadden een meer integrerende redeneerstijl (criterium 3b). Daarmee durven wij de stelling aan dat zij ook een hogere kwaliteit van oordelen hadden.

4. Recency bias

Een andersoortige maat voor kwaliteit van beoordelen is het (kunnen) weerstaan van vertekening door biases in de beoordeling. Eerder onderzoek, waaronder dat van Tetlock (1983), heeft laten zien dat de volgorde waarin mensen informatie tot zich nemen, invloed heeft op hun oordelen. Dat is deels onvermijdelijk. Je kunt nu eenmaal niet alles tegelijk lezen en moet ergens beginnen, dus dit soort effecten treedt al snel op. Ook is het vaak automatisch in werkprocessen ingebouwd. Docenten lezen tentamens of papers die op een stapel liggen en er is altijd het risico (of de kans!) voor de student dat de positie in de stapel invloed heeft op de beoordeling. Hetzelfde zien we bij strafzaken waarin agenten beginnen met een redelijk vermoeden dat iemand schuldig is en zo al met de belastende informatie starten. Maar ook in een rechtszaak ten overstaan van de rechter(s), bijvoorbeeld ten tijde van een strafzitting, een mondelinge behandeling of een civiel pleidooi, is er altijd iemand die als eerste aan de beurt is om het woord te doen (Giesen 2015, nr. 486; Van der Post & Van Toor, 2019). Daarbij komt nog dat degene die pleit, zelf kan bepalen welke informatie hij of zij als eerste of juist als laatste onderdeel van het eigen betoog naar voren brengt (Gravett 2018, met name p. 184-186). Beide elementen kunnen vooringenomenheid en tunnelvisie in de hand werken.

De literatuur (o.a. Enneking e.a., 2013; Van der Post & Van Toor, 2019) onderscheidt zowel het primacy effect (datgene wat als eerste binnenkomt heeft een groot effect) als ook het recency effect (datgene wat als laatste is gelezen heeft de grootste invloed). Het primacy effect speelt vooral op langere termijn, het recency effect op korte termijn. Om dit effect op de beoordeling te kunnen onderzoeken, was de volgorde waarin de informatie over de casus werd gepresenteerd in onze experimenten random gevarieerd. Sommige deelnemers ontvingen als eerste alle belastende informatie, sommige deelnemers juist eerst alle ontlastende informatie en vervolgens was er (in de eerste twee deelexperimenten) een middengroep die de informatie gemengd ontving. Op die manier is mogelijk te zien of de volgorde van de informatievoorziening invloed heeft op de beoordelingen van de casus.

Analyse liet vervolgens zien dat er bij de eerste twee groepen inderdaad sprake was van een vertekening door biases.[4] Meer precies werd het recency effect zichtbaar: de deelnemers die als laatste belastende informatie ontvingen, kwamen wat strenger uit in hun gemiddelde oordeel dan de deelnemers die als laatste ontlastende informatie ontvingen. De opzet en manipulatie was voor de deelnemers duidelijk, zo bleek ook in de workshops; men had natuurlijk meteen door dat de informatie in een bepaalde volgorde was gepresenteerd. Desondanks bleek dat er een significante correlatie was tussen de volgorde van informatie en het uiteindelijke oordeel, waardoor een bias zichtbaar wordt. Nadere analyse bevestigde dit verband[5] en laat zien dat de deelnemers aan de eerste twee deelexperimenten zich lieten beïnvloeden door de volgorde waarin de informatie werd gepresenteerd. Dat gold dus voor toezichthouders in vergelijkbare mate als voor studenten.



Figuur 1: Recency-effect bij toezichthouders (exp. 1)

Deze uitkomst stemt tot nadenken voor professionele beoordelaars, zoals rechters. De manier waarop wij onze juridische procedures hebben ingericht of zouden kunnen inrichten, of het nu om strafrecht, civiel recht of bestuursrecht gaat, kan blijkbaar van invloed zijn op de beoordeling, omdat die inrichting ook bepalend is voor de volgorde waarin informatie binnenkomt. Hoor en wederhoor is een groot goed, en in procedures wordt daarom sterk bewaakt dat elke procesdeelnemer aan bod komt, maar dat beginsel laat onverlet dat die procesdeelnemers in een bepaalde volgorde aan de beurt komen. Meervoudig in plaats van enkelvoudig beslissen is een instrument dat benut kan worden om bepaalde biases (o.a. tunnelvisie) te corrigeren, maar als elk van de rechters onder invloed van eenzelfde bias (volgorde-effecten) staat, is dat probleem misschien niet zonder meer van de baan door op ruimere schaal meervoudige kamers in te richten. Toezichthouders en de rechterlijke macht zouden zich ten minste van dit alles bewust moeten zijn (want hier liggen valkuilen); de advocatuur net zozeer (want hier liggen kansen, o.a. Gravett, 2018). De wetgever zou er waarschijnlijk goed aan doen om dit soort kennis mee te nemen bij het ontwerpen of herzien van een procedurevorm (zoals momenteel bij de modernisering van het Wetboek van Strafvordering). Is het bijvoorbeeld nodig dat eerst het OM de zaak van A tot Z aanbrengt en onderbouwt, en dat daarna de verdediging aan de beurt komt, of is het ook een optie om de inbreng van beide zijden iets meer te vermengen? Daarbij is het ook goed om te kijken naar passende vormen van professionele verantwoording, want dat blijkt

3. Voor de zekerheid: de validiteit van het door de respondenten opgebouwde betoog is niet nader gescreend.

4. Bij de derde groep kon dit niet zichtbaar worden door een andere opzet: de informatie moest op papier worden aangeleverd in het derde experiment terwijl het de eerste keren digitaal kon. Digitaal kon de eerst gegeven informatie 'verdwijnen', op papier lukt dat niet.

5. De one-way ANOVA – dat is een variantieanalyse waarmee in experimenteel onderzoek vaak verschillen tussen meerdere groepen worden geanalyseerd – werkte hier want: $F = 6{,}446$, $p = {,}013$.

			Zorgvuldigheid: vermijden fouten	Differentiatie (aantallen argumenten)
Spearman's rho	Verantwoording (1) / niet-verantwoording (0) (duur)	Correlatie Coëfficiënt	-,257*	,340**
		Sig. (2-tailed)	,019	,001
		N	83	85

Tabel 5: de relatie tussen verantwoording en oordeelskwaliteit

inderdaad een positief effect op oordeelskwaliteit te hebben, zoals we hierna toelichten. Beknibbelen op de rechterlijke motiveringsplicht zou bijvoorbeeld uit den boze moeten zijn; sterker, die motiveringsplicht zou een nadere versterking door meer aandacht en een concretere invulling ervan verdienen.

5. Verantwoording verbetert beoordelingen

De analyse hiervoor heeft laten zien dat er significante verschillen zichtbaar zijn in de beoordelingskwaliteit van verschillende deelnemers (par. 3) en ook dat er sprake is van een bias, in het bijzonder het recency-effect (par. 4). De verwachting vooraf was dat deelnemers onder condities van verantwoording tot betere beoordelingen zouden komen, geïnspireerd door het oorspronkelijke experiment van Tetlock (1983). Zoals vermeld werkte de manipulatie in de twee deelexperimenten niet maar in het derde experiment met rechters en gerechtelijk secretarissen werkte deze wel.

In de vereenvoudigde opzet was er inderdaad een sterkere en meer realistische manipulatie van verantwoording voor een deel van de deelnemers doordat zij ter plekke verantwoording zouden kunnen moeten afleggen ten overstaan van hun collega's en de aanwezige hoogleraren. Onze hypothese werd vervolgens bevestigd: de geanticipeerde verantwoording leidde inderdaad tot oordelen van hogere kwaliteit, zowel in termen van het resumeren van meer correcte argumenten als ook in het vermijden van fouten.[6] Hierboven in Tabel 5 zijn de correlaties weergegeven.

Net als in de eerste twee deelexperimenten zien we een samenhang tussen verschillende aspecten van oordeelskwaliteit, die ook samenhangen met beoordelingsinzet in termen van tijd.

De manier waarop wij onze juridische procedures hebben ingericht of zouden kunnen inrichten kan blijkbaar van invloed zijn op de beoordeling

Kort gezegd, de 41 respondenten die verwachtten dat zij hun oordeel in het openbaar moeten toelichten ten overstaan van hun peers en externe experts, komen tot betere beoordelingen met minder fouten dan hun 44 collega's die meenden dat ze anoniem oordelen. Dit effect is conform de verwachting en is van acuut belang voor beoordelaars in de publieke sector. Vreemde ogen dwingen, zo suggereert dit resultaat. Dat is een bevinding die in veel politiek-bestuurlijke contexten natuurlijk wel is verdisconteerd in checks and balances, tegenlezers, overleg, het vier-ogen principe of in andere vormen. Het ligt dan ook in de rede dat meervoudige kamers tot kwalitatief betere beslissingen komen (Bauw e.a., 2013), want daarbij wordt er meteen onderling verantwoording afgelegd. Tegelijk kunnen wij ons voorstellen dat in de dagelijkse, routinematige gang van zaken dergelijke checks and balances kunnen verzwakken en dat sommige beoordelingen effectief door individuen of vaste groepen kunnen worden gemaakt, zonder veel tegenmacht. Ons onderzoek onderstreept dus nogmaals het belang van 'vreemde ogen' bij het nemen van complexe besluiten, waartoe de ogen van de 'peers' zeker ook behoren, zoals is gebleken. In de juridische context onderstreept deze bevinding het belang dat (terecht) gehecht wordt aan en goede, gedegen motivering van een rechterlijke uitspraak. Misschien moeten we dat belang echter nog scherper over het voetlicht brengen (Vgl. Giesen 2015, nr. 495). Dat in het strafrecht een zogenoemd verkort vonnis mogelijk is (artikel 365a Sv; kritisch Mevis, 2001), hetgeen wil zeggen dat er geen uitgewerkte motivering van een uitspraak volgt tenzij er hoger beroep wordt ingesteld, laat zich vanuit dit perspectief eigenlijk lastig denken (maar die verbazing komt wel van een bestuurskundige en een civilist, die uiteraard ook inzien dat de belasting van de rechterlijke macht hier van groot belang is). Uit de professionele standaarden voor strafrechters volgt overigens dat een motivering standaard is (zie Standaard 2.8 onder 1) maar wat die motivering ('op een wijze die recht doet aan de zaak') dan precies behelst en omvat, is lastig te duiden door de ruime formulering van die standaard.[7]

6. Overeenkomsten en verschillen tussen professionals en studenten

6.1. Inleiding: van 1983 naar 2019

De conclusies bevestigen tot zover grotendeels die van Tetlock uit 1983. Zijn we dan iets opgeschoten? We menen van wel. Allereerst is van belang te constateren dat patronen die eerder zijn gevonden met studenten in onrealistische settings nu worden bevestigd in onderzoek in een meer realistische setting en ook met professionele beoordelaars die een voor hen vaak herkenbaar soort van beoordeling moeten uitvoeren, zelfs al betreft dit experiment dan geen 'echt' vonnis.[8] Het laat zien dat ook ervaren beoordelaars vatbaar zijn voor beoordelingsbiases en dat ook bij hen geldt dat de anticipatie op verantwoording tot betere beoordelingen leidt. Dit is een belangrijke conclusie voor beoordelaars in de publieke sector, inclusief de rechterlijke macht. Het is ook een extra aanmoediging om onderzoek

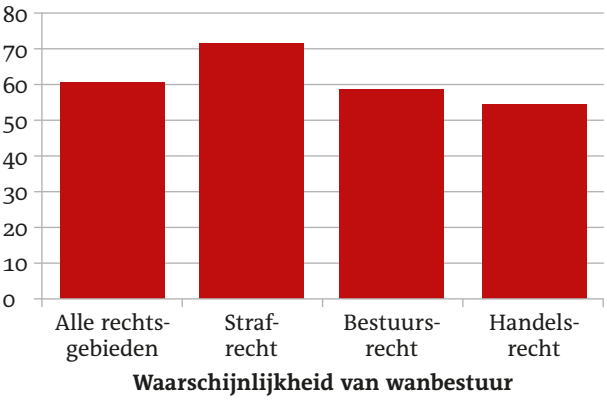
De verschillen in patronen suggereren dat verschillende groepen andere redeneerpatronen kunnen hanteren, hetgeen de relevantie van professionele socialisatie ook in relatie tot oordeelsvorming onderstreept

naar boordelingen onder studentenpopulaties serieus te nemen voor de professionele praktijk en te verkennen wat zorgwekkende of hoopgevende resultaten uit dit soort onderzoek zouden kunnen betekenen voor die praktijk. Dat gezegd hebbende zijn er ook enkele verschillen tussen de groepen die de moeite waard zijn om bij stil te staan.

6.2. Professioneel-afwijkende oordeelspatronen

Individuele rechters en toezichthouders moeten zelf tot een oordeel komen op grond van geldende wetten, regels en normen. Tegelijk maken ze onderdeel uit van specifieke organisaties en organisatieonderdelen maar ook van professionele gemeenschappen, waarbinnen gedeelde praktijken en professionele mores ontstaan. Binnen professionele groepen is sprake van socialisatiepraktijken waarbinnen professionals elkaar beïnvloeden en waarin ze elkaar leren hoe te kijken en te interpreteren (Schillemans, 2016). Verwacht kan daarom worden dat de oordelen van professionals uit dezelfde professionele groep onderling op elkaar lijken. Dat fenomeen lijkt zichtbaar te worden in ons onderzoek. Het bleek allereerst uit een inhoudelijke analyse van de toelichtingen op de oordelen, hoewel dit verschil slechts in beperkte mate te kwantificeren is. De eerste groep (toezichthouders) koos overwegend voor een *integratieve* wijze van redeneren. De meerderheid koos ervoor om de argumenten pro- en contra gezamenlijk te bespreken en onderling af te wegen. De studenten kozen relatief vaker voor een *additieve* wijze van redeneren: argumenten werden meer los op een rijtje gezet waarbij ze als het ware 'toe-praatten' naar hun conclusie. De derde, meer homogeen samengestelde, groep van rechters en gerechtssecretarissen viel op door juist een meer *reductieve wijze van redeneren* te hanteren. Veel vaker dan in de andere groepen beenden deze respondenten de informatiestroom als het ware uit tot de kernfeiten die voor hen cruciaal waren. Zo schreef een van hen bijvoorbeeld na drie feiten correct te hebben benoemd: 'Alle andere informatie is voor mijn oordeelsvorming nu volstrekt irrelevant.' Deze verschillen in patronen suggereren dat verschillende groepen andere redeneerpatronen kunnen hanteren, hetgeen de relevantie van professionele socialisatie ook in relatie tot oordeelsvorming onderstreept.

Dit bleek helemaal doordat er ook significante verschillen tussen sommige groepen waren, waarbij hier met name de verschillen tussen verschillende typen rechters interessant zijn. Oordeelt een strafrechter anders dan de bestuursrechter of een civiele rechter? Het antwoord op basis van dit beperkte onderzoek is: 'ja, er lijken verschillen te zijn' en die verschillen zijn ook bij deze relatief kleine sample statistisch significant. Desalniettemin moeten we ook voor ogen houden dat we ons baseren op onderzoek met een kleine groep deelnemers, wat aanleiding geeft tot terughoudendheid in interpretatie en generalisatie. Figuur 2 hieronder brengt een belangrijk verschil naar voren.



Figuur 2: Verschillen tussen rechtsgebieden

In dit onderzoek heeft de groep strafrechters een eigen profiel: die groep is scherper negatief in zijn oordeel over de rol van de bestuurder en wijkt daarmee significant af van de andere groepen rechters. Daartoe komt men overigens niet zomaar, want men heeft ook meer woorden gebruikt (wat duidt op meer inzet) en meer correcte argumenten gereproduceerd. Tegelijk zijn in die woordenstroom ook meer kleine foutjes gemaakt. De wat grotere groep bestuursrechters valt vooral op doordat zij minder fouten maken, men zou kunnen concluderen dat zij als groep iets zorgvuldiger zijn. De handelsrechters als groep (de groep waarvan tevoren het meeste affiniteit met de problematiek van de bestuurdersaansprakelijkheid mocht worden verwacht), wijkt significant af op het punt van de beoordeling: zij achten de bestuurder relatief minder schuldig, blijkend uit de lagere getalsmatige inschatting. Kunnen we daarmee zeggen dat het type zaken dat

6. De one-way Anova werkt voor het resumeren van correcte argumenten ($F = 7{,}421$, $p = 0{,}008$) en ook voor het vermijden van fouten ($F = 5{,}106$, $p = 0{,}027$). Deze analyse van de relatie is alleen hier uitgevoerd en niet bij de eerste twee deel-experimenten, omdat daar geen correlatie was tussen verantwoording en oordeelskwaliteit, zoals besproken in de tekst.

7. Deze standaard luidt (zie www.rechtspraak.nl/SiteCollectionDocuments/professionele-standaarden-strafrecht.pdf): Een uitspraak van een strafrechter is, ongeacht de inhoud van of aanleiding tot de beslissing, gemotiveerd op een wijze die recht doet aan de zaak.

8. Dit is te meer van belang in de context van de replicatiecrisis waar de psychologie mee kampt en waarin soms spectaculaire en inspirerende effecten niet opnieuw kunnen worden aangetoond.

men doet, de rechter verder conditioneert? Zijn strafrechters meer belust op schuld toedelen? Dat is niet te zeggen, daarvoor is toch echt meer onderzoek nodig. De gevonden verschillen bij deze kleine groepen suggereren vooral dat dergelijk onderzoek relevant kan zijn.

6.3. Professionals: 'beter' maar ook mogelijk overmoedig

Naast verschillen binnen de groep van rechters waren er ook enkele relevante verschillen tussen studenten en professionele beoordelaars. Allereerst bleek dat de professionals eenvoudig 'beter' oordeelden. Ze waren beter in staat om argumenten correct te memoreren en ze waren vooral veel preciezer. De studenten maakten meer fouten, gemiddeld dubbel zo veel, als de professionals. En waar bijna de helft van de professionals foutloos oordeelde daar waren slechts 6 van de 77 studenten geheel foutloos. Dit is, hoewel wellicht een punt van aandacht voor de studenten en hun docenten, vooral een teken van bevestiging en geruststelling. Meer ervaren beoordelaars blijken ook betere en meer zuivere beoordelaars te zijn.

Tegenover deze positieve bevinding ten aanzien van de kwaliteit van professionele beoordelingen staat ook een kanttekening: de professionele beoordelaars waren ook meer uitgesproken en extreem in hun oordelen dan de studenten. De spreiding in antwoorden en de daarbij geëtaleerde zelfverzekerdheid in de beoordelingen was opvallend hoog bij de professionals. De onderlinge verschillen in mate van (on)schuld was veel groter dan bij studenten en de professionals durfden ook de meer extreme oordelen (richting 0 of 100) aan, waar studenten wat meer naar het midden tendeerden. Professionals durfden bijvoorbeeld significant vaker meer extreem (0-15 of 85-100) te oordelen dan de wat meer voorzichtige studenten (12% versus 8%). Tegelijk waren de verschillen tussen de professionals ook veel groter wat tot uiting komt in een grotere spreiding (met een standaarddeviatie van 18 versus een standaarddeviatie van 14). Professionals zijn in dit experiment dus tegelijkertijd meer uitgesproken én meer verschillend. Professionals waren dus stellig, extremer in hun oordelen maar onderling ook meer verschillend in weging van de casus. Hierin komt mogelijk een risico voor professionele beoordelaars naar voren: zij *moeten durven* oordelen, wat een bepaalde professionele zelfverzekerdheid met zich meebrengt die juist weer kwetsbaar maakt en tot zelfoverschatting kan leiden (Tetlock & Gardner, 2015).

7. Tot besluit

In de afgelopen decennia is in de gedragswetenschappen een keur aan onderzoeken uitgevoerd naar het oordeelsvermogen van en biases bij mensen en de factoren die daar een belemmerende of stimulerende rol bij vervullen, zoals verantwoording. Dit leverde conclusies op voor de juridische en toezichtspraktijk, daar toezichthouders en rechters ook mensen zijn, die tegelijk nog in meer realistische omstandigheden zouden moeten worden getoetst. Dit onderzoek heeft hierin een eerste stap willen zetten. De resultaten leveren bevestiging, geruststelling én nieuwe inzichten op.

Verantwoording is misschien wel eens irritant en levert dan stress op, maar het ontbreken van verantwoording heeft zichtbaar negatieve gevolgen voor de inzet van mensen en voor de kwaliteit van oordelen

De resultaten *bevestigen* allereerst dat ook professionele beoordelaars, waaronder rechters, vatbaar zijn voor beoordelingsbiases. Op kernpunten zijn de bevindingen bij professionals in een meer realistische setting vergelijkbaar met die bij studenten in minder realistische settings. Ook blijkt uit ons experiment opnieuw dat de anticipatie op latere verantwoording een gunstig effect heeft op de oordelen van professionals. Dit betekent ook dat er extra reden is om hoopgevende of zorgwekkende bevindingen uit psychologisch beoordelingsonderzoek met studenten serieus te nemen voor professionele praktijken van toezichthouders en rechters. Ook zijn de resultaten deels *geruststellend* daar professionele beoordelaars duidelijk zuiverder en preciezer oordeelden dan de studenten. Dat is zoals gehoopt en verwacht maar toch goed om te constateren: professionele ervaring loont. Tegelijk levert het onderzoek ook *nieuwe inzichten* op. Met name de constatering dat er duidelijke verschillen zijn tussen maar ook binnen de groepen en het risico van professionele zelfoverschatting. Dit is een goede voedingsbodem voor vervolgonderzoek. De praktische consequentie is vooral dat toezichthouders en rechters er goed aan doen te garanderen dat zij opereren in een passende verantwoordingscontext; voor de rechterlijke praktijk zou dat bijvoorbeeld kunnen betekenen dat de eisen die gesteld worden aan de (omvang van de) rechterlijke motivering ten minste gehandhaafd en eerder verstevigd dan afgezwakt moeten worden. Verantwoording is misschien wel eens irritant en levert dan stress op, maar het ontbreken van verantwoording heeft zichtbaar negatieve gevolgen voor de inzet van mensen en voor de kwaliteit van oordelen. •

Referenties

Aleksovska, M., Schillemans, T., & Grimmelikhuijsen, S. (2019), 'Lessons from five decades of experimental and behavioral research on accountability: A systematic literature review', *Journal of Behavioral Public Administration, 2*(2), 1-18, DOI: 10.30636/.

Baron, J. (1998), *Judgment misguided: Intuition and error in public decision making*, Oxford University Press.

Baron, J. (2008), *Thinking and Deciding,* Cambridge University Press.

Bauw, E., Dijk, F. van, & Sonnemans, J. (2013), 'De waarde van meervoud', *NJB* 2013/292, afl. 6.

Van Boom, W. H., Giesen, I., & Verheij, A. J. (2013), *Capita Civilologie. Handboek empirie en privaatrecht*, *6*, Den Haag: Boom Juridische uitgevers.

Cushing, B. E., & Ahlawat, S. S. (1996). 'Mitigation of recency bias in audit judgment: The effect of documentation', *Auditing, 15*(2), 110.

Doney, P. M., & Armstrong, G. M. (1995), 'Effects of accountability on symbolic information search and information analysis by organizational buyers', *Journal of the Academy of Marketing Science 24*(1), 57-65.

Enneking, L. F. H., Giesen, I. & Rijnhout, R. (2013), 'Bewijswaardering en psychologische inzichten', in: Van Boom, W. H., Giesen, I., & Verheij, A. J., *Capita Civilologie*, 1017-1085, Den Haag: Boom juridische uitgevers.

Feitsma, J. N. P. (2016), 'Meer dan een nudge: Gedragsexperts bij de Nederlandse overheid', *Bestuurskunde 25*(3), 24.

Giesen, I. (2015), *Asser Procesrecht. Deel 1. Beginselen van burgerlijk procesrecht*, Deventer: Wolters Kluwer.

Gravett, W. (2018), 'Subconscious Advocacy – Part 2: Verbal communication in the courtroom and ethical considerations', *Stellenbosch Law Review* 2018 (2), 175-198.

Kahneman, D., Slovic, P., & Tversky, A., *Judgment under uncertainty: Heuristics and biases*, Cambridge University Press.

Kahneman, D., & Egan, P. (2011), *Thinking, fast and slow, 1*, New York: Farrar, Straus and Giroux.

Van Koppen, P., de Keijser, J. W., Horselenberg, R., & Jelicic, M. (eds.). (2017), *Routes van het Recht: Over de rechtspsychologie*, Den Haag: Boom juridisch.

Majone, G. (1994), 'The rise of the regulatory state in Europe', *West European Politics, 17*(3), 77-101.

Mevis, P. A. M. (2001), 'Artikel 365a Sv', in: A. L. Melai, M. S. Groenhuijsen e.a. (eds.), *Wetboek van Strafvordering*, Deventer: Kluwer (losbladig).

Ottow, A. T. (2015), 'De lessons learned van toezichtrapporten', *Tijdschrift voor Toezicht, 6*(2), 44-52.

Peecher, M. E., & Piercey, M. D. (2008), 'Judging audit quality in light of adverse outcomes: Evidence of outcome bias and reverse outcome bias', *Contemporary accounting research, 25*(1), 243-274.

Pohl, R. (ed.) (2004), *Cognitive illusions*, Psychology Press.

Van der Post, N., & Toor, D. van. (2019), 'Volgorde effecten in het Nederlandse strafproces?', *Expertise en Recht, 1*, 41-54.

Raad van de rechtspraak (2018), *Jaarverslag 2017*, Den Haag 2018. https://jaarverslagrechtspraak.nl/downloads/Jaarverslag_2017.pdf

Rachlinski, J.J. (2012), 'How Judges make Decisions', in: R. Giard (ed.), *Judicial decision making in civil law*, The Hague: Eleven Publishing 2012, 87-105.

Schillemans, T. (2007), *Verantwoording in de schaduw van de macht: Horizontale verantwoording bij zelfstandige bestuursorganen*, Den Haag: Boom.

Schillemans, T. (2016), 'Calibrating Public Sector Accountability: Translating experimental findings to public sector accountability', *Public Management Review, 18*(9), 1400-1420.

Schillemans, T., & M. Bovens. (2018), 'Governance, Accountability and the Role of Public Sector Boards', *Policy & Politics*, DOI: 10.1332/030557318X15296526490810.

Scott, W. A. (1962), 'Cognitive complexity and cognitive flexibility', *Sociometry*, 405-414.

Suedfeld, P., & Tetlock, P. (1977), 'Integrative complexity of communications in international crises', *Journal of conflict resolution, 21*(1), 169-184.

Tetlock, P. E. (1983), 'Accountability and the perseverance of first impressions', *Social Psychology Quarterly,* 285-292.

Thaler, R., & C. Sunstein (2008), *Nudge: Improving decisions about health, wealth, and happiness*, Yale University Press.

Ten Velden, F. S. & De Dreu, C. K. W., *Sociaalpsychologische determinanten van straf rechtelijke besluitvorming*, Raad voor de rechtspraak.

WRR, (2014), *Van tweeluik naar driehoeken: versterking van interne checks and balances bij semipublieke organisaties*, Amsterdam: Amsterdam University Press.