

De uitkomsten van een forensisch-gedragsdeskundig onderzoek kunnen het verschil maken in de besluitvorming over schulduitsluiting en het toekennen van een straf dan wel een maatregel. Dat onderzoek kan grote gevolgen hebben voor het leven van een verdachte. Tamara De Beuf en Peter van Koppen toetsen in de gevalsstudie van Thijs Hermans de testdiagnostiek in een complexe strafzaak aan de huidige wetenschappelijke en beroepsmatige richtlijnen.

EEN GEVALSSTUDIE VAN FORENSISCH TESTGEBRUIK

HET TESTEN VAN THIJS

Een forensisch psycholoog doet, vaak samen met een forensisch psychiater, uitspraken over een persoon die verdacht wordt van een misdrijf. Een belangrijk instrument van de psycholoog is de testbatterij. Aan de hand van psychologische tests trachten deskundigen een antwoord te formuleren op de vragen die hen door de rechter worden voorgelegd. De tests variëren van tests van intelligentie, neuropsychologisch en emotioneel functioneren, psychische stoornissen tot persoonlijkheidstests en risicotaxatie-instrumenten. Daarnaast worden in de forensische context standaard vragenlijsten afgenomen om na te gaan of de onderzochte mogelijk symptomen over- of onderrapporteert.

Met testdiagnostiek verzamelen forensisch psychologen veel informatie over de onderzochte. De waarde die aan de verzamelde informatie kan worden gehecht, is afhankelijk van een aantal factoren zoals onder andere de kwaliteit van de gebruikte tests, de manier waarop de tests zijn afgenomen en de vertaalslag van testcores naar conclusies. Eerder onderzoek heeft laten zien dat testdiagnostiek in forensische gedragsdeskundige rapportages niet altijd en overal volgens de laatste wetenschappelijk inzichten dan wel beroepsrichtlijnen gebeurt (De Clercq & Vander Laenen, 2019; Neal et al., 2019; Schimmel & Van Koppen, 2017). In de onderhavige gevalsstudie toetsen wij de testdiagnostiek in een complexe strafzaak aan de hand van de huidige wetenschappelijke en beroepsmatige richtlijnen.

ROL VAN FORENSISCH GEDRAGSDESKUNDIGE

RAPPORTAGES IN STRAFZAKEN Volgens het jaarverslag van het Nederlands Instituut voor Forensische Psychiatrie en Psychologie (NIFP) zijn in 2021 bijna 12.000 gedragsdeskundige rapportages uitgevoerd voor verschillende doeleinden, zoals consultaties, strafzaken, voogdijzaken en tbs-verlenging (NIFP, 2022b). Dergelijke aantallen laten zien dat forensisch gedragsdeskundige rapportages behoorlijk ingeburgerd zijn in de Nederlandse strafrechtspleging.

De meest voorkomende vragen die forensisch psychiaters en psychologen voorgeschiedt krijgen, gaan over de toerekeningsvatbaarheid van de verdachte voor het misdrijf (Rassin, 2017). Dat is niet zo gek, want de rechter moet onderzoeken of de verdachte strafbaar is voor het ten laste gelegde feit nadat het feit bewezen is verklaard. Naast de eventuele aanwezigheid van een stoornis en het verband met de feiten is een rechter ook geïnteresseerd in het recidiverisico. Die informatie gebruikt de rechtbank bij de afweging of een tbs-maatregel kan worden opgelegd. De uitkomsten van het forensisch gedragsdeskundig onderzoek kunnen zo het verschil maken in de besluitvorming over schuldtoelating en het toekennen van een straf of een maatregel. Zodoende kan dat onderzoek grote gevolgen hebben voor het leven van een verdachte (Van Esch, 2012).

Het beoordelen van toerekeningsvatbaarheid is een complexe taak waarbij verschillende deskundigen in eenzelfde zaak niet altijd tot dezelfde conclusie komen. De

psychologen Guarnera & Murrie (2017) deden een systematische review naar de betrouwbaarheid van beslissingen over toerekeningsvatbaarheid en vonden dat onafhankelijke deskundigen in ongeveer 25 à 35% van de bestudeerde dossiers tot een andere conclusie kwamen (i.e., toerekeningsvatbaar versus niet toerekeningsvatbaar).

De redenen voor die verschillen zijn divers. Op basis van soms beperkte, soms tegenstrijdige informatie moet de deskundige conclusies trekken over de geestestoestand van de onderzochte ten tijde van de feiten, hetgeen maanden tot jaren geleden kan zijn (Guarnera & Murrie, 2017; Van Koppen, 2004). Bij dergelijke complexe besluitvorming kan bij voorbaat een matige overeenstemming worden verwacht (Mossman, 2013). Dat laat onverlet dat deskundigen een rol hebben in het bevorderen van de validiteit van hun rapportages. Dat gebeurt bijvoorbeeld met training en certificering, gebruik van gestructureerde en gestandaardiseerde methodes en het vermijden van stokpaardjes – de neiging om steeds in een bepaalde richting te concluderen (Guarnera et al., 2017; Murrie & Warren, 2005; Peters et al., 2017). Om de betrouwbaarheid en validiteit van forensisch psychologische evaluaties te verhogen en de invloed van de deskundige te verkleinen, worden kwaliteitseisen gesteld aan forensische rapportages.

KWALITEITSEISEN VOOR FORENSISCH PSYCHOLOGISCHE EVALUATIES Recent publiceerden forensisch psycholoog Tess Neal en collega's (Neal, Martire, et al., 2022) acht *best practices* voor forensisch gedragsdeskundige rapportages. Ze baseerden zich op richtlijnen uit standaardwerken in de forensische psychologie, richtlijnen opgesteld door beroepsorganisaties, alsook recente internationale ontwikkelingen in het bredere vakgebied van de forensische wetenschap. Met

de acht *best practices* benadrukken ze niet alleen het belang van de validiteit van de gebruikte methodes. Ze hebben ook aandacht voor bias, communicatiestijl, transparantie en algemene ethische principes zoals het volgen van beroepscodes (zie tabel 1).

In Nederland heeft het NIPF een richtlijn opgesteld voor forensisch psychologische onderzoek en rapportage, met als doel het onderzoek uniform en transparanter te maken (NIPF, 2022a).¹ De richtlijn beschrijft de context en randvoorwaarden van het psychologisch onderzoek. Er wordt ook een opzet gegeven voor de verschillende onderdelen in het onderzoek (namelijk anamnese, klinisch onderzoek, differentiaaldiagnose, beantwoording van de vragen). Een aanzienlijk deel van de richtlijn gaat over de testdiagnostiek waarbij deskundigen geïnformeerd worden over de aanbevolen tests, onderverdeeld per stoornis en probleem.

Het Nederlands Register Gerechtelijke Deskundigen (NRGD) toetst en registreert forensisch psychologen die zich beschikbaar stellen om als deskundige in een strafzaak te worden ingezet. Aan die registratie zijn eisen verbonden. Zo moeten de geregistreerde psychologen beschikken over 'testtheoretische kennis en kennis van de mogelijkheden en beperkingen van testmaterialen' en tests goed kunnen interpreteren (NRGD, 2021). Ze moeten ook goed de betrouwbaarheid en validiteit van de tests kunnen beoordelen. Onder meer op basis van de eisen die het NRGD stelt, schreven Van Esch en collega's (2021) een leidraad om juristen te ondersteunen bij het beoordelen van forensisch gedragsdeskundige rapportages. Ook voor gedragsdeskundigen is het een bruikbaar naslagwerk, omdat het aanbevelingen bevat voor het testpsychologisch onderzoek gebaseerd op of overeenkomstig met de bovenstaande richtlijnen.

WETENSCHAPPELIJK ONDERZOEK NAAR FORENSISCHE TESTDIAGNOSTIEK Onderzoek in de vs heeft laten zien dat, ondanks het bestaan van beroepsmatige richtlijnen, de kwaliteit van de testdiagnostiek in forensisch psychologisch onderzoek soms te wensen over laat. Neal en collega's (2019) maakten een overzicht van de tests die psychologen gebruiken in een forensische context. De meerderheid (90%) van de tests bleek empirisch te zijn geëvalueerd, maar slechts 40% was als goed beoordeeld door het *Mental Measurements Yearbook* (vergelijkbaar met de Commissie Testaangelegen-

In mei 2019 stak Thijs op twee verschillende dagen en locaties drie wandelaars dood die hun hond aan het uitlaten waren

1 Er bestaat een afzonderlijke richtlijn voor psychiatrisch onderzoek en rapportage in strafzaken, opgesteld door de Nederlandse Vereniging voor Psychiatrie (2013). Deze richtlijn wordt momenteel geüpdatet.

heden Nederland ofwel COTAN). Bovendien bleek dat tests werden gebruikt die niet breed geaccepteerd waren binnen de psychologie, zoals de *Buss-Durkee Hostility Inventory* (BDHI; voor de volledige lijst zie <https://osf.io/td8sm>).

Schimmel en Van Koppen (2017) deden in Nederland gelijkaardige vaststellingen bij forensisch testdiagnostiek uitgevoerd tussen 2008 en 2016. Zij toetsten de kwaliteit van de gebruikte tests aan de hand van beoordelingen door de COTAN, een commissie van het Nederlands Instituut van Psychologen (NIP) die minutieus tests beoordeelt. De COTAN let onder andere op de betrouwbaarheid (zoals foutenmarges van testscores), begripsvaliditeit (zoals meet de test wat het pretendeert te meten), criteriumvaliditeit (zoals de voorspellende waarde van de testscores) en de aanwezigheid van normen (zie voor gedetailleerde uitleg Evers et al., 2010). Bij de inventarisatie van de testdiagnostiek vonden Schimmel en Van Koppen (2017) dat slechts 28% van de 101 gebruikte tests door de COTAN als voldoende was beoordeeld.

Het is noemenswaardig dat de forensische rapporten die de Schimmel en Van Koppen onderzochten, waren goedgekeurd door de NRGD. Ze waren destijds immers gebruikt om de betrokken forensisch psychologen tot het NRGD-register toe te laten. Hoogstwaarschijnlijk volgden de deskundigen de NIPF-richtlijn, waarin tests worden aanbevolen die door de COTAN als onvoldoende worden beoordeeld.

Tot slot vonden Schimmel en Van Koppen dat de keuze van de testbatterij weinig werd gemotiveerd en dat informatie over de gebruikte tests en het verloop van de testafname vaak ontbrak.

DE ONDERHAVIGE STUDIE

Hoe ziet de situatie er vandaag uit? Wordt er voldoende onderbouwing gegeven bij de selectie van tests? Worden alleen valide en breed geaccepteerde tests gebruikt? Communiceren deskundigen transparant over het onderzoek en de testresultaten? Met andere woorden: Worden de richtlijnen in de hedendaagse praktijk gevolgd?

Voor ons onderzoek hadden wij een steekproef kunnen trekken uit een groot aantal forensische gedragskundige rapportages in strafzaken. Dat had wellicht een indruk kunnen geven over de vragenlijsten en tests die in heden daags forensisch Nederland worden gebruikt, maar dat had ons weinig diepgaand inzicht gegeven in het hoe en waarom van het psychologisch testonderzoek. Dergelijke rapporten kenmerken zich juist door een grote aandacht voor het individu van de verdachte en alle nuances die dat met zich meebrengt.

TABEL 1. ACHT BEST PRACTICES VOOR FORENSISCH PSYCHOLOGISCH ONDERZOEK VOLGENS NEAL, MARTIRE, ET AL. (2022).

1. Fundamentele validiteit (inclusief accuraatheid, herhaalbaarheid, reproduceerbaarheid)
2. Toegepaste validiteit (inclusief expertise van de deskundige en validiteit in de praktijk)
3. Beheersen en beperken van bias
4. Aandacht voor kwaliteitsborging
5. Juiste communicatie van gegevens, resultaten en meningen
6. Expliciete beschouwing van beperkingen en aannames
7. Afwegen van meningsverschillen of alternatieve visies
8. Volgen van beroepscodes, richtlijnen en regelgeving

Daarom kozen wij voor een gevalsstudie waarvoor de strafzaak tegen Thijs Hermans zich bij uitstek leent.² Die zaak, die hieronder verder wordt toegelicht, is ten eerste interessant omdat het forensisch-gedragsdeskundige onderzoek van groot belang was voor het oordeel van de rechter. Men mag verwachten dat de deskundigen om die reden zo zorgvuldig mogelijk hebben gewerkt. Ten tweede zijn er in die zaak drie forensisch-gedragsdeskundige rapporten uitgebracht. Dat biedt de mogelijkheid te onderzoeken hoe de rapporten zich tot elkaar verhouden en daarbij het testgebruik te vergelijken.

Concreet worden de volgende vragen over de testdiagnostiek in het strafdossier beantwoord:

1. Welke tests worden gebruikt en zijn die tests als voldoende beoordeeld door de COTAN?
2. Wordt er een onderbouwing gegeven voor de gekozen testbatterij en worden de gekozen tests voldoende toegelicht? Worden tests geselecteerd op basis van vooraf opgestelde hypothesen?
3. Worden de resultaten op een transparante en neutrale manier gecommuniceerd?
4. Passen de conclusies bij de bevindingen van de testdiagnostiek?
5. Hoe verhouden de verschillende rapporten zich ten opzichte van elkaar?

² Wij gebruiken de volledige naam van Thijs Hermans omdat dat ook op zijn verzoek gebeurde in de publicaties over zijn zaak in *NRC Handelsblad* door Bas Haan (2022, 2023a, 2023b).

GEVALSSTUDIE

Thijs werd vervolgd voor het om het leven brengen van drie mensen (voor een uitgebreide beschrijving van de zaak, zie Haan, 2022, 2023a, 2023b). In mei 2019 stak Thijs op twee verschillende dagen en locaties drie wandelaars dood die hun hond aan het uitlaten waren. De eerste doding vond plaats in een Haags natuurgebied, waarbij Thijs een vrouw om het leven bracht met tientallen messteken. Een aantal dagen later doodde hij, kort na elkaar, een vrouw en een man op de Brunssummerheide nabij Heerlen. Ook zij overleden door meerdere steekwonden.

In de eerste verhoren ontkende Thijs dat hij iets met de misdrijven te maken had, maar na drie maanden bekende hij. Als reden voor de ontkenning vertelde hij later dat hij van 'het systeem' niets mocht zeggen. Volgens Thijs verkeerde hij ten tijde van de feiten in een psychose waarbij hij in zijn waan ervan overtuigd was dat hij de moorden moest plegen om zichzelf en zijn familie te beschermen tegen 'het systeem'.

Dat was de grote inzet van de strafzaak: verkeerde Thijs inderdaad in een psychose ten tijde van de feiten? Is er een direct verband tussen de diagnose en de feiten en wat betekent dat voor de toerekeningsvatbaarheid? Zoals kon worden verwacht, werd Thijs voor observatie naar het Pieter Baan Centrum (PBC) verwezen.

De deskundigen van het PBC kwamen tot de conclusie dat Thijs ten tijde van de feiten in een psychotische toestand verkeerde en adviseerden volledige ontoerekeningsvatbaarheid.³ In juli 2020 werd Thijs door de rechtbank van Maastricht echter met verminderde toerekeningsvatbaarheid veroordeeld voor drievoudige moord en hij kreeg achttien jaar gevangenisstraf met terbeschikkingstelling opgelegd.⁴ De rechtbank week daarmee af van het advies van het PBC. Thijs ging in hoger beroep tegen de uitspraak.

In de aanloop van de behandeling door het hof werden nog twee forensisch-gedragsdeskundige onderzoeken uitgevoerd: een in opdracht van de verdediging, bemiddeld via het NIFP (hierna: het tweede rapport), en een in opdracht van de raadsheer-commissaris (hierna: het derde rapport). Het tweede onderzoek leidde tot hetzelfde advies als het PBC-rapport, terwijl in het derde rapport de deskundigen zich onthielden van een advies over toerekeningsvatbaarheid.

In maart 2022 kwam het hof in Den Bosch evenals de rechtbank tot de conclusie van verminderde toerekeningsvatbaarheid en veroordeelde Thijs tot 22 jaar gevangenisstraf met terbeschikkingstelling.⁵ De verdediging diende daarop een cassatieverzoek in dat in oktober 2023 door de Hoge Raad werd afgewezen. De Hoge Raad oordeelde dat het gerechtshof geen fouten heeft gemaakt en hield de veroordeling in stand.⁶

METHODE

Wij richtten ons in deze gevalsstudie op het testonderzoek van de drie rapporten. Wij onderzochten welke tests werden afgenomen, hoe erover werd gerapporteerd en welke conclusies werden getrokken uit de testresultaten. In de regel vormen het testgebruik en de rapportage daarover een separate paragraaf in het gezamenlijke rapport van de psycholoog en de psychiater. Niettemin moet het testgebruik worden gezien in de context van het bredere rapport en daarom werden ook onderdelen bestudeerd zoals de onderzoeksopzet, de forensisch-gedragsdeskundige beschouwing en de beantwoording van de vragen.

DRIE FORENSISCH GEDRAGSDESKUNDIGE RAPPORTEN

Wij maakten voor ons onderzoek gebruik van de drie rapporten waarnaar reeds werd verwezen. Het PBC-rapport werd in opdracht van de rechter-commissaris opgesteld en door de psycholoog en psychiater getekend op 4 december 2019. Het telt 136 bladzijden. Het tweede rapport, van 24 juni 2021, werd opgesteld in opdracht van de advocaten van Thijs en telt 100 pagina's. Het derde rapport, van 114 pagina's, werd geschreven in opdracht van de raadsheer-commissaris en uitgebracht op 13 januari 2022. Dat rapport kwam er op het verzoek van het hof vanwege de tijd die was verstreken sinds het afronden van het PBC-rapport en het afwijken van de rechtbank van de adviezen in dat rapport.⁷

De vraagstelling die werd voorgelegd aan de forensisch gedragsdeskundigen was voor allen dezelfde (zie tabel 2). Het was de standaardvraagstelling bij een forensisch-gedragsdeskundig onderzoek. Een uitzondering daarop was de extra

3 Als de rechter ook oordeelt dat de verdachte ontoerekeningsvatbaar is, dan heeft dat tot gevolg dat hem geen straf wordt opgelegd. Wel kan de rechter besluiten om een maatregel op te leggen, zoals de TBS-maatregel.

4 Rechtbank Limburg, 30 juli 2020, ECLI:NL:RBLIM:2020:5621.

5 Hof's-Hertogenbosch, 17 maart 2022, ECLI:NL:GHSHE:2022:868.

6 Hoge Raad, 17 oktober 2023, ECLI:NL:HR:2023:1295.

7 Hoe meer tijd is verstreken tussen misdrijf en het onderzoeken van de verdachte, hoe minder kan worden gezegd over de psychische toestand van de verdachte ten tijde van het misdrijf en hoe minder relevant de testdiagnostiek is voor de beslissing die de rechter moet nemen. In dat opzicht is de beslissing van het hof vanuit forensisch psychologisch perspectief merkwaardig.

TABEL 2. VRAAGSTELLING GERICHT AAN DE FORENSISCH GEDRAGSDESKUNDIGEN.

- o. Als verdachte weigert onderzocht te worden, tot welke overwegingen van de onderzoekers geeft die weigering aanleiding?
1. Is onderzochte lijdende aan een ziekelijke stoornis en/of gebrekkige ontwikkeling van de geestvermogens (geestesstoornis) en zo ja, hoe is dat in diagnostische zin te omschrijven?
2. Hoe was dit ten tijde van het plegen van het ten laste gelegde?
 - a. Kunt u aangeven, indien u ten tijde van het plegen van het ten laste gelegde een psychische stoornis heeft vastgesteld, of en in hoeverre deze stoornis een gevolg is van door verdachte gebruikte psychofarmaca, alcohol en/of drugs?
 - b. Betrokkene heeft lopende het strafrechtelijk onderzoek verklaard geen of beperkte herinneringen omtrent het ten laste gelegde te hebben, en later aangegeven een geheugenstoornis geveinsd te hebben. Kunt u aangeven, indien u een ziekelijke stoornis en/of gebrekkige ontwikkeling ten tijde van het ten laste gelegde heeft vastgesteld, op welke wijze u onderzocht heeft dat deze ziekelijke stoornis of gebrekkige ontwikkeling niet [te] duiden is als simulatie.
3. Beïnvloedde de eventuele ziekelijke stoornis en/of gebrekkige ontwikkeling van de geestvermogens onderzochte gedragskeuzes en gedragingen ten tijde van het ten laste gelegde?
4. Zo ja, kunt u dan gemotiveerd aangeven:
 - a. op welke manier dat gebeurde,
 - b. of dit leidt tot het advies om het ten laste gelegde in een verminderde mate dan wel in het geheel niet toe te rekenen, en,
 - c. indien geadviseerd wordt om in een verminderde mate toe te rekenen, preciezer dit gedragsdeskundig.
5.
 - a. Welke verwachting heeft u, gelet op de hiervoor beschreven stoornis, ten aanzien van het risico op recidive?
 - b. Welke beschermende functies in de persoonlijkheid of functioneren dienen hierbij in ogenschouw te worden genomen?
 - c. Welke contextuele, situatieve of andere condities dienen hierbij in ogenschouw te worden genomen?
 - d. Is er iets te zeggen over eventuele onderlinge beïnvloeding van deze factoren en condities?
6.
 - a. Welke aanbevelingen van gedragsdeskundige en van andere aard zijn te doen voor interventies die het eventuele recidivegevaar kunnen beperken?
 - b. Binnen welk(e) juridisch(e) kader(s) zouden deze gerealiseerd kunnen worden?

Aanvullende vraag in het derde rapport: 'Bijzondere vraagstelling'

Wat kunt u - vanuit uw expertise op gedragsdeskundig gebied - zeggen over de argumenten en het oordeel van de rechtbank, zoals vermeld in het vonnis, over de toerekenbaarheid van het tenlastegelegde aan verdachte, waarin door de rechtbank wordt afgeweken van het advies van de deskundigen van het Pieter Baan Centrum, te weten het advies om verdachte volledig ontoerekeningsvatbaar te verklaren?

vraag die voor het derde onderzoek door de raadsheer-commissaris aan de standaardvraagstelling was toegevoegd. De gedragsdeskundigen werd gevraagd commentaar te geven op de beslissing van de rechter af te wijken van de conclusie van het advies van de deskundigen van het Pieter Baan Centrum. Dat is een bijzondere vraag. Wij kennen geen andere zaken waarin een hof aan een deskundige vraagt om te reflecteren op de beslissing van de rechtbank.

ANALYSE Wij analyseerden de rapporten en vooral het testpsychologisch onderzoek aan de hand van richtlijnen die zijn opgesteld door het NIPP, het NRGD en door Neal en collega's (Neal, Martire, et al., 2022; Neal et al., 2019). Martire en Neal (2022) hebben de bovengenoemde acht *best practices* omgezet in vragen die gesteld kunnen worden bij het beoordelen van een rapport. Bijvoorbeeld: 'Has the practitioner provided citations to the research or professional literature for each of the assertions they make in their report?' Ze stelden in totaal 117 vragen op, maar zoals ze zelf schrijven, overlappen de vragen elkaar

nogal eens. De vragen van Martire en Neal (2022) hebben wij daarom gebundeld tot 81 vragen en we hebben die voor elk rapport beantwoord.⁸ Onze bevindingen worden hieronder besproken.

RESULTATEN

In het PBC-onderzoek werd het grootste aantal tests afgenomen (24; zie tabel 3). In de tweede onderzoek waren dat er zes en in het derde onderzoek twaalf. De tests hebben betrekking op verschillende domeinen zoals intelligentie, cognitieve functies, persoonlijkheid, copingvaardigheden, ervaren klachten en het risico op toekomstig geweld. Daarnaast werden in alle drie onderzoeken tests gebruikt die de validiteit van de antwoorden toetsen en meten of de persoon de neiging heeft om klachten en symptomen te overdrijven of te verbergen. Dergelijke tests zijn vooral belangrijk als de

8 De 81 door ons gecondenseerde vragen kunnen worden geraadpleegd op het Open Science Framework, <https://osf.io/zqje5>.

psychodiagnostiek plaatsvindt in een forensische context, zoals hier (Merckelbach & Dandachi-FitzGerald, 2021).

In het PBC-rapport en in het derde rapport werden ook projectieve tests gebruikt die pretenderen op een indirecte manier persoonlijkheid te peilen. Het gebruik van projectieve tests is controversieel, zeker daar waar geen gestandaardiseerde codering beschikbaar is en er daardoor een grote interpretatieruimte voor de onderzoeker ontstaat (NIPF, 2022a; Rassin, 2017). In totaal werden dertig verschillende tests gerapporteerd waarvan de meerderheid wordt aanbevolen door het NIPF (27; 90%). Negen van de dertig tests (30%), waarvan sommigen welbekende forensisch psychologische tests – zoals de STAXI-2 en HCR-20 – waren niet of nog niet beoordeeld door de COTAN.

Beoordeling door de COTAN

De tests die wel zijn beoordeeld volgens het COTAN-beoordelingssysteem⁹ laten een wisselend beeld zien (zie tabel 3). Vijf tests (24%; bijvoorbeeld PCL-R) scoren voldoende tot goed op de belangrijkste psychometrische kenmerken, namelijk betrouwbaarheid, begripsvaliditeit en criteriumvaliditeit. Vijf andere tests (bijvoorbeeld Rivermead BMT) scoren dan weer onvoldoende op de drie psychometrische kenmerken. De overige tests laten een mix van onvoldoende en voldoende psychometrische kwaliteit zien.

Bij de tests waarvoor er een COTAN-beoordeling beschikbaar is, heeft de meerderheid (16; 80%) geen of verouderde normen. Het probleem van verouderde of afwezige normen wordt besproken in de NIPF-richtlijn, waarbij men opmerkt dat een onvoldoende beoordeling op dat punt niet meteen betekent dat de test niet kan worden gebruikt. Volgens die richtlijn is het vooral belangrijk dat de deskundige op de hoogte is van de beperkingen en, met kennis van de laatste wetenschappelijke ontwikkelingen, het testgebruik kan verantwoorden. Als er voor sommige psychologische domeinen geen tests met een goede beoordeling voor de forensische praktijk beschikbaar zijn, dan kan de deskundige daar transparant over communiceren.

Onderbouwing van de testselectie

De selectie van de gebruikte tests wordt in de rapporten slechts in beperkte mate en in algemene termen onder-

bouwd. In het PBC-rapport wordt bijvoorbeeld wel toegelicht dat de GIT-2 afgenomen werd om de betrouwbaarheid van de resultaten van de WAIS-IV-NL te toetsen (zie tabel 3). De laatste test werd namelijk tien maanden eerder afgenomen, terwijl aangeraden wordt om minstens één jaar te wachten met testherhaling. Een leereffect kan immers resulteren in een positiever beeld van de intelligentie dan de werkelijke intelligentie (Lichtenberger & Kaufman, 2013).

Bij de andere tests blijft de onderbouwing beperkt tot het refereren naar de meetpretentie ('TAT is afgenomen om meer zicht te krijgen op de beleving van betrokkene'). Daarbij wordt de keuze niet afgewogen tegenover andere potentiële tests die hetzelfde pretenderen te meten en soms van betere kwaliteit zijn. In het tweede en derde rapport onderbouwen de deskundigen waarom zij menen dat er geen nieuw intelligentie- en neuropsychologische onderzoek nodig is, maar verder wordt geen onderbouwing gegeven anders dan hier en daar een verwijzing naar de meetpretentie. In het tweede rapport beargumenteert men wel nog dat persoonlijkheidsonderzoek maar beperkt nodig is omdat de toestand van Thijs ten tijde van hun testonderzoek nog steeds ten goede leek te veranderen. Kennelijk wordt bedoeld dat nieuw persoonlijkheidsonderzoek beperkt tot nieuwe inzichten zou kunnen leiden.

Volgens de NIPF-richtlijn kan de keuze van het instrumentarium worden verantwoord vanuit hypothesen die de deskundige wenst te toetsen. Uit de rapportages konden wij echter niet afleiden dat men van hypothesen gebruik heeft gemaakt. In dit dossier stond de vraag of Thijs al dan niet handelde onder invloed van een psychose centraal en via het opstellen van een hypothese hieromtrent zouden relevante tests kunnen zijn afgenomen. Men heeft echter geen tests gebruikt die psychotische symptomen onderzoeken. Er zijn nochtans instrumenten beschikbaar om symptomen van schizofrenie en psychotische stoornissen te onderzoeken zoals de *Positive And Negative Syndrome Scale* (PANSS, Løvgren et al., 2020), de vragenlijsten voor psychosegevoeligheid ontwikkeld door Chapman en collega's (zie voor een beschrijving Peters et al., 2008) en screeningsinstrumenten om syndromen in kaart te brengen, zoals de SCID (Arntz et al., 2018).

Normgroepen

Bij sommige tests kunnen deskundigen gebruikmaken van normgroepen waarmee de testscores van de verdachte kunnen worden vergeleken. Ook daarbij is het relevant te weten waarom een bepaalde normgroep wordt gekozen. Zo

9 De helft van de COTAN-beoordelingen is inmiddels meer dan 10 jaar oud. De oudste beoordeling dateert uit 1991 betreffende de Zin Aanvul Test en de meest recente uit 2016 betreffende de Nederlandse Klinische Persoonlijkheidsvragenlijst.

hadden de deskundigen bij het onderzoek van Thijs met de klachtenlijst (SCL-90-R; PBC-rapport) en de vragenlijst over copingvaardigheden (CISS; derde rapport) kunnen kiezen tussen verschillende normgroepen: algemene bevolking, psychiatrische patiënten en studenten. In geen van beide rapportages wordt echter toegelicht waarom een bepaalde normgroep werd gekozen. Meer nog, voor het interpreteren van de persoonlijkheidsvragenlijst NKPV maken de deskundigen een andere keuze: in het PBC-rapport wordt de vergelijking gemaakt met twee normgroepen (algemeen en psychiatrische patiënten), terwijl in het derde rapport alleen de vergelijking wordt gemaakt met de normgroep 'algemeen'. Het waarom van die keuze wordt uit de rapporten niet duidelijk, terwijl het gevolgen kan hebben voor het beeld dat over Thijs wordt gepresenteerd. In het tweede rapport wordt niet vergeleken met normgroepen, terwijl er voor sommige vragenlijsten wel degelijk normgroepen beschikbaar waren (bijvoorbeeld CISS).

Toelichting bij de testbatterij

Hoewel de testkeuze in het PBC-rapport nauwelijks wordt verantwoord, worden de tests wel toegelicht in een bijlage bij het rapport. Voor elke gehanteerde test, behalve FEAST, is er een omschrijving beschikbaar. Die omschrijving bevat echter geen informatie over de psychometrische kwaliteiten van de tests, met uitzondering van de MMPI-2 en WAIS-IV-NL.

In de andere twee rapporten wordt minder uitleg gegeven bij de gebruikte tests. Hier en daar wordt een test toegelicht met een enkele zin ('Deze lijst betreft een zelfbeoordelingsvragenlijst met betrekking tot posttraumatische stressklachten'; derde rapport), maar veelal valt men met de deur in huis ('Op deze zelfrapportagetest scoort betrokkene in vergelijking met de normgroep [...]'; derde rapport). De bespreking van de psychometrische kwaliteiten van de gebruikte tests blijft in het tweede rapport beperkt tot een vermelding van de betrouwbaarheid en validiteit van het intelligentieonderzoek uitgevoerd door het PBC. Verder wordt er in de testdiagnostiek van het tweede en derde rapport niets geschreven over de psychometrische kwaliteiten van de tests.

Geen van de drie rapportages verwijst naar de COTAN. Gebruikers van de rapportage hebben op die manier geen zicht op de kwaliteit van de tests. Zo kunnen rechters minder goed beoordelen welk bewijs voldoende valide is om mee te wegen in hun beslissing (zie voor een bespreking in verband met het Angelsaksische recht Neal et al., 2019). Zonder informatie over de kwaliteit van de gebruikte tests kunnen

rechters die poortwachtersfunctie minder goed vervullen.

In de testonderzoeken en de rapportage in het algemeen wordt evenmin gebruikgemaakt van of verwezen naar wetenschappelijke bronnen. De afwezigheid van wetenschappelijke referenties is in overeenstemming met de NIPP-richtlijn. Daarin wordt beargumenteerd dat het vermelden van wetenschappelijke referenties het rapport een schijn van wetenschappelijkheid zou geven. Er wordt gesteld dat men ervan mag uitgaan dat deskundigen volgens de laatste wetenschappelijk inzichten werken, conform de gedragscode van de NRGD (NIPP, 2022a). In de *best practices* van Neal en collega's wordt wel gewezen op het belang van wetenschappelijke bronvermeldingen vanwege transparantie over wat de onderliggende bronnen zijn waarop uitspraken en stellingen in een rapportage gestoeld zijn (Neal, Martire, et al., 2022). Het toevoegen van dergelijke informatie zou de leesbaarheid van een rapport kunnen verminderen. Om dat te voorkomen, zou de achtergrondinformatie over het instrumentarium in een bijlage kunnen worden gebundeld. Gebruikers, zoals gedragsdeskundigen die een contra-expertise uitvoeren, kunnen daar dan informatie terugvinden over de testkeuze, keuze van normgroepen, psychometrische kwaliteiten en de wetenschappelijke bronnen waarop die informatie is gebaseerd.

Toelichting bij de testomstandigheden

Naast de beperkte onderbouwing en omschrijving van de gebruikte tests wordt er weinig informatie gegeven over de testomstandigheden. In welke setting en onder welke omstandigheden werden de tests afgenomen? In welke volgorde zijn de tests afgenomen? Informatie over het toestandsbeeld van de betrokkene, maar ook storende omgevingsfactoren, kan immers relevant zijn voor de interpretatie van de testresultaten. Ook de volgorde waarin tests zijn afgenomen, bijvoorbeeld in functie van de moeilijkheidsgraad of mate van cognitieve inspanning, kan betekenisvol zijn voor het begrijpen van de testresultaten.

In de onderzochte rapporten wordt doorgaans wel gerapporteerd door wie en wanneer het testpsychologisch onderzoek werd uitgevoerd en hoe Thijs zich presenteerde, maar er wordt geen informatie gegeven over de testvolgorde, de duur van het testonderzoek, de ruimte waarin en het tijdstip waarop de tests werden afgenomen. In het derde rapport wordt een bijzonderheid vermeld, namelijk dat Thijs tussen twee contactmomenten een aantal (vermoedelijk vijf) vragenlijsten heeft meegekregen om in te vullen. Het is niet duidelijk of het invullen onder toezicht gebeurde, zoals de

TABEL 3. AFGENOMEN TESTS PER FORENSISCH GEDRAGSDESKUNDIG RAPPORT.

	PBC-RAPPORT ^A	TWEEDE RAPPORT ^B	DERDE RAPPORT ^C	COTAN-BEOORDELING ^D
Aantal tests	24	6	12	
<i>Symptoomvaliditeit</i>				
1. Amsterdamse Korte Termijn Geheugentest (AKTG)	X			+ criteriumvaliditeit - betrouwbaarheid, begripsvaliditeit, normen
2. Structured Inventory of Malingered Symptomatology (SIMS)	X	X	X	+ betrouwbaarheid, criteriumvaliditeit - begripsvaliditeit, normen
3. Test of Memory Malingering (TOMM)	X			n.v.t.
4. Supernormality Scale Revised (SSR)		X	X	n.v.t.
<i>Intelligentie</i>				
5. Wechsler Adult Intelligence Scale IV-NL (WAIS-IV-NL)	X			++ betrouwbaarheid + begripsvaliditeit, normen - criteriumvaliditeit
6. Groninger Intelligentietest II (GIT-II)	X			++ begripsvaliditeit + betrouwbaarheid, criteriumvaliditeit, normen
<i>Neuropsychologisch</i>				
7. Boston Naming Test (BNT) ^e	X			n.v.t.
8. Nederlandse Leestest voor Volwassenen (NLV)	X			+ betrouwbaarheid, begripsvaliditeit - criteriumvaliditeit, normen
9. 15-Woordentest	X			++ betrouwbaarheid, begripsvaliditeit - criteriumvaliditeit, normen
10. Rivermead BMT	X			- alle psychometrische eigenschappen, normen
11. Complexe Figuurtest van Rey (CFT)	X			+ betrouwbaarheid, begripsvaliditeit - criteriumvaliditeit, normen
12. Wisconsin Card Sorting Test (WCST)	X			n.v.t.
13. Trail Making Test (TMT) ^e	X			- alle psychometrische eigenschappen, normen
14. Stroop kleur-woord test	X			++ betrouwbaarheid + begripsvaliditeit - criteriumvaliditeit, normen
15. D2 aandachts- en concentratietest	X			- alle psychometrische eigenschappen, normen
16. Dierentuin plattegrondtest (BADs)	X			- alle psychometrische eigenschappen, normen
17. FEAST ^e	X			n.v.t.
18. Screeningslijst voor ADHD bij Volwassenen (onduidelijk welke)			X	n.v.t.

	PBC-RAPPORT ^A	TWEEDE RAPPORT ^B	DERDE RAPPORT ^C	COTAN-BEOORDELING ^D
Klachten				
19. Symptom Checklist 90-Revised (SCL-90-R)	X			++ betrouwbaarheid, begrips- en criteriumvaliditeit - normen
20. Zelf Inventarisatie Lijst (ZIL)			X	+ betrouwbaarheid, begripsvaliditeit - criteriumvaliditeit, normen
Persoonlijkheid				
21. Minnesota Multiphasic Personality Inventory- 2 (MMPI-2)	X		X	+ betrouwbaarheid - construct- en criteriumvaliditeit, normen
22. Nederlandse Klinische Persoonlijkheids-vragenlijst (NKPV)	X		X	++ normen en betrouwbaarheid + construct- en criteriumvaliditeit
23. Nederlandse Persoonlijkheids-vragenlijst 2 herzien (NPV-2-R)	X			++ normen, betrouwbaarheid, begripsvaliditeit. + criteriumvaliditeit.
24. Psychopathy Checklist-Revised (PCL-R)	X	X		+ betrouwbaarheid, begrips- en criteriumvaliditeit - normen
Copingvaardigheden				
25. Coping Inventory for Stressful Situations (CISS)	X	X	X	+ betrouwbaarheid en begripsvaliditeit. - criteriumvaliditeit, normen
26. STAXI-2		X	X	n.v.t.
Projectieve tests				
27. Thematic Apperception test (TAT)	X		X	n.v.t.
28. Zinnen Aanvullen Test (ZAT)			X	Groningse versie: onvoldoende op alle vlakken
Risicotaxatie				
29. HCR-20 V3	X	X	X	n.v.t.
30. SAPROF			X	n.v.t.
a. Het PBC-rapport was gebaseerd op tests die zijn afgenomen tussen 26 augustus en 11 oktober 2019. b. Het tweede rapport was gebaseerd op tests die zijn afgenomen op 12 februari 2021. c. Het derde rapport was gebaseerd op tests die zijn afgenomen tussen 5 en 23 juli 2021 en op 3 november 2021. d. ++ = beoordeeld als goed; + = beoordeeld als voldoende; - = beoordeeld als onvoldoende; n.v.t. = niet beoordeeld door COTAN. e. * = Niet opgenomen in de NIFP-richtlijn.				

NIFP-richtlijn voorschrijft. Het is in het algemeen *bad practice* om een test niet onder begeleiding en toezicht af te nemen, en wel om voor de hand liggende redenen. Als daarvan wordt afgeweken, moet dat worden gemotiveerd. Nu dat niet is gebeurd, is dat nefast voor de waarde die aan de uitkomsten

van die vragenlijsten kan worden gehecht en dus voor de transparantie van het onderzoek.

Communiceren van testcores

In geen van de drie rapporten worden testcores, zoals

De discipline van de forensische psychologie heeft vooralsnog geen glazen bol waarmee het verleden en de toekomst kristalhelder worden

schaalscores, totaalscores of profielscores, gerapporteerd. Alleen de vier indexscores van de intelligentietest worden genoemd; initieel in het PBC-rapport en daarna overgenomen in de volgende twee rapporten. Het lijkt erop dat men niet eerst de ruwe testcores bij het PBC heeft opgevraagd, wat formeel wel mogelijk is, zoals vastgelegd in het NRGD-beoordelingskader (NRGD, 2021). Zonder testcores kunnen de interpretaties van de testresultaten niet worden getoetst en zijn de conclusies uit de testdiagnostiek niet controleerbaar.

Een voorbeeld daarvan in de zaak van Thijs is het niet vermelden van de profielscores van de MMPI-2. Zodoende kunnen de grote verschillen in de persoonlijkheidsomschrijving tussen het PBC-rapport en het derde rapport niet worden beoordeeld. We kunnen niet nagaan of de testonderzoeken op een ander profiel uitkwamen of dat de profielscore anders werd geïnterpreteerd.

Heldere communicatie van bevindingen

Een goede communicatie van testbevindingen houdt ook in dat een vlot leesbaar rapport wordt afgeleverd. Ten eerste kan een duidelijke en inzichtelijke structuur positief bijdragen aan de leesbaarheid en de toegankelijkheid van een rapport. Ook daarin verschillen de rapporten van elkaar. In het PBC-rapport wordt evident aandacht besteed aan de tekststructuur met behapbare paragrafen, tussenkopjes én een inhoudsopgave. Dat staat in contrast met de andere twee rapporten waar alinea's soms zonder onderbreking meerdere pagina's bestrijken. Wat betreft het testpsychologisch onderzoek worden zowel in het PBC-rapport als het derde rapport de tests overzichtelijk een voor een gepresenteerd, terwijl de tests in het tweede rapport in een lopende tekst beschreven worden.

Ten tweede is taalgebruik belangrijk. Een goed leesbaar rapport bevat geen spelfouten en jargon, en maakt gebruik van neutrale formuleringen (Neal, Martire, et al., 2022; NRGD, 2021). In elk rapport wordt in meer of mindere mate jargon gebruikt zonder bijkomende uitleg of een verklarende woordenlijst. Vooral in het derde rapport maakt men gebruik

van jargon en onduidelijke taalgebruik, in het bijzonder als de deskundigen schrijven over de persoonlijkheidsopbouw van Thijs. Daarin worden – zonder uitleg – termen gebruikt zoals 'structuurdiagnostiek', 'splitting', 'projectieve identificatie' en 'werkelijkheidsdomeinen'. Er is sprake van vage en moeilijk te volgen formuleringen. We geven een voorbeeld:

'Het antwoord op die vraag bestaat in een klinisch oordeel, dat uit de aard der zaak – omdat het klinische oordeel zich op niet-discursieve wijze, als een soort "gut feeling", in het contact openbaart – niet argumentatief is en zich ook niet ondubbelzinnig, wellicht zelfs helemaal niet, laat onderbouwen in zgn. "objectieve" termen, d.w.z. in termen die refereren aan werkelijkheidsdomeinen die gelegen zijn buiten het domein van het klinische zelf' (cursivering in origineel).

De formuleringen in het derde rapport wekken ook op verschillende momenten de indruk dat Thijs pathologischer wordt voorgesteld dan de bevindingen rechtvaardigen. Zo scoort Thijs op de vragenlijst over boosheid (STAXI-2) binnen het normale bereik. Toch schrijven de deskundigen: 'In relatieve zin scoort hij hoog op de neiging om naar buiten gerichte boosheid te voorkomen.' Relatief ten opzichte van wat?

Een ander voorbeeld uit het derde rapport is de interpretatie van de projectieve test TAT. In die test krijgen personen een reeks plaatjes te zien waarbij ze een narratief moeten bedenken. Dit narratief wordt vervolgens volgens Freudiaans denken geëvalueerd om iets over de persoonlijkheid van het subject te weten te komen. In hun analyse focussen de deskundigen op hoe Thijs reageerde op de 'negatieve elementen' in de plaatjes en laten ze eventuele positieve of neutrale aspecten buiten beschouwing.

Discrepanties in bevindingen en conclusies

We onderzochten de conclusies van het testonderzoek en de antwoorden op de vraagstelling van de opdrachtgevers. Zoals eerder gezegd worden in de rapporten, behalve bij de intelligentietests, geen testcores gecommuniceerd.

Daardoor is het onmogelijk na te gaan of de interpretaties en conclusies van de deskundigen gerechtvaardigd zijn. Desondanks noemen we een aantal discrepanties die vragen oproepen maar in de rapporten niet worden geduid. Zo schrijven de deskundigen in alle drie rapporten – elk in hun eigen bewoordingen – dat Thijs mogelijk (dis)simuleert, hoewel dat niet blijkt uit de testresultaten die ze rapporteren. Een ander voorbeeld van een ogenschijnlijke inconsistentie die niet wordt uitgelegd, is dat de deskundigen in het derde rapport concluderen dat Thijs agressieregulatieproblemen heeft, terwijl hij binnen het normale bereik scoort op de STAXI-2 waarmee onder meer de controle van boosheid wordt gemeten.

Tot slot merkten we op dat in de forensisch gedragsdeskundige beschouwing en de beantwoording van de standaardvraagstelling weinig expliciet wordt verwezen naar het testpsychologische onderzoek, terwijl daarvoor wel gelegenheid was. Zo formuleert men in het tweede rapport conclusies over het al dan niet simuleren door Thijs zonder te verwijzen naar de symptoomvaliditeitstesten, of spreekt men in het derde rapport over de persoonlijkheid van Thijs zonder te verwijzen naar de MMPI of andere uitgevoerde persoonlijkheidstests. Wellicht zullen deskundigen beargumenteren dat de testpsychologische bevindingen impliciet in de conclusies vervat liggen en een expliciete verwijzing overbodig is. Toch is het aangeraden om conclusies te objectiveren via de uitgevoerde testdiagnostiek (Neal, Martire, et al., 2022). Uit onderzoek blijkt immers dat de validiteit van goede tests superieur is aan het klinische oordeel van psychologen en psychiaters (Garb et al., 2012).

Gepaste onzekerheid bij het formuleren van conclusies

Conclusies in deskundigenverslagen dienen met een gepaste uitdrukking van onzekerheid te worden geformuleerd. Men kan beargumenteren dat stellige conclusies met een definitieve formulering – zoals: een kenmerk is wel of niet aanwezig – niet gepast zijn als het gaat over het schatten van een psychische toestand in het verleden of in de toekomst (Neal, Lienert, et al., 2022; Rassin, 2017). De discipline van de forensische psychologie heeft vooralsnog geen glazen bol waarmee het verleden en de toekomst kristalhelder worden (zie Van Koppen, 2004).

De rapporten verschillen van elkaar wat betreft de stelligheid van de conclusies. In het PBC-rapport en het tweede rapport wordt een mix van definitieve en meer voorzichtige uitspraken gebruikt. De deskundigen zijn bijvoorbeeld stellig over de aanwezigheid van psychotische

ontregeling ten tijde van de ten laste gelegde feiten: 'In de dagen voorafgaand aan en ten tijde van de ten laste gelegde feiten was betrokkene dan ook ernstig psychotisch ontregeld.' (PBC-rapport). Bij andere formuleringen wordt wel rekening gehouden met enige mate van onzekerheid; bijvoorbeeld: 'Met betrekking tot de Dexamfetamine is betoogd dat het, [...], zeer aannemelijk is dat dit de psychose, [...], heeft aangejaagd en verergerd.'

De deskundigen van het derde rapport zijn voorzichtiger in hun conclusies. Dat is zichtbaar in formuleringen zoals '(niet) kunnen vaststellen', 'niet suspect voor het bestaan van' en 'dat het aannemelijk is'. In dit rapport lijkt de voorzichtigheid echter over te gaan in volledige onzekerheid zodat de deskundigen vragen onbeantwoord laten ondanks alle bestudeerde informatie. Zo lijken de deskundigen zich bij de risicotaxatie in bochten te wringen om geen eindoordeel te moeten formuleren. Ze schrijven namelijk dat uit de risicotaxatie met behulp van de HCR-20 V3 en SAPROF 'geen evident hoog risico op gewelddadig recidive' blijkt, zonder daarbij de eindoordelen te vermelden. Ze vervolgen: 'Dat wil naar de mening van deskundigen niet zeggen dat dat risico daarom ook laag is'. Door te refereren naar hun mening leggen de deskundigen de uitkomst van de tests naast zich neer en gaan zij af op hun klinische oordeel, wat ingaat tegen de aanbevelingen van evidencebased risicotaxatie (Heilbrun et al., 2021). De risicofactoren uit de instrumenten en hun eventuele onderlinge samenhang worden niet gecommuniceerd, zoals wel aanbevolen wordt in de NIPF-richtlijn.

Risicotaxatie

Het beschrijven en wegen van risicofactoren ontbreekt ook in het PBC-rapport. De deskundigen beschrijven vooral hun ongestructureerde klinische oordeel en verwijzen slechts naar de gestructureerde risicotaxatie met de HCR-20 in de zin dat het hun klinisch beeld bevestigt, zonder daarbij de relevante risicofactoren uit het instrument te noemen. Het omgekeerde gebeurt ook: deskundigen van zowel het PBC-rapport als het tweede rapport vermelden beschermende factoren zonder dat ze daarvoor een risicotaxatie-instrument zoals de SAPROF hebben gebruikt.

In het tweede rapport wordt de risicoanalyse wel volgens de NIPF-richtlijn en het gestructureerde professioneel oordeel gecommuniceerd door te vermelden onder welke omstandigheden het recidiverisico toeneemt en wat nodig is om het risico te verminderen. De HCR-20 V3 wordt echter foutief geïntroduceerd als een instrument dat vooral is ontwikkeld om geweld te voorspellen bij mensen met een

persoonlijkheidsstoornis. Dat klopt niet: de handleiding zegt niet dat de onderzochte een stoornis of een formele diagnose moet hebben (De Vogel et al., 2022).

Hoe verhouden de rapporten zich tot elkaar?

De deskundigen van het tweede en het derde rapport namen kennis van het eerste rapport. Zo namen zij testbevindingen over van het PBC-onderzoek. De deskundigen van het tweede rapport verwijzen naar het PBC-rapport voor bevindingen over intelligentie en cognitieve vaardigheden. De psychologe schrijft dat zij eerst het testonderzoek van het PBC-rapport heeft bestudeerd alvorens zelf tests af te nemen. Hetzelfde geldt voor het derde rapport: de deskundigen maken gebruik van het intelligentie-onderzoek dat werd uitgevoerd door het PBC en het lijkt erop dat ook zij eerst de vorige twee rapporten hebben gelezen alvorens hun eigen testonderzoek aan te vangen. Dat laatste wordt echter niet expliciet gerapporteerd. In het tweede noch het derde rapport wordt over die afhankelijkheid van het PBC-rapport gereflecteerd: Welke gevolgen zou dat kunnen hebben? Kan deze voorkennis het eigen onderzoek hebben gekleurd? Wat hebben ze eraan gedaan om die gevolgen te beperken?

Dat de rapporten elkaar in tijd opvolgen, biedt de mogelijkheid om de bevindingen met elkaar te vergelijken en tegen elkaar af te zetten. Hoewel dat niet opgenomen is in de vraagstelling kan het toch interessant zijn voor het onderzoek. De deskundigen van het tweede rapport namen bijvoorbeeld vier tests af die ook al door het PBC waren afgenomen en zich dus lenen voor vergelijking (SIMS, CISS, PCL-R, HCR-20 V3). Hoewel men bij de bespreking van de SIMS en de CISS wel verwijst naar de bevindingen in het PBC-rapport, ligt de nadruk op de overeenkomsten. Terwijl ook verschillen informatief kunnen zijn voor diagnostische overwegingen: kunnen de verschillen gerelateerd worden aan psychische toestand van Thijs? Passen de verschillen bij de behandeling die Thijs inmiddels kreeg? Of zijn de verschillen te duiden vanuit de onderzoeksopzet?

In het derde rapport wordt geen vergelijking gemaakt met het testonderzoek uit de eerste twee rapporten ondanks dat men voor verschillende tests tot andere bevindingen komt. Het niet vergelijken van de testpsychologische bevindingen tussen resultaten is een gemiste kans. Zoals we net schreven, kunnen verschillen relevant zijn vanuit diagnostische overwegingen, maar ze kunnen ook te wijten zijn aan de eigenschappen van de gebruikte tests. Hoe ongestructureerder, subjectiever en intuïtiever een test, hoe groter de kans dat er verschillen zijn tussen twee afnames

(Neal, Martire, et al., 2022). Het is aan deskundigen die verschillen en hun oorzaken te herkennen en te bespreken. Zo kunnen deskundigen laten zien dat ze kritisch en rigoureus te werk zijn gegaan (Neal, Martire, et al., 2022). Er is in de wetenschappelijke literatuur namelijk veel te doen over de betrouwbaarheid van forensische diagnostiek (zie voor een uiteenzetting Guarnera et al., 2017). Bovenal is een dergelijke duiding zinvol voor de gebruikers van de rapportages, zoals rechters die moeten oordelen over de bruikbaarheid van de onderzoeken (Drees et al., 2020).

DISCUSSIE

De vragen die in forensisch-gedragsdeskundige rapportages moeten worden beantwoord, zijn lastig. Deskundigen wordt immers gevraagd een schatting te maken van iemands toestand in het verleden én in de toekomst. Er bestaan vooralsnog geen tests om retrospectief te diagnosticeren, laat staan om een causaal verband vast te stellen tussen die diagnose en het delict dat heeft plaatsgevonden (Van Koppen, 2004). Forensisch gedragsdeskundigen roeien met de riemen die ze hebben en een belangrijk instrument in hun toolbox is het testpsychologisch onderzoek. Het is daarbij belangrijk dat deskundigen expliciet hypothesen formuleren en toetsen, gebruikmaken van betrouwbare en valide tests en transparant communiceren over het proces en de resultaten (Neal, Martire, et al., 2022; Peters et al., 2017). Op die manier kunnen ze een relevante en wetenschappelijk onderbouwde bijdrage leveren aan de besluitvorming van de rechter.

Deze gevalsstudie biedt een unieke en gedetailleerde inkijk in de testdiagnostiek van forensisch gedragsdeskundigen rapporten van een strafzaak die inmiddels veel aandacht heeft gekregen bij de betrokken beroepsgroepen zoals strafrechters en gedragsdeskundigen. De zaak is hoogstwaarschijnlijk niet representatief voor andere zaken waarin forensisch gedragsdeskundigen wordt gevraagd hun expertise te verlenen. Het is namelijk bijzonder dat er drie forensisch gedragsdeskundige rapporten beschikbaar zijn en alle bovendien van aanzienlijke omvang. Desalniettemin kunnen de bevindingen van deze gevalsstudie interessant zijn voor al wie beroepsmatig met dergelijke rapporten te maken krijgt.

Elk rapport dat in deze gevalsstudie onderzocht werd, heeft sterke punten die in overeenstemming zijn met de meest recente nationale en internationale richtlijnen voor forensisch gedragsdeskundige rapportage. Zo vonden wij in het PBC-rapport een diversiteit aan tests, waarbij het testgebruik bovendien overzichtelijk wordt gepresenteerd,

Ondanks dat bias en het beperken van bias bij forensisch onderzoek in de wetenschappelijke literatuur de laatste jaren veel aandacht heeft gekregen, is het onderwerp niet opgenomen in de NIFP-richtlijn

inclusief een bijlage met een omschrijving van de gebruikte tests. Het tweede rapport kenmerkt zich door een goede risicocommunicatie waarin transparant wordt gecommuniceerd over de relevante risicofactoren en de context waarin een hoog recidiverisico zich kan manifesteren. In het derde rapport zien we, net als in het RVC-rapport, een overzichtelijke weergave van het testpsychologisch onderzoek en weerhouden de deskundigen zich van te stellige uitspraken. De meerderheid van de gebruikte tests in de drie rapporten wordt bovendien aanbevolen in de NIFP-richtlijn.

Desondanks konden wij nog een aantal verbeterpunten identificeren die – als ze gevolgd worden – de rapporten meer zouden afstemmen met de *best practices* die in de inleiding werden aangehaald. Zo vonden we dat de rapporten meer toelichting kunnen geven over de selectie van tests en wat hun meetpretenties en psychometrische kwaliteiten zijn. Op die manier wordt transparantie over de gekozen testmethode gegarandeerd en kunnen lezers en gebruikers van de rapporten zich zelf een oordeel vormen over de waarde van de testdiagnostiek. Daarnaast zouden de interpretaties en conclusies die uit de testdiagnostiek volgen, controleerbaar moeten zijn – bijvoorbeeld voor contra-expertise. In de rapporten die in deze gevalstudie onderzocht zijn, waren de ruwe testcores niet beschikbaar en was er slechts beperkte informatie over de testomstandigheden. De afwezigheid van dergelijke informatie maakt het lastig om de conclusies van de testdiagnostiek te toetsen.

De mogelijkheid om interpretaties en conclusies te toetsen, wordt met name relevant als ogenschijnlijke tegenstrijdigheden niet worden geduid. Wij identificeerden verschillende discrepanties zowel tussen testresultaten onderling, testresultaten en andere bronnen, als tussen testresultaten en conclusies die volgden uit de testdiagnostiek. Discrepanties en inconsistenties hoeven niet noodzakelijk een probleem te zijn, maar ze moeten wel worden benoemd en geduid. Dat gebeurde weinig in de drie rapporten.

We identificeerden ook discrepanties tussen de rapporten. Zoals we hierboven reeds schreven, kunnen deskundigen inconsistenties aangrijpen om potentieel interessante inzichten te genereren, zowel inhoudelijk als over de gevolgde testprocedure. Een beschouwing daarvan kan gebruikers ondersteunen bij het interpreteren van de testdiagnostiek, bijvoorbeeld door de testresultaten in de context van elk specifiek onderzoek te plaatsen.

Tot slot laat deze gevalstudie zien dat aandacht voor de leesbaarheid van forensisch-gedragdeskundige rapporten nog steeds relevant is. Instructies zoals zorgen voor een overzichtelijke structuur, duidelijke en neutrale formuleringen en het weglaten van jargon zijn nog steeds nodig. Dat dergelijke instructies niet altijd gevolgd worden, werd eerder al aangetoond in onderzoek door Van Esch (2012) en door Schimmel en Van Koppen (2017).

BREDERE KWALITEITSEISEN VOOR RAPPORTAGES In de huidige gevalstudie lag de nadruk op de testdiagnostiek, maar er zou een analyse van de volledige rapportages kunnen worden gemaakt aan de hand van de *best practices* die zijn opgesteld door Neal en collega's. In die richtlijnen – gebaseerd op internationale richtlijnen uit de forensische psychologie en forensisch technische disciplines – wordt bijvoorbeeld ook het belang van openheid over de achtergrond van de deskundige genoemd. Heeft de deskundige de juiste scholing en ervaring om de vraagstelling te beantwoorden? Is de deskundige vakbekwaam? Is de deskundige up-to-date? In de huidige rapportages wordt louter verwezen naar de BIG- en NRGD-registraties van de deskundigen. Het is niet bekend of ze geschoold zijn in het afnemen van tests zoals bijvoorbeeld de PCL-R of de HCR-20 V3 of ze recent bijscholing in hun vakgebied hebben gevolgd, enzovoort. Van Esch en collega's (2021) raden aan om als bijlage een verkort cv aan de rapportage toe te voegen. Ook de Hoge Raad oordeelde dat informatie over de opleiding, functies, kennis, ervaring en publicaties van de deskundige wenselijk is.¹⁰

Een ander belangrijk onderwerp waaraan Neal en collega's veel aandacht besteden, is het probleem van bias. Bias is een onvermijdelijk bijproduct van ons menselijk denkproces, ook voor deskundigen zoals forensisch psychologen en psychiaters (Cooper & Meterko, 2019; Dror & Murrie, 2018; Neal, Lienert, et al., 2022). Het is belangrijk dat deskundigen zich bewust zijn van hun vatbaarheid voor bias zodat ze transparant kunnen communiceren over invloeden die hun bevindingen kunnen kleuren en de acties die ze ondernomen hebben om bias te beperken (Neal, Martire, et al., 2022). Ondanks dat bias en het beperken van bias bij forensisch onderzoek in de wetenschappelijke literatuur de laatste jaren veel aandacht heeft gekregen, is het onderwerp niet opgenomen in de NIPP-richtlijn.

CONCLUSIE

Forensisch gedragsdeskundigen die de complexe onderzoeksopdracht in het kader van ontoerekeningsvatbaarheid uitvoeren, worden daarbij ondersteund door richtlijnen opgesteld door beroepsorganisaties. De forensisch psychologische praktijk heeft de afgelopen decennia op dat vlak grote vorderingen gemaakt, zowel internationaal als nationaal. Toch blijkt uit deze gevalstudie dat de adviezen en kwaliteitseisen uit de wetenschappelijke literatuur niet altijd hun weg vinden naar de Nederlandse praktijk. Een eerste stap om daarmee om te gaan, is het verder aanscherpen en uitbreiden van de Nederlandse richtlijn voor forensisch psychologische onderzoeken en rapportages op basis van de internationale *best practices*. Om een effect in de praktijk terug te zien, is het echter noodzakelijk dat de rapportages actief worden getoetst op het naleven van de richtlijnen, bijvoorbeeld bij herregistraties voor het NRGD of door systematisch elk rapport aan een kwaliteitsbeoordeling te onderwerpen zoals dat in Noorwegen gebeurt (Løvgren et al., 2022). Voor nu kunnen juristen en gedragsdeskundigen kritisch toezien op de kwaliteit van forensisch-gedragsdeskundige rapportages met checklists (zie Van Esch et al., 2021; Neal, Martire, et al., 2022). Een rapport dat onvoldoende conform de checklist is, zou niet aan het oordeel van de rechter mogen bijdragen.

OVER DE AUTEURS

Dr. Tamara L.F. De Beuf is postdoctoraal onderzoeker aan de Faculteit Rechtsgeleerdheid en Criminologische Wetenschappen, Katholieke Universiteit Leuven en de Faculteit Psychologie en Neurowetenschappen, Universiteit Maastricht. E-mail: tamara.debeuf@kuleuven.be.

Prof. dr. Peter J. van Koppen is emeritus hoogleraar aan de Faculteiten der Rechtsgeleerdheid, Vrije Universiteit Amsterdam en Universiteit Maastricht. E-mail: p.j.van.koppen@vu.nl.

Verantwoording: Wij ontvingen de besproken rapporten en een aantal andere stukken van de strafzaak tegen Thijs Hermans van diens toenmalige advocaat mr. Job Knoester. Wij hebben gesproken met Thijs noch enig eigen onderzoek aan hem gedaan. Wij kregen van zijn ouders akkoord voor de passages in het artikel met privacygevoelige informatie. Wij danken Robert Horselenberg, Roosmarijn van Es en Vivienne de Vogel voor hun commentaar op een eerdere versie van dit artikel.

Summary

TESTING THIJS: A CASE STUDY OF FORENSIC PSYCHOLOGICAL TESTING

T.L.F. DE BEUF & P.J. VAN KOPPEN

The questions to be answered in forensic mental health reports are challenging. Experts have the task to assess a defendant's past as well as their current and future mental state. To support forensic mental health experts in that complex task, best practices have been developed nationally and internationally. In the Netherlands, there are guidelines on forensic psychological assessment and reporting, including guidelines on forensic psychological testing. In this study we aimed to understand the extent to which these best practices are applied in a Dutch criminal case that has received considerable attention among criminal justice and forensic mental health professionals. Central to the case is the question whether the defendant can be held criminally responsible. With three comprehensive forensic mental health reports in the criminal file, our case study offers elements that can be improved in these kinds of assessments.

Literatuur

- Arntz, A., Kamphuis, J.H. & Derks, J.L. (2018). Gestructureerd klinisch interview voor DSM-5 Syndroomstoornissen Vragenlijst (SCID-5-SV). Boom.
- Clercq, M. De & Laenen, F. Vander (2019). The use of test instruments in forensic report writing: Explorative research in Flanders. *International Journal of Law and Psychiatry*, 66, 1014-54.
- Cooper, G.S. & Meterko, V. (2019). Cognitive bias research in forensic science: A systematic review. *Forensic Science International*, 297, 35-46.
- Drees, M., Huisman, C., Jong, J. de, Olde Keizer, M., Rodenboog, S., Tongeren, P. van, Wildt, K. de & Wolf, M. van der (2020). De gedragsdeskundige contra-expertise in straf- en tbs-zaken: Problemen en oplossingen. *Expertise en Recht*, 5, 178-187.
- Dror, I.E. & Murrie, D.C. (2018). A hierarchy of expert performance applied to forensic psychological assessments. *Psychology, Public Policy, and Law*, 24, 11-23.
- Esch, C. van (2012). *Gedragsdeskundigen in strafzaken*. Van Gorcum (diss. Universiteit Leiden).
- Esch, C. van, Hummelen, J.W. & Blaauw, E. (2021). *Het kritisch beoordelen van gedragsdeskundige rapportages in strafzaken*. Paris.
- Evers, A., Lucassen, W., Meijer, R. & Sijtsma, K. (2010). COTAN beoordelingsstelsel voor de kwaliteit van tests. Nederlands Instituut van Psychologen (<https://psynip.nl/wp-content/uploads/2021/11/COTAN-Beoordelingsstelsel-2010.pdf>).
- Garb, H.N., Lilienfeld, S.O. & Fowler, K.A. (2012). *Psychological assessment and alinical judgment* (3rd ed.). Routledge.
- Guarnera, L.A. & Murrie, D.C. (2017). Field reliability of competency and sanity opinions: A systematic review and meta-analysis. *Psychological Assessment*, 29, 795-818.
- Guarnera, L.A., Murrie, D.C. & Boccaccini, M.T. (2017). Why do forensic experts disagree? Sources of unreliability and bias in forensic psychology evaluations. *Translational Issues in Psychological Science*, 3, 143-152.
- Haan, B. (2022). De vernietigende psychose van Thijs Hermans. *NRC Handelsblad* (17 september), 1.
- Haan, B. (2023a). Advies: Veroordeling van Thijs H. in stand houden. *NRC Handelsblad* (21 juni), 8.
- Haan, B. (2023b). Veroordeelde Thijs Hermans, die in 2019 een drievoerdige moord pleegde, betwist onder meer de onafhankelijkheid van het derde deskundigenrapport. *NRC Handelsblad* (12 juli), 10.
- Heilbrun, K., Yasuhara, K., Shah, S. & Locklair, B. (2021). Approaches to violence risk assessment: Overview, critical analysis, and future directions. In K.S. Douglas & R.K. Otto (Eds.), *Handbook of violence risk assessment* (2nd ed., pp. 3-27). Routledge.
- Koppen, P.J. van (2004). Weg van de toerekeningsvatbaarheid: Over rapportages over de verdachte. *Trema*, 27, 221-228.
- Lovgren, P.J., Laake, P., Reitan, S.K. & Narud, K. (2020). Assessing psychotic symptoms in forensic evaluations of criminal responsibility: A pilot study using Positive And Negative Syndrome Scale. *Journal of Forensic Psychiatry and Psychology*, 31, 490-502.
- Lovgren, P.J., Laake, P., Reitan, S.K. & Narud, K. (2022). Use of assessment instruments in forensic evaluations of criminal responsibility in Norway. *BMC Psychiatry*, 22, 235.
- Martire, K.A. & Neal, T.M.S. (2022). *Rigorous forensic psychological assessment practices* (Report 1). The Joint Federal/Provincial Commission into the April 2020 Nova Scotia Mass Casualty (<https://masscasualtycommission.ca/files/commissioned-reports/COMMO058934.pdf?t=1655479169>).
- Merckelbach, H. & Dandachi-FitzGerald, B. (2021). Symptoomvaliditeits-tests in de rechtszaal. *Expertise en Recht*, 5, 105-111.
- Mossman, D. (2013). When forensic examiners disagree: Bias, or just inaccuracy? *Psychology, Public Policy, and Law*, 19, 40-55.
- Murrie, D.C., Boccaccini, M.T., Guarnera, L.A. & Rufino, K.A. (2013). Are forensic experts biased by the side that retained them? *Psychological Science*, 24, 1889-1897.
- Murrie, D.C. & Warren, J. I. (2005). Clinician variation in rates of legal sanity opinions: Implications for self-monitoring. *Professional Psychology: Research and Practice*, 36, 519-524. <http://dx.doi.org/10.1037/0735-7028.36.5.519>
- Neal, T.M.S., Lienert, P., Denne, E. & Singh, J.P. (2022). A general model of cognitive bias in human judgement and systematic review specific to forensic mental health. *Law and Human Behavior*, 46, 99-120.
- Neal, T.M.S., Martire, K.A., Johan, J.L., Mathers, E.M. & Otto, R.K. (2022). The law meets psychological expertise: Eight best practices to improve forensic psychological assessment. *Annual Review of Law and Social Science*, 18, 17.11-17.24.
- Neal, T.M.S., Slobogin, C., Saks, M.J., Faigman, D.L. & Geisinger, K.F. (2019). Psychological assessments in legal contexts: Are courts keeping "junk science" out of the courtroom? *Psychological Science in the Public Interest*, 20, 135-164.
- NIPF (Nederlands Instituut voor Forensische Psychiatrie en Psychologie) (2022a). *NIPF-richtlijn, Forensisch psychologisch onderzoek en rapportage in het strafrecht: Volwassenen en jeugdigen*. Ministerie van Justitie en Veiligheid (<https://www.nifp.nl/documenten/publicaties/2022/04/20/nifp-richtlijn-forensisch-psychologisch-onderzoek-en-rapportage-in-het-strafrecht>).
- NIPF (Nederlands Instituut voor Forensische Psychiatrie en Psychologie) (2022b). *Uitgelicht: Het NIPF in 2021*. Ministerie van Justitie en Veiligheid (<https://www.nifp.nl/documenten/jaarverslagen/2022/05/09/uitgelicht-het-nifp-in-2021>).
- NRGD (Nederlands Register Gerechtelijke Deskundigen). (2021). *Beoordelingskader FPPO*. <https://www.nrgd.nl/deskundigheidsgebieden/fppo-psychiatrie-psychologie/documenten/publicaties/2021/10/28/beoordelingskader-fppo>.
- Peters, M.J.V., Giesbrecht, T. & Jelicic, M. (2017). Forensische psychodiagnostiek: Tests en schalen in de forensische context. In P.J. van Koppen, J.W. de Keijser, R. Horselenberg & M. Jelicic (Eds.), *Routes van het recht: Over de rechtspsychologie* (pp. 935-949). Boom juridisch.
- Peters, M.J.V., Giesbrecht, T., Jelicic, M. & Merckelbach, H. (2008). Psychosegevoeligheid. In T. Giesbrecht, C. de Ruiter & M. Jelicic (Eds.), *Forensisch psychodiagnostisch gereedschap: Malingering, psychopathie, en andere persoonlijkheidstreken*. Harcourt.
- Rassin, E. (2017). Forensisch psychologische rapportages. In P.J. van Koppen, J.W. de Keijser, R. Horselenberg & M. Jelicic (Eds.), *Routes van het recht: Over de rechtspsychologie* (pp. 951-963). Boom juridisch.
- Schimmel, L.M.C. & Koppen, P.J. van (2017). Verdachten testen: Testgebruik in de forensische psychologie. *De Psycholoog*, 52, 34-42.
- Vogel, V. de, Beuf, T. De, Shepherd, S. & Schneider, R.D. (2022). Violence risk assessment with the HCR-20 V3 in legal contexts: A critical reflection. *Journal of Personality Assessment*, 104, 252-264.