Erasmus School of
Social and
Behavioural Sciences

Rechtbank Den Haag
Team Handel - Algemene Zaken
Postbus 20302
2500 EH Den Haag

Datum
April 26, 2022

Onderwerp
Deskundigenrapport (derde versie)
zaaknummer/ rekestnummer
C/09/588009/ HA RK 20-67

Ons kenmerk

Uw kenmerk

Pagina
1/1

Bijlagen
Deskundigenrapport (3-voud)
Brief met reacties (3-voud)
Declaraties deskundigen

Afdeling
Department of Psychology, Education and Child Studies

Bezoekadres
Burgemeester Oudlaan 50
Mandeville Building
Room T13-28

Postadres
Postbus 1738
3000 DR Rotterdam

T +31 10 408 8799
E m.ph.born@essb.eur.nl
W www.eur.nl/essb

Geachte Rechtbank,

In drievoud treft u hierbij het deskundigenrapport (derde versie) aan dat door dr. A.V.A.M. Evers, prof.dr. D.J. Veltman en mijzelf in de functie van door u aangezochte deskundigen is geschreven in zaaknummer/ rekestnummer C/09/588009/ HA RK 20-67 van de Koninklijke Nederlandse Jagers Vereniging en de Koninklijke Nederlandse Schietsport Associatie versus de Staat der Nederlanden.

Naar aanleiding van de reacties van de advocaten van bovengenoemde twee partijen op de tweede versie van het rapport (de versie van 15 november 2021), zijn een aantal wijzigingen doorgevoerd die hebben geleid tot de voorliggende (derde) versie van het rapport, gedateerd 21 april 2022. (Zie ook sectie 'verantwoording van het onderzoek'.)

In een afzonderlijk document (in drievoud bijgevoegd, gedateerd 21 april 2022, en in briefvorm gericht aan Mw. Van Asch) zijn onze reacties op de vragen van de advocaten te vinden.

Bijgevoegd treft u ook volgens afspraak de declaraties van de drie deskundigen aan voor de aanvullende werkzaamheden, te weten het beantwoorden van de brieven van de advocaten van beide partijen en het doorvoeren van hieruit voortvloeiende wijzigingen in het rapport.

Met vriendelijke groet, mede namens dr. A.V.A.M. Evers en prof.dr. D.J. Veltman,

Prof.dr. M.Ph. Born

Datum: donderdag 21 april 2022

Aan: Rechtbank Den Haag/Team handel

Betreft: zaaknummer C/09/588009 / HA RK 20-67

Geachte mw. Van Asch,

In antwoord op uw brief d.d. 17 januari 2022, met aanvullend de beschikking van de rechtbank d.d. 24 februari 2022 met betrekking tot verhoging van het voorschot en het verzoek d.d. 28 maart 2022 tot voortzetting van onze werkzaamheden, berichten wij u als volgt.

Op 14 december 2021 ontvingen wij een brief van de landsadvocaat Pels Rijcken (mr. M. Dijkstra) namens de Staat en op 16 december 2021 een brief van het advocatenkantoor Van Oosten Schulz De Korte (mr. J. Ph. De Korte) namens de Jagersvereniging & KNS over het door ons in november aangeleverde deskundigenrapport met betrekking tot de E-screener. Beide partijen formuleerden een aantal vragen en – deels tegenstrijdige – wensen met betrekking tot wijzigingen in het rapport (verwijderen resp. toevoegen van informatie). De rechtbank gaf desgevraagd als oordeel (17-1-2022) geen aanleiding te zien voor nadere aanwijzingen betreffende de beantwoording door de deskundigen van de aanvullende vragen van beide partijen naast hetgeen vermeld was in de eerdere beschikking d.d. 18 maart 2021 onder 3.6 b. en c. (geen informatie opnemen die kan leiden tot aantasting van de betrouwbaarheid van de E-screener resp. door aanvragers en/of trainingsinstellingen gebruikt kan worden om aanvragers voor te bereiden op de test).

Hieronder volgen puntsgewijs de antwoorden op de vragen en opmerkingen van beide partijen in de volgorde waarin de brieven van de partijen werden ontvangen. In onze antwoorden gebruiken wij de term 'deskundigenrapport' om te verwijzen naar het voorlopige deskundigenrapport van vorig jaar november.

Hoogachtend,

de deskundigen

Prof. dr. M. Ph. Born

Dr. A. V. A. M. Evers

Prof. dr. D. J. Veltman

Brief Pels Rijcken (vertegenwoordigende de Staat) d.d. 14-12-2021 (hierna te noemen: de verweerder)

1. *Algemeen: geschiktheid en wetenschappelijke exclusiviteit van de COTAN-systematiek.*

(i) Wettelijk criterium

In de eerste twee alinea's wordt door de verweerder een verband gelegd tussen de uitkomst en de mogelijke gevolgen van het deskundigenrapport en de vraag of het COTAN-beoordelingssysteem voor de kwaliteit van tests wel een geschikt middel is om de kwaliteit van de E-screener te beoordelen. Het lijkt de deskundigen niet correct om de geschiktheid van het COTAN-beoordelingssysteem ter discussie te stellen in het licht van de uitkomsten van de beoordeling. Bovendien heeft verweerder eerder aangegeven akkoord te zijn met het inzetten van deze methode.

De inhoudelijke argumentatie om de toepasbaarheid van het COTAN-systeem op de E-screener te betwijfelen is naar onze mening onjuist.

Ten eerste wordt door verweerder gesteld dat de validiteit van de E-screener niet hoeft te worden aangetoond, immers objectiviteit wordt gesteld boven statistisch wetenschappelijke bewijsbaarheid (p. 2, 4e alinea). Echter, de E-screener is een instrument bestaande uit 12 onderdelen en een totaalscore. De combinatie van deze onderdelen, de totaalscore en het gebruik van het instrument in de situatie van het toekennen van een wapenverlof zijn nieuw. Op grond van de scores wordt een voor de cliënt (de aanvrager van een wapenverlof) en voor de maatschappij belangrijke beslissing gebaseerd. Het is evident dat hierbij onderzoek noodzakelijk is dat de werkzaamheid van de E-screener kan onderbouwen, hoe lastig uitvoerbaar dit onderzoek ook is. Dit is zowel in het belang van de cliënten als in het belang van de maatschappij en het is onjuist dat aan dit algemene principe wordt getwijfeld. Het voorbeeld dat wordt gegeven met betrekking tot de variabele suïcidaliteit ("Wanneer iemand laat weten zichzelf van het leven te willen beroven is er geen deugdelijke statistiek voorhanden die laat zien ... zal iedereen met gezond die verstrekking stellig achterwege laten en geen wapenverlof verlenen") is niet van toepassing. Zoals te lezen is in het deskundigenrapport hoefde Suïcidaliteit (evenals Psychiatrische opnamegeschiedenis) zich niet meer te "bewijzen" omdat de experts in het vooronderzoek grote overeenstemming vertoonden over de relevantie van deze vraag. Uiteraard dient wel de effectiviteit te worden aangetoond van nieuwe variabelen die voor het eerst in deze context worden gebruikt en, voor zover deze variabelen uit meerdere items bestaan, welke kritische scores het meest valide zijn.

Ten tweede wordt gesteld, dat omdat de E-screener een instrument is voor risico-taxatie en het COTAN-systeem bedoeld is voor psychodiagnostische instrumenten, het COTAN-systeem niet van toepassing is. Ook dit is onjuist: Psychodiagnostiek is een onderzoeksproces gericht op het doen van psychologische uitspraken over personen met veelal als doel voorspellingen over toekomstig gedrag van deze personen mogelijk te maken: Risico-taxatie past perfect in deze definitie.

(ii) Criteriumvaliditeit

Bij dit punt stelt verweerder de vraag of het überhaupt mogelijk is om in het geval van de E-screener onderzoek naar de criteriumvaliditeit te verrichten dat aan de eisen van het COTAN-beoordelingssysteem voldoet. Zoals blijkt uit het deskundigenrapport bestaat er bij de deskundigen veel waardering voor het onderzoek dat door TNO is verricht, maar kent het onderzoek ook een aantal tekortkomingen. De deskundigen geven aan dat onderzoek dat aan alle mogelijke standaarden voldoet waarschijnlijk ondoenlijk is, maar dat betekent nog niet dat dan maar van alle onderzoek moet worden afgezien. Zoals bij het vorige punt aangegeven betreft het

een nieuw instrument en zonder onderzoeksresultaten die enige aanwijzing kunnen geven voor de criteriumvaliditeit (hoeveel er ook op dat onderzoek aan te merken moge zijn) zou het – gelet op de consequenties voor de betrokkenen – onverantwoord zijn om de E-screener af te nemen. Met de teksten die door verweerder worden geciteerd is door de deskundigen bedoeld aan te geven, dat één onderzoek naar de criteriumvaliditeit in dit geval niet voldoende zal zijn gelet op de beperkingen waarmee dergelijk onderzoek te maken heeft. Slechts cumulatie van onderzoeksgegevens zou voldoende ondersteuning voor de criteriumvaliditeit kunnen opleveren. Maar onderzoek naar de criteriumvalidteit om een juiste werking van de E-screener aan te tonen blijft zeer noodzakelijk. Zo zou een herhaling van het huidige onderzoek met een aantal aanpassingen bijvoorbeeld waardevolle informatie, in casu al dan niet een bevestiging van de resultaten, kunnen opleveren. De deskundigen spreken ook geen oordeel uit over het al dan niet afschaffen of opschorten van de E-screener. Zoals in het rapport is te lezen hebben de deskundigen met name bezwaar tegen het gebruik van afzonderlijke persoonlijkheidsschalen als beslissingsgrond ongeacht de score op een of meer andere schalen, omdat de betrouwbaarheid van de meerderheid van deze schalen te kort schiet. Wanneer een instrument volgens het COTAN-systeem op meerdere onvoldoendes uitkomt betekent dit niet dat het instrument onbruikbaar is, maar dat aan de deskundigheid van de gebruiker hogere eisen worden gesteld (COTAN Beoordelingssysteem voor de kwaliteit van tests, p. 3).

(iii) Constructvaliditeit

Verweerder stelt dat de meeste onderdelen van de E-screener afkomstig zijn van instrumenten waarvan de validiteit al eerder is onderzocht. In de rapporten van TNO wordt verwezen naar publicaties waarin de wetenschappelijke kwaliteit (betrouwbaarheid en validiteit) van de opgenomen schalen is beschreven. Echter, TNO geeft zelf geen beschrijving van dit door anderen verrichte onderzoek. Hierdoor is niet na te gaan of dat onderzoek deugdelijk was en/of de in dat onderzoek verkregen validiteitsgegevens generaliseerbaar zijn naar de situatie waarin de E-screener wordt gebruikt. Op één na gaat het om schalen die door anderen dan TNO zijn vertaald en/of bewerkt voor gebruik in Nederland. De enige uitzondering is een door TNO zelf vertaalde schaal ten behoeve van opname in de E-screener. Buitenlands onderzoek naar de begripsvaliditeit (= constructvaliditeit) kan niet automatisch worden gegeneraliseerd naar de Nederlandse situatie vanwege de mogelijke invloed van culturele verschillen en de invloed van vertalingen/bewerkingen. Voor zover door de betreffende auteurs Nederlands onderzoek naar de begripsvaliditeit is verricht wordt dit onderzoek niet beschreven. Ook hierbij geldt dus dat niet duidelijk is in hoeverre specifieke kenmerken van de gebruikte onderzoeksgroepen en de onderzoekssituatie de verkregen gegevens met betrekking tot de begripsvaliditeit hebben beïnvloed en of deze generaliseerbaar zijn naar de groepen en de situatie waarvoor de E-screener is bedoeld. Het door TNO zelf verrichte onderzoek is zoals beschreven in het rapport als 'onvoldoende' beoordeeld. Verweerder verwijst onder dit punt (Constructvaliditeit) ook naar door TNO zelf uitgevoerd betrouwbaarheidsonderzoek. De resultaten van dit onderzoek zijn in het deskundigenrapport besproken, waarbij is vastgesteld dat de betrouwbaarheid voor sommige schalen te laag is (zie pagina's 20-22 van het deskundigenrapport). Ook voor de elders gepubliceerde en in de TNO-rapporten overgenomen betrouwbaarheidscoëfficiënten geldt dat deze niet zonder meer van toepassing zijn op de situatie waarin de E-screener wordt gebruikt. Betrouwbaarheidsgegevens dienen immers te worden vastgesteld in de groepen waarvoor het instrument ook werkelijk wordt gebruikt, omdat in deze groepen de beslissing – in het onderhavige geval over het verlenen van een wapenvergunning – wordt genomen. Het is dan ook van belang om te weten hoe betrouwbaar de informatie is waarop de beslissingen worden gebaseerd.

(iv) Beslisregels: niet statistische afwegingen

De tekst van verweerder onder dit kopje is vaag en de bedoeling is ons niet duidelijk. Als met de tekst wordt gesuggereerd dat men beter groepen van experts zou kunnen inzetten in plaats van de E-screener: het is niet aan de deskundigen om daar iets over te zeggen. Als wordt bedoeld dat meerdere experts de uitslag van de E-screener zouden moeten bekijken, dan lijkt dat een zinvolle suggestie, maar de vraag is daarbij wel wat onder 'experts' wordt verstaan. De E-screener is een complex instrument en in het deskundigenrapport is dan ook de suggestie opgenomen om de E-screener door psychodiagnostisch geschoolde professionals te laten gebruiken. Wanneer daar meerdere psychodiagnostisch geschoolde professionals bij worden betrokken zou dat de kwaliteit van het beslissingsproces kunnen verbeteren.

(iv) Onuitvoerbaar onderzoek

In het eerste punt onder dit kopje vestigt verweerder er de aandacht op dat enkele achtergrondgegevens van de personen die de E-screener invullen (zoals regio, sekse, en leeftijd), die noodzakelijk zijn om te kunnen bepalen of de onderzoeksgroep representatief is, niet gevraagd mogen worden in verband met de AVG. Dit bezwaar van verweerder geldt in ieder geval niet voor de via het onderzoeksbureau geworven deelnemers en de vraag is of de andere deelnemers daar ook geen toestemming voor kunnen verlenen als deze gegevens – geanonimiseerd – alleen voor onderzoeksdoeleinden worden gebruikt.
Het tweede punt betreft het feit dat Trimbos en TNO sommige kritieke grenzen baseren op gegevens uit de wetenschappelijke literatuur en de deskundigen stellen dat niet wordt verantwoord waarom deze grenzen in de situatie van de E-screener geldig zouden zijn. Er wordt niet beschreven voor welke situatie en met behulp van welke onderzoeksgroepen deze grenzen zijn vastgesteld. Deze normen kunnen dan ook niet zonder meer van toepassing worden verklaard op de situatie waarin de E-screener wordt gebruikt. Verweerder stelt dat Trimbos en TNO niet zien hoe zij dit kunnen verduidelijken en dat de deskundigen hier ook geen oplossing voor aandragen. Vooropgesteld dat het niet de opdracht was aan de deskundigen om hierin te voorzien, zou een suggestie kunnen zijn om een op deze specifieke situatie ingerichte standaardbepalingsprocedure op te zetten.
In het derde punt stelt verweerder dat informatie over de representativiteit van de hoog risicogroep nooit kan worden verkregen. Dit is een correcte constatering, dat lijkt inderdaad nagenoeg onmogelijk.
Bij het vierde aandachtspunt stelt verweerder dat de test-hertestbetrouwbaarheid niet kon worden onderzocht, omdat de aanvragers tot op heden slechts maximaal éénmaal de E-screener hebben ondergaan. Deze vaststelling bevestigt inderdaad dat dit onderzoek nog niet heeft plaatsgevonden, maar dat betekent niet dat dit niet had gekund. Aan de aanvragers van een wapenverlof had bijvoorbeeld kunnen worden verzocht om enige tijd later vrijwillig de E-screener nogmaals in te vullen.
In de laatste alinea vraagt verweerder of het nooit verantwoord is – wetenschappelijk, maatschappelijk – om een instrument dat met de geconstateerde problemen kampt te gebruiken bij beslissingen over wapenverloven? Het antwoord op deze vraag is "nee", zie het antwoord onder het kopje criteriumvaliditeit hierboven en het antwoord op punt 3 van Van Oosten Schulz De Korte.

(v) Correcties

Onder dit kopje maakt verweerder twee opmerkingen. De eerste is: "Bij vraag 7.3 merkt TNO op, dat de kandidaten in de hoog risicogroep zijn geselecteerd op een combinatie van tenminste twee

risicofactoren, nooit op maar één." Naar aanleiding van deze informatie is de tekst in het deskundigenrapport bij vraag 7.3.b aangepast. De zin in het deskundigenrapport waarin kritiek wordt geuit op het feit dat selectie op basis van slechts één risicofactor heeft plaatsgevonden is verwijderd.
De tweede opmerking van verweerder is dat sommige berekeningen wel degelijk door TNO zijn uitgevoerd, maar niet zijn gepubliceerd in de documenten die aan de deskundigen ter beschikking zijn gesteld. De deskundigen zijn ervan uitgegaan dat alle beschikbare informatie in de toegezonden documentatie voorhanden zou (moeten) zijn. Op sommige punten is om verduidelijking en extra informatie gevraagd, maar niet op punten waarvan de deskundigen niet het vermoeden hadden dat deze informatie voorhanden was (Overigens hebben de deskundigen veel waardering voor de snelle en coöperatieve manier waarop de vertegenwoordigers van TNO op verzoeken om extra informatie hebben gereageerd). Verweerder vraagt zich af of het verstrekken van deze gegevens alsnog invloed zou kunnen hebben op de beoordeling, met name de beoordeling van de Normen en de beoordeling van de Begripsvaliditeit. Indien gegevens over de representativiteit van de *normgroep in het pilot-onderzoek* beschikbaar zouden zijn, zou dit wel degelijk invloed kunnen hebben op de beoordeling van deze normen. Het zou dan ook helpen als informatie over de standaardmeetfout en/of de betrouwbaarheidsintervallen van de scores beschikbaar zou komen. Voor deze laatste informatie geldt dan wel dat deze op een of andere manier bij de interpretatie van de scores moet worden betrokken. De beoordeling van de *normen ten behoeve van de rapportage* zou kunnen veranderen als meer informatie wordt gegeven in antwoord op de vragen 4.4 t/m 4.7 van het COTAN-beoordelingssysteem. Overigens zijn de deskundigen van mening dat het onwenselijk is dat ten behoeve van de beslissing en van de rapportage aan de cliënt verschillende normen worden gehanteerd. Hoewel het wenselijk is dat gegevens over de factorstructuur worden verstrekt, zal het verstrekken van deze gegevens waarschijnlijk geen invloed hebben op de hoogte van de beoordeling van de Begripsvaliditeit.

2. *Voortgezet gebruik van de e-screener*

Verweerder stelt de vraag of het huidige gebruik van de E-screener overeenkomt met de aanbeveling om bij het gebruik psychologische of psychiatrische experts in te schakelen. In de huidige procedure is nu de mogelijkheid tot contra-expertise ingebouwd; dit betekent dat bij een negatieve uitslag de aanvrager zelf een deskundige moet inhuren om de beslissing aan te vechten. Deze procedure is niet in overeenstemming met de aanbeveling van de deskundigen. De cliënt, in dit geval de aanvrager van een wapenverlof, heeft recht op een (gratis) nabespreking van het rapport met de uitslag met een vertegenwoordiger van de externe opdrachtgever, in dit geval de Staat (zie bijvoorbeeld de Algemene Standaard Testgebruik NIP 2017) (NB. Dit persoonlijk rapport ten behoeve van de cliënt zou dan ook de gegevens moeten bevatten die zijn gebruikt voor de daadwerkelijke beslissing.) Gelet op de aard en complexiteit van de E-screener kan deze nazorg alleen worden verleend door een psychodiagnostisch geschoolde professional, zoals een psycholoog of psychiater. De opmerking "dit gebeurt al", die door verweerder is gemaakt, is dan ook onjuist.

3. *Vertrouwelijkheid*

De deskundigen zijn zich ten volle bewust van het feit dat de waarde van de E-screener zou kunnen worden aangetast door te veel over de inhoud en de werking van de E-screener bekend te maken. Daarom zijn in de eerste versie van het deskundigenrapport (gedateerd 21-09-2021) een aantal wijzigingen doorgevoerd. Deze wijzigingen betroffen vooral het schrappen van de namen

van de meeste schalen die in de E-screener zijn opgenomen, het aantal items per schaal en de termen waarmee twee groepen variabelen in het beslisproces worden onderscheiden. Een verantwoording van deze wijzigingen staat onderaan pagina 6 en bovenaan pagina 7 van de versie die het resultaat was van deze wijzigingen (gedateerd 09-11-2021) en die hier ter bespreking voorligt. De waarschuwing met betrekking tot de vertrouwelijkheid van het deskundigenrapport betrof de eerste versie van het deskundigenrapport en is abusievelijk in de herziene versie blijven staan. De betreffende zin is nu uit het deskundigenrapport verwijderd.

(i) Vertrouwelijkheid van de bouwstenen van de e-screener
In deze alinea wordt verzocht om de namen van de schalen, de bronnen, de psychometrische gegevens en de beslisregels uit het deskundigenrapport te verwijderen teneinde oefenen en of faken van de E-screener te voorkomen.
Namen van de schalen: De meeste schalen die in de E-screener zijn opgenomen worden in het deskundigenrapport niet met name genoemd. Zoals op pagina 7 van het rapport wordt vermeld is dit met uitzondering van Psychiatrische opnamegeschiedenis, Suïcidaliteit, Alcohol- en drugsgebruik, Psychose (in combinatie met medicijngebruik), Emotionele stabiliteit en Sociale wenselijkheid. De eerste vier betreffen schalen of variabelen die in het verleden ook al in de procedure voor de aanvraag van een wapenvergunning zijn gebruikt (en die dus bekend mogen worden verondersteld). Wat betreft Emotionele stabiliteit en Sociale wenselijkheid zou de bespreking in het deskundigenrapport onmogelijk zijn zonder de namen van de schalen te noemen en/of uit de bespreking zou de inhoud van de schalen zijn af te leiden (Emotionele stabiliteit blijkt hoog met veel andere schalen te correleren en heeft een minimale bijdrage aan de gewogen somscore: het is niet mogelijk deze resultaten te duiden zonder de naam van de schaal te noemen; Sociale wenselijkheid heeft een speciale positie in de procedure: ook dit kan niet worden besproken zonder de naam van de schaal te noemen). Overigens worden de namen van alle schalen in het persoonlijk rapport vermeld dat alle cliënten na het invullen van de E-screener ontvangen. De namen van de schalen kunnen dus reeds bekend worden verondersteld.
Bronnen: We nemen aan dat hiermee wordt bedoeld dat een referentie naar de vragenlijst of de publicatie waaraan de schaal is ontleend in het deskundigenrapport is opgenomen. Dat is niet het geval; de herkomst van de schalen is dus op basis van het deskundigenrapport niet te traceren. Zo komt bijvoorbeeld de schaal Psychose in veel vragenlijsten voor en is uit de naam van de schaal niet te herleiden uit welk instrument deze schaal afkomstig is. Een uitzondering vormde de schaal Emotionele stabiliteit, maar de naam van de betreffende vragenlijst waarvan deze schaal een onderdeel is, is nu ook uit het deskundigenrapport verwijderd.
Psychometrische gegevens: Het is niet voorstelbaar hoe het vermelden van bijvoorbeeld betrouwbaarheidscoëfficiënten of de correlaties met andere vragenlijsten kunnen leiden tot het oefenen met of het faken van de E-screener. De betrouwbaarheidscoëfficiënten die in het deskundigenrapport worden vermeld zijn op de door TNO verzamelde gegevens berekend; het betreft niet de coëfficiënten uit de bronpublicaties. Met behulp van deze gegevens zijn deze schalen derhalve ook niet in literatuur te identificeren. Afgezien van deze overwegingen: de psychometrische gegevens zijn de basis voor het bepalen van het oordeel over de kwaliteit van de E-screener. Zonder vermelding van deze gegevens zou niet duidelijk zijn waarop het oordeel van de deskundigen is gebaseerd.
Beslisregels: De beslisregels zijn essentieel voor het gebruik van de E-screener. Bij een bespreking van bijvoorbeeld de Normen is het onmogelijk om niet nader in te gaan op de inhoud van de beslisregels. Overigens wordt door de deskundigen met opzet bij geen enkele schaal de ruwe score vermeld die de kritische grens vormt tussen het afwijzen of toekennen van een wapenvergunning. Ook daar waar de 95%/5%-grens wordt genoemd is hier geen ruwe score aan

te koppelen. Voor zover cliënten zouden willen oefenen of zouden willen faken is er dus geen direct aanknopingspunt met betrekking tot een score waar zij naar zouden moeten streven of die zij zouden moeten vermijden.
De conclusie van de deskundigen is dat de in het deskundigenrapport vermelde gegevens nodig zijn voor een verantwoorde bespreking van de kwaliteit van de E-screener. Tevens zijn zij van mening dat de vermelde gegevens, in het licht van wat reeds over de E-screener bekend is, geen invloed zullen hebben op de mogelijkheid van cliënten om zich voor te bereiden op de invulling van de E-screener. Op dit punt zien de deskundigen dan ook geen noodzaak verdere wijzigingen in de inhoud van het deskundigenrapport aan te brengen.

(ii) Vertrouwelijkheid van persoonsgegevens
In deze alinea wordt door verweerder verzocht de namen van medewerkers van Trimbos en de Politieacademie te anonimiseren. De namen van medewerkers van de Politieacademie waren echter al zwart gelakt. Het verzoek om namens Trimbos de naam van hun vertegenwoordiger en de inhoud van **Bijlage 1** te verwijderen is nieuw. Hoewel hier niet eerder door de betreffende vertegenwoordiger om is gevraagd, de betreffende persoon ervan op de hoogte was dat het interviewverslag in het rapport zou worden opgenomen, de tekst van het interview (dat wil zeggen voor zover het de antwoorden betreft) door deze persoon zelf is opgesteld en dit verzoek ook geen deel uitmaakt van de beschikking van de Rechtbank uit maart 2021, is aan dit verzoek tot verwijdering gevolg gegeven. Hierbij wordt door de deskundigen aangetekend dat een argument dat pleit tegen verwijdering van de tekst is dat in dit interview enkele relevante kritische kanttekeningen bij de E-screener worden gemaakt. In het deskundigenrapport is nu de naam van de vertegenwoordiger van het Trimbos zwartgelakt en de door deze persoon gegeven antwoorden zijn uit **Bijlage 1** verwijderd (waarmee overigens ook de al eerder zwartgelakte namen van medewerkers van de Politieacademie zijn verwijderd). Net zoals voor **Bijlage 2** geldt hierbij dat de tekst van de vragen is gehandhaafd en dat de inhoud van de gegeven antwoorden elders in het deskundigenrapport wel wordt gebruikt onder verwijzing naar het interview.

4. *Bijlage 2*

In deze bijlage wordt door verweerder concreet aangegeven wat er uit het rapport verwijderd zou moeten worden. Bij de bespreking van het punt "Vertrouwelijkheid van de bouwstenen van de E-screener" hierboven is al aangegeven dat wat de deskundigen betreft er op een enkel punt na geen verdere wijzigingen in het rapport noodzakelijk of gewenst zijn. Geen van de in deze bijlage gewenste veranderingen is dan ook doorgevoerd, met uitzondering van het zwart lakken van de naam van de vertegenwoordiger van het Trimbos instituut.
Nog enkele opmerkingen bij de drie aandachtspunten van verweerder:
Ad. 1. Verweerder stelt onder dit punt dat in het deskundigenrapport teveel over de werking van de E-screener wordt prijs gegeven zodat een deskundige (N.B. "deskundige" hier te onderscheiden van de deskundigen die dit rapport hebben opgesteld) zou kunnen herleiden hoe de E-screener zou kunnen werken en waarmee een kandidaat groen of rood zou scoren. Hierbij wordt door verweerder onder andere de term knockout-variabele gebruikt. In het deskundigenrapport wordt deze term juist met opzet nergens genoemd om de cliënt geen inzicht te geven in de speciale status van dergelijke variabelen; deze speciale status wordt hier door verweerder nu juist wel prijsgegeven. Om diezelfde reden, zoals hierboven onder (i) Beslisregels al is aangegeven, worden in het deskundigenrapport geen formules voor het berekenen van de totaalscores en geen concrete aftestgrenzen vermeld; ook door "deskundigen" is dus niet te achterhalen wanneer een kandidaat precies groen of rood zal scoren.

Ad. 2. Onder dit punt stelt verweerder: "Bepaalde schalen staan weliswaar in een Kamerbrief en andere documentatie ..." en "Anderzijds staan de risicofactoren ook op de rapportage die de aanvragers krijgen....". Feitelijk wordt hier aangegeven dat het niet noemen van de schalen een wassen neus is, omdat deze ook in antwoord op Kamervragen zijn vermeld en deze ook op het rapport aan de cliënt staan. Deze zijn dus al bekend. Verder stelt verweerder dat door het vermelden van de psychometrische gegevens van de schalen, een specialist deze in de literatuur zou kunnen terugvinden. Zoals hierboven onder #3 (i) Bronnen en #3 (i) Psychometrische gegevens is betoogd, is dat volgens de deskundigen niet het geval.
Ad. 3. Hier wordt door verweerder gesteld dat alle namen van experts en andere betrokkenen uit het deskundigenrapport dienen te worden verwijderd. Zoals hierboven onder #3 (ii) Vertrouwelijkheid persoonsgegevens is aangegeven, is aan dit verzoek gevolg gegeven; na het verwijderen van de antwoorden uit het interview met Trimbos hoefde alleen nog de naam van de vertegenwoordiger van Trimbos te worden verwijderd.
Ad Voetnoot: In deze voetnoot wordt verwezen naar twee e-mails van TNO waarin door TNO wordt gewezen op de afspraken die er bestaan tussen TNO en het Ministerie van J&V over de vertrouwelijkheid van de verstrekte informatie. In het bovenstaande is reeds in extenso aangegeven hoe met de verstrekte informatie is omgegaan. Aan het verzoek van TNO om de inhoud van het interview niet op te nemen is gevolg gegeven. De beslisregels zijn ontleend aan andere door TNO toegestuurde documentatie en het bespreken hiervan is zoals eerder vermeld essentieel.

Brief Van Oosten Schulz De Korte (vertegenwoordigende de KJV & de KNSA) d.d. 16-12-2021 (hierna te noemen: de verzoekers)

1. *Vertrouwelijkheid*

   Verzoekers stellen dat het deskundigenrapport niet als vertrouwelijk dient te worden beschouwd binnen de door de rechtbank aangegeven beperkingen, te weten het noemen van betrokkenen bij TNO vanwege eerdere incidenten (bedreigingen), en opnemen van informatie die de validiteit van de E-screener zou kunnen compromitteren. De opmerking over vertrouwelijkheid in het deskundigenrapport was blijven staan uit een eerdere versie van het rapport en is in de huidige versie verwijderd. Vgl. het uitvoerige antwoord aan verweerder #3.

2. *Het noemen van de namen van alle schalen die zijn opgenomen in de E-screener*

   Verzoekers geven aan dat de twaalf schalen in het deskundigenrapport genoemd zouden moeten worden, omdat het publieke informatie betreft. In reactie op dit punt verwijzen wij graag naar ons antwoord aan verweerder #4.

3. *Beroepscode en Algemene Standaard Testgebruik (AST) van het Nederlands Instituut van Psychologen (NIP)*

   Verzoekers zijn van mening dat de E-screener naast toetsing aan het COTAN-Beoordelingssysteem ook aan bijvoorbeeld de Beroepscode van het NIP en de Algemene Standaard Testgebruik (AST) van het NIP getoetst zou moeten worden. Onderscheid moet worden gemaakt tussen de (kwaliteit) van de test zelf en het gebruik van de test. In de beschikking van de Rechtbank wordt aangegeven (onder 2.4) dat de Rechtbank beoogt zoveel mogelijk tegemoet te komen aan de wens van verzoekers om de E-screener te beoordelen volgens COTAN-maatstaven voor de

kwaliteit van tests en eventuele vergelijkbare normen. Hierbij gaat het dus uitdrukkelijk om de kwaliteit van de test, i.c. de E-screener; andere criteria worden in de beschikking van de Rechtbank niet genoemd. In de inleiding van het COTAN-beoordelingssysteem wordt wel naar de AST verwezen en in een van de vragen van het beoordelingssysteem (vraag 3.7) wordt gevraagd of eisen worden gesteld aan de deskundigheid van de testgebruiker. Er is uiteraard een verband tussen de aard van de test en het testgebruik. In het algemeen kan worden gesteld dat hoe complexer de test, hoe belangrijker de beslissingen die met de test worden genomen en hoe meer de kwaliteit van de test te wensen overlaat, des te hogere eisen er aan de deskundigheid van de testgebruiker worden gesteld. In dit verband is met name de volgende passage uit de AST van belang: "Maar dat wil niet zeggen dat instrumenten die op een bepaald criterium geen 'voldoende' beoordeling hebben verworven niet zinvol en verantwoord zouden kunnen worden gebruikt .... De psycholoog dient zich bewust te zijn van eventuele onvolkomenheden van het instrument .... en er bij de interpretatie van de resultaten rekening mee te houden" (AST NIP, 2017, p. 25). Het is om deze redenen dat naar de AST is verwezen, maar het ligt buiten de opdracht aan de deskundigen om het gebruik van de E-screener langs de meetlat van de AST en of de Beroepscode van het NIP te leggen.

4. *Toepassen van E-screener voor het aanvragen versus verlengen van een wapenvergunning*

Verzoekers geven aan dat de E-screener zowel bij eerste aanvragen als aanvragen voor verlenging van een wapenvergunning gebruikt is, en verder dat de toepassing ervan bij aanvragen voor verlenging vanaf 29 oktober 2019 is opgeschort. In het deskundigenrapport wordt hierin inderdaad geen onderscheid gemaakt, 'aanvrager' kan zowel op een eerste aanvraag als een verlenging betrekking hebben aangezien de E-screener bij beide wordt gebruikt. De vraag van verzoekers om ook het historische gebruik van de E-screener te toetsen valt buiten het bestek van de opdracht aan deskundigen (vgl. antwoord op vraag #3 van verzoekers).

5. *Lijsten risicofactoren*

De verzoekers vragen of de deskundigen de lijsten van mogelijke risicofactoren hebben opgevraagd. Deze lijsten werden door de deskundigen opgevraagd bij de vertegenwoordiger van Trimbos, die aangaf dat deze niet konden worden geleverd. Deze partij gaf aan dat de lijsten onderweg tijdens het project nogal eens werden aangepast en niet goed zijn terug te vinden.

6. *E-screener of pre-screener?*

Verzoekers wijzen op een onjuistheid bij het gebruik van de begrippen E-screener resp. pre-screener. Deze terminologie is in het deskundigenrapport aangepast.

7. *Nieuwe rekenregels*

Onder punt 7 wordt gevraagd naar de noodzaak en de reden van de nieuwe rekenregels ingevoerd op 13 januari 2020. De reden voor het wijzigen van de rekenregels worden niet expliciet vermeld door TNO. Hierbij kunnen echter de volgende opmerkingen worden gemaakt. De wijzigingen betreffen alleen de samenstelling en de berekening van de gewogen somscore. Bij deze herberekening in de vorm van een regressieanalyse zijn uit de gewogen somscore drie schalen verwijderd die ook bij eerdere regressieanalyses al een lage – niet-significante – bijdrage aan de somscore leverden. Vanwege het feit dat de bijdrage van deze schalen niet significant is

zouden zij al niet in de somscore opgenomen moeten zijn. Een van deze drie schalen betrof bovendien de schaal Sociale Wenselijkheid die volgens de deskundigen niet in de somscore thuishoort maar de functie van een controle-variabele zou moeten hebben. Kortom, men zou kunnen stellen dat de wijzigingen deze twee onvolkomenheden in feite hebben hersteld. Omdat de bijdrage van de drie schalen klein was, is het effect van hun verwijdering ook beperkt. Dit blijkt onder meer uit het feit dat de sensitiviteit, de specificiteit en het contrast van de gewogen somscore voor en na de wijzigingen nauwelijks verschillen (zie pagina 19 van het deskundigenrapport), waarbij wel de specificiteit (in dit geval de mate waarin aanvragen voor een wapenvergunning terecht worden gehonoreerd) toeneemt van 80% naar 84%.

8. *Zijn psychose en eerdere psychiatrische opname knock-outfactoren?*

Eerdere psychiatrische opname is een knock-outfactor, net als suïcidaliteit en psychopathie. Psychose is dat alleen in combinatie met medicijngebruik. In het rapport is erop gewezen (p. 21) dat dit laatste problematisch kan zijn omdat de psychiatrische literatuur suggereert dat met name *onbehandelde* psychose een risicofactor vormt (vgl. de casus Tristan van der V.).

9. *E-mailwisseling met TNO*

Onder punt 9 wordt gevraagd naar de referentie van een e-mailwisseling tussen een van de deskundigen en een van de vertegenwoordigers van TNO. Deze e-mailcorrespondentie heeft tussen 17 en 20 september 2021 plaatsgevonden. De e-mails betroffen verduidelijking van de werking van de E-screener. De referentie naar deze correspondentie is naar aanleiding van de vraag van verzoekers op pagina 5 en pagina 8 van het deskundigenrapport opgenomen. Tevens wordt nu op specifieke punten in het deskundigenrapport naar deze correspondentie als Document 20 verwezen. Voor de inhoud van deze correspondentie geldt echter dat deze in verband met de door TNO gewenste vertrouwelijkheid niet letterlijk in het deskundigenrapport kon worden opgenomen. Bij dit punt wordt door verzoekers de naam van de betreffende vertegenwoordiger en het exacte tijdstip van een van de e-mails genoemd. Hierbij dient te worden aangetekend dat deze informatie niet van de deskundigen afkomstig is aangezien beide zaken niet in het deskundigenrapport worden vermeld.

Aan: Rechtbank Den Haag
Betreft: zaaknummer/rekestnummer C/09/588009 / HA RK 20-67
Datum: 21-4-2022

Rapport

In opdracht van de Rechtbank Den Haag, in de zaak van de Koninklijke Nederlandse Jagersvereniging (KNJV) en de Koninklijke Nederlandse Schietsport Associatie (KNSA), beide gevestigd in Amersfoort, tegen de Staat der Nederlanden, meer in het bijzonder het Ministerie van Justitie en Veiligheid (MJ&V), gevestigd in Den Haag, hebben ondergetekenden, **prof. dr. M.Ph. Born**, psycholoog, **dr. A.V.A.M. Evers**, psycholoog, en **prof. dr. D.J. Veltman**, psychiater, BIG-geregistreerd onder nummer 79024468401, een onderzoek uitgevoerd naar de z.g. E-screener in gebruik bij beoordeling aanvragen voor wapenverlof.

Inhoudsopgave

Formulering vraagstelling

In de Akte van Uitlating d.d. 18 februari 2021 wordt de volgende vraagstelling geformuleerd:

1. Wat zijn de theoretische achtergronden van de E-screener (wat wordt er gemeten)?
2. In hoeverre is de E-screener psychometrisch betrouwbaar?
3. In hoeverre is de E-screener psychometrisch gevalideerd?
4. a. Hoe is de cesuur bepaald (dat wil zeggen hoe is bepaald wanneer iemand voldoet of niet voldoet)?
   b. Hoe werkt de cesuur?
   c. Hoe is men tot de vaststelling van deze cesuur gekomen?
5. Voldoet deze vorm van afname (in hoeverre is er een kans op 'faken')?
6. In hoeverre is de vragenlijst mogelijk niet fair voor bepaalde groepen (bijvoorbeeld ouderen – jongeren)?
7. Hoe beoordeelt u alle aspecten van de wijze van samenstelling en verzameling van gegevens van de 'hoogrisicogroep', al dan niet middels de 'pre-screener'? Welke invloed hebben samenstelling en verzamelde gegevens van de 'hoogrisicogroep' op de resultaten van de E-screenertest?
8. Welke wijzigingen zijn er na 13 februari 2018 precies aan de E-screener toegevoegd, door wie en wanneer? Welke wijzigingen aan de E-screener heeft de Staat zelfstandig toegevoegd en wanneer zijn die doorgevoerd? Hoe verhouden deze wijzigingen zich met de conclusies en aanbevelingen van de TNO-managementrapportage van 13 februari 2018? Welke invloed heeft ieder van deze wijzigingen (per wijziging) gehad op de resultaten die de E-screener vanaf 1 oktober 2019 heeft gegenereerd?
9. Voldoet de E-screener aan het COTAN beoordelingssysteem voor de kwaliteit van tests (2010) en eventuele vergelijkbare normen?
10. Na verloop van welke termijn kan een E-screenerresultaat redelijkerwijs zijn werking verloren hebben?
11. Welke gegevens bevat de rapportage van een E-screener resultaat, ofwel de 'output' van de gemaakte test die (naar verzoeksters begrijpen) aan het ministerie van Justitie en Veiligheid wordt gezonden?
12. Is de E-screener, gemeten naar professionele maatstaven, een deugdelijk hulpmiddel om de mogelijke aanwezigheid van risicofactoren te detecteren?
13. Is het in uw visie mogelijk verschillen in relevantie aan te geven tussen de risicofactoren die de E-screener onderzoekt? Verschillen deze risicofactoren in mate van voorspelbaarheid van incidenten bij wapengebruik?
14. Is een negatieve score op de met behulp van de E-screener afgenomen test, bezien tegen de achtergrond van het antwoord op de vorige vraag, een deugdelijke reden om te zeggen dat er (geringe) twijfel mogelijk is aan het verantwoord zijn van vuurwapenbezit van de aanvrager?
15. Heeft u verder nog aan- of opmerkingen met betrekking tot de E-screener?

In het navolgende zullen deze vragen puntsgewijs worden beantwoord, waarbij de beantwoording van vraag 9 in een apart hoofdstuk is ondergebracht, voorafgaand aan de beantwoording van de verdere vragen.

Overzicht van ter beschikking gestelde gegevens

Schriftelijk dossier, eerste katern:
5-2-2020: verzoekschrift van Oosten Schultz de Korte namens KNJV/KNSA aan Rechtbank Den Haag tot het bevelen van een voorlopig deskundigenbericht, met toegevoegd:

- 23-12-2019: dagvaarding kort geding eisers,
- 8-7-2015: Akte van statutenwijziging KNJV;
- 2016/7: emailwisseling MV&J betr. opdracht kalibratie E-screener (2x);
- 26-10-2016: Offerteaanvraag MV&J kalibratie E-screener;
- 9-1-2017: nota MV&J, aankondiging demonstratie E-sreener;
- 19-1-2017: brief staatssecretaris MV&J aan KNSA;
- 24-1-2017: voorstel validatie E-screener (grotendeels weggelakt);
- 15-6-2017: Akte van statutenwijziging KNSA;
- 29-6-2017: brief staatssecretaris MV&J aan KNSA;
- 2017: emailwisseling TNO betr. validatie/ kalibratie E-screener (2x);
- 31-10-2017: brief MV&J aan Directoraat-Generaal Politie betr. kalibratie E-screener
- 2018: emails TNO betr. rapport validatie E-screener
- 12-3-2108: Communicatiestrategie Plan van aanpak
- 18-9-2018: Samenwerkingsovereenkomst MV&J en KNSA
- 11-12-2018: notitie digitale inclusie Min. Binnenlandse Zaken en Koninkrijksrelaties
- 7-10-2019: brief politie, uitnodiging E-screener
- 9-10-2019: brief Minister J&V aan KNJV
- 10/11-2019: mediaberichten betr. E-screener
- 23-10-2019: brief politie, intrekking wapenverlof
- 24-10-2019: brief politie, intrekking wapenverlof
- 24-10-2019: Kamervragen CDA betr. E-screener, beantwoording 14-11-2019
- 15-11-2019: emailwisseling TNO KNJV
- 5-11-2019: vragen en beantwoording Statencommissie Provincie Limburg betr. e-screening jagers
- 29-10-2019: brief Minister J&V aan Tweede Kamer betr. bijstelling uitvoeringspraktijk E-screener
- 28-10-2019: brief KNSA aan Min. J&V betr. implementatie E-screener
- 19-12-2019: notitie Aangepaste examinering
- Z.d.: Screeningsformulier wapenverlofaanvraag
- 14-1-2020: brief Pels Rijcken namens de Staat betr. kort geding KNJV/KNSA tegen MJ&V, toegevoegd: email TNO en managementsamenvatting evaluatierapport TNO
- 17-1-2019: brief van Oosten Schultz de Korte betr. ter beschikking stellen stukken
- 28-1-2020: Conclusie van antwoord, voorzieningenrechter Rechtbank Den Haag. Toegevoegd: uitspraak Gerechtshof Den Haag betr. civiele procedure in beroep tegen Politieregio Holland Midden inz. schietincident Alphen a/d Rijn in 2011, en uitspraak Hoge Raad in cassatie. Verder: werkinstructie Politie gebruik E-screener, emailwisseling betr. kalibratie E-screener, WOB-verzoek aan MV&J, brief Min. J&V aan Tweede Kamer, uitspraak voorzieningenrechter inz. intrekking wapenverlof d.d. 20-12-2019, toelichting Politie besluit intrekking wapenverlof d.d. 14-1-2020, mail R Meijer d.d. 19 en 24-1-2020, twee uitspraken MJ&V inz. beroep intrekking wapenverlof, afhandeling verzoek inzage deelnemer E-screenertest, beroepscode psychologen, afwijzing aanvraag wapenverlof/jachtakte. Tenslotte de pleitnota's d.d. 28-1-2020 van Oosten Schultz de Korte resp. Pels Rijcken.

Tweede katern:
3-3-2020: zitting Rechtbank Den Haag, verweerschrift Pels Rijcken namens de Staat inz. gelasten deskundigheidsbericht, met toegevoegd:

- 11-2-2020: vonnis Rechtbank Den Haag in kort geding eisers (waaronder KNJV/KNSA) tegen de Staat, met name MJ&V,

- Z.d: CV's enkele wetenschappers;

Derde katern:
16-11-2020: Pels Rijcken aan Rechtbank Den Haag, aanvullende stukken:
- 13-2-2020: TNO-rapport Tussentijdse evaluatie E-screener
- 16-4-2020: Werkinstructie MJ&V aan Politie betr. beoordeling aanvraag wapenverlof

Vierde katern:
26-11-2020: Werkaantekeningen De Korte/Wilts namens van Oosten Schultz de Korte voor zitting Rechtbank Den Haag

Vijfde katern:
18-2-2021: Beschikking Rechtbank Den Haag, beveelt onderzoek door drie deskundigen

Zesde katern:
3-3-2021: Reactie Pels Rijcken aan Rechtbank Den Haag inz. betrokkenheid COTAN en wenselijke anonimiteit TNO-medewerkers

Zevende katern:
4-3-2021: Akte uitlating van Oosten Schultz de Korte

Achtste katern:
18-3-2021: Beschikking Rechtbank Den Haag, benoeming deskundigen en specificering onderzoeksopdracht.

Aanvullende (door TNO en het Trimbos instituut beschibaar gestelde) stukken:
- 30-11-2012: Trimbos instituut, ontwikkeling E-screener voor risicotaxatie bij aanvraag wapenverlof, conceptversie
- 9-7-2013: Trimbos instituut, powerpoint presentatie E-screener wapenverlof
- 16-7-2013: Trimbos instituut, powerpoint presentatie start-up E-screener
- Z.d., Trimbos instituut, powerpoint presentatie betr. ontwikkelde E-screener
- Z.d., Trimbos instituut, mail betr. expertbijeenkomst
- Z.d., TNO/Trimbos instituut, mail betr. bijeenkomst evaluatie E-screener
- Z.d., TNO-rapport Validatie E-screener
- 22-11-2017: TNO-rapport validatie E-screener
- 22-1-2018: Trimbos instituut, Review conceptrapport TNO E-screener
- 2-2018: TNO-rapport Validatie e-screening
- 9-2019: TNO-rapport Inhoudelijke aanpassingen aan de E-screener
- 13-2-2020: TNO-rapport Tussentijdse evaluatie E-screener
- 4-2020: TNO: Update aanpassingen E-screener
- 14-2-2020: Kamerbrief Minister van Veiligheid en Justitie (ter informatie ingebracht door TNO)
- E-screener - technische toelichting op de afleiding van de einduitslag (dit document werd ter beschikking gesteld door TNO naar aanleiding van het interview).
- E-screener - technische toelichting op de afleiding van de einduitslag bijlage 30 augustus 2021 (dit document werd na een verzoek van de deskundigen ter beschikking gesteld door TNO).
- E-mailwisseling met TNO 17-20-september 2021

Verantwoording van het onderzoek

Ter beantwoording van de bovenomschreven vraagstelling is de E-screener geëvalueerd, waarbij ook aanvullend literatuuronderzoek is gedaan en interviews zijn gehouden met de ontwikkelaars (het Trimbos instituut, vertegenwoordigd door [redacted][1]) en de uitvoerders van de validatie van de E-screener (TNO, vertegenwoordigd door [redacted][2]). Het gesprek met [redacted] (Trimbos) vond plaats op 10-8-2021, het gesprek met [redacted] (TNO) op 18-8-2021. Beide gesprekken duurden ongeveer een uur. De gespreksverslagen werden ter aanvulling en eventuele correctie voorgelegd aan de vertegenwoordigers van Trimbos/TNO. De definitieve en geaccordeerde versie van het gespreksverslag met TNO werd ontvangen op 25-8-2021, het verslag van het Trimbos op 10-8-2021. Beide verslagen zijn door de geïnterviewden zelf opgesteld. Het verslag van TNO is door de interviewers ongewijzigd geaccepteerd. Het verslag van Trimbos is op drie punten door de interviewers aangevuld; deze aanvullingen konden niet worden gecheckt met de geïnterviewde omdat deze niet meer bereikbaar was. Hoewel in eerste instantie door de geïnterviewde van Trimbos geen bezwaar werd gemaakt tegen opname van het verslag in dit rapport werd via de advocaat van de Staat (brief d.d. 14-12-2021) verzocht het verslag alsnog uit het rapport te verwijderen. In bijlage 1 vindt men daarom alleen de aan Trimbos gestelde vragen. Het verslag van het interview met TNO is op verzoek van de geïnterviewden eveneens niet opgenomen in dit rapport, omdat hiermee teveel over de inhoud en de werking van de E-screener bekend zou worden. Ook in bijlage 2 vindt men daarom alleen de aan TNO gestelde vragen. Na het interview heeft tussen 17 en 20 september 2021 nog een e-mailwisseling met [redacted] van TNO plaatsgevonden waarin om verduidelijking werd gevraagd van enkele aspecten met betrekking tot het rapport dat wordt opgesteld na afname van de E-screener. De informatie uit beide interviews en de e-mailwisseling hebben geleid tot enkele aanpassingen in de eindversie van het rapport.
Volgens de beschikking (d.d. 21-04-2021) van de rechtbank van Den Haag dienden bij de beoordeling de maatstaven van het beoordelingssysteem van de COTAN (Commissie Test Aangelegenheden van het Nederlands Instituut van Psychologen, NIP) leidend te zijn. De beoordeling is uitgevoerd door prof. dr. M. Born, dr. A. Evers en prof. dr. D. Veltman (hierna te noemen: de deskundigen). Dr. Evers is lid van de COTAN en Prof. Born is voormalig COTAN-lid. Beiden hebben ruimschoots ervaring met de beoordelingssystematiek van de COTAN. Prof. Veltman is als derde deskundige benaderd vanwege zijn psychiatrisch-bestuursrechtelijke expertise (als lid van de Nederlandse Vereniging voor Medisch Specialistische Rapportage, de NVMSR). Ten behoeve van de beoordeling is de E-screener ook door de deskundigen zelf ingevuld. Hoewel gebruik wordt gemaakt van het beoordelingssysteem van de COTAN, is de COTAN zelf niet bij deze beoordeling betrokken. Alle drie de deskundigen leveren hun bijdrage op persoonlijke titel zonder verdere ruggenspraak. Terzijde zij opgemerkt dat beoordelaars bij een beoordelingsprocedure die door de COTAN wordt uitgevoerd anoniem blijven en in eerste instantie ook niet van elkaar weten dat zij dezelfde test beoordelen. Hun beoordelingen worden, nadat consensus is bereikt, geïntegreerd door de senior-redacteur testbeoordelingen (dr. A. Evers was in het verleden senior-redacteur). Op deze procedurele aspecten wijkt de beoordeling van de E-screener dus af van een reguliere testbeoordeling door de COTAN.
Een eerste versie van het rapport werd op 21 september 2021 ingediend bij de Rechtbank Den Haag. Naar aanleiding van een daaropvolgende e-mailwisseling met de rechtbank werd het rapport nogmaals gecheckt op gegevens die de validiteit van de E-screener zouden kunnen compromitteren en werd een aantal aanpassingen doorgevoerd. Deze tweede versie van het deskundigenrapport is gedateerd 9 november 2021. De aanpassingen hadden tot doel de geheimhouding van de inhoud en de beslisregels zoveel mogelijk te waarborgen. Om deze reden werden ook de meeste schalen die in de E-screener zijn opgenomen in deze versie van het rapport niet met name genoemd, met uitzondering van Psychiatrische opnamegeschiedenis, Suïcidaliteit, Alcohol- en drugsgebruik,

[1] Op verzoek van de advocaten van de Staat wordt de naam van de vertegenwoordiger van Trimbos niet genoemd.
[2] Op verzoek van de rechtbank worden de namen van de vertegenwoordigers van TNO niet genoemd.

Psychose (in combinatie met medicijngebruik), Emotionele stabiliteit en Sociale wenselijkheid. De uitzonderingen betreffen schalen of variabelen die in het verleden ook al in de procedure voor de aanvraag van een wapenvergunning zijn gebruikt (en die dus toch al bekend mogen worden verondersteld) en enkele schalen waarvan de namen bij het bespreken van de E-screener niet gemist konden worden. Hier wordt benadrukt dat de beoordeling en conclusies met betrekking tot de E-screener ongewijzigd zijn gebleven. In dit rapport wordt regelmatig verwezen naar door het Trimbos-instituut en TNO beschikbaar gestelde documenten (zie de lijst op voorgaande pagina). Deze documenten zijn niet openbaar en niet door derden te raadplegen. De schrijvers van dit rapport hebben deze verwijzingen niettemin opgenomen, omdat de door hen gedane uitspraken in principe verifieerbaar dienen te zijn. Op grond van de reacties van de advocaten van beide partijen zijn in de tweede versie van het rapport nog een aantal wijzigingen doorgevoerd. Dit heeft geleid tot de voorliggende (derde) versie van het rapport, gedateerd 21 april 2022. Ook deze wijzigingen hadden voornamelijk betrekking op de vertrouwelijkheid van de verstrekte gegevens en hebben niet geleid tot een wijziging van de beoordeling en conclusies met betrekking tot de E-screener.

Beoordeling van de E-screener volgens het beoordelingssysteem van de COTAN

Internationaal gezien is het COTAN-systeem het meest uitgewerkte systeem voor de beoordeling van de kwaliteit van tests (waaronder vragenlijsten). Enigszins vergelijkbaar is het *EFPA Review Model for the description and evaluation of psychological and educational tests* (European Federation of Psychologists' Associations, 2013) dat deels op het COTAN-systeem is gebaseerd. Beoordeling met dit Europese systeem zou dan ook tot vergelijkbare resultaten hebben geleid.

De door het Trimbos-instituut en TNO beschikbaar gestelde en bij de beoordeling betrokken stukken:

Ter beschikking gesteld door het Trimbos-instituut

1. De offerte voor het maken van een E-screener (Word document)
2. Het PSU (project start up) intern document aan de start van een project voor de uitvoering van dat project (Powerpoint)
3. Het eind Rapport E-screener - met relevante bijlagen (PDF)
4. Slotpresentatie over de E-screener voor de opdrachtgever (Powerpoint)
5. De Technische Handleiding voor gebruik van de E-screener (PDF)
6. Instructie conjunctmeting (Word)
7. Instructie conjunctmeting (Powerpoint)
8. De beoordelingsscores van de experts in de ontwikkelingsfase van de E-screener (SPSS-databestand)
9. TNO-rapport Validatie E-screener d.d. 22-11-2017
10. E-mail TNO-Trimbos betreffende bespreking validatierapport
11. Review TNO-rapport over de E-screener

Ter beschikking gesteld door TNO

12. 2017 – 2018: Validatieonderzoek E-screener (zie 18R10219) en meegestuurde Managementsamenvatting
13. 2019: Afstelling E-screener (reken- en beslisregels) op basis van het onderzoek en beleidskeuzes binnen een werkgroep J&V/Politie (zie 19R11285; bevat ook de E-screener zelf)
14. 2019 – 2020: Monitoring E-screener (tussentijdse data-analyses, niet allemaal meegestuurd) en bijstelling op basis van de evaluaties (zie 20R10174; intern memo 2020)
15. 2020 – heden: Structurele monitoring en evaluatie E-screener (en eventuele bijstelling), adviseur Ministerie van J&V

Interviewverslagen en nagestuurde informatie

16. Verslag interview met [redacted] van het Trimbos instituut
17. Verslag interview met [redacted] van TNO
18. E-screener - technische toelichting op de afleiding van de einduitslag (dit document werd ter beschikking gesteld door TNO naar aanleiding van het interview)
19. E-screener – Technische toelichting op de afleiding van de einduitslag Bijlage 30 augustus 2021 (dit document werd ter beschikking gesteld door TNO na een verzoek van de deskundigen)
20. E-mailwisseling met TNO 17-20 september 2021

De deskundigen hebben de indruk dat zij van alle relevante stukken kennis hebben kunnen nemen. In het hiernavolgende zal soms naar specifieke stukken uit bovenstaande lijst worden verwezen onder gebruikmaking van de nummers 1 t/m 20. Bij een beoordeling volgens het COTAN-systeem staat de informatie die in de testhandleiding wordt vermeld altijd centraal. Hier zullen met name de documenten 3, 5 en 12 tot en met 20 als handleiding worden beschouwd. Hierbij moet worden

aangetekend dat sommige informatie in de chronologisch eerdere documenten is achterhaald door aanpassingen die later hebben plaatsgevonden (zie ook de opmerkingen bij vraag 8 van de vraagstelling van de rechtbank).

Het gehanteerde COTAN-beoordelingssysteem kent zeven criteria, elk bestaande uit een aantal vragen. Deze zeven criteria zijn:
- Uitgangspunten van de testconstructie
- Kwaliteit van het testmateriaal
- Kwaliteit van de handleiding
- Normen
- Betrouwbaarheid
- Begripsvaliditeit
- Criteriumvaliditeit

Op de beoordeling van deze zeven criteria zal hieronder een toelichting worden gegeven. Het beoordelingssysteem kent ook een addendum met betrekking tot fairness, dat bij sommige van de bovengenoemde zeven criteria een rol speelt. Het aspect fairness zal zowel bij de afzonderlijke criteria (indien van toepassing) als samenvattend na het zevende criterium worden besproken. Voor de beoordeling van elk criterium worden in het beoordelingssysteem een aantal vragen gesteld. De beoordeling op elk van deze vragen kan zijn 'onvoldoende', 'voldoende' of 'goed'. De eindoordelen op de criteria kunnen eveneens 'onvoldoende', 'voldoende' of 'goed' zijn. Het vaststellen van de eindoordelen per criterium op basis van de antwoorden op de afzonderlijke vragen gebeurt volgens vaste voorschriften. Bij de totstandkoming van de eindoordelen per criterium in het onderstaande worden deze regels niet expliciet vermeld. Ze zijn echter vermeld in de tekst van het beoordelingssysteem. Het beoordelingssysteem is in te zien en te downloaden via de website van het Nederlands Instituut van Psychologen NIP (zie psynip.nl).

Uitgangspunten van de testconstructie

Vraag 1.1.a: Is er aangegeven welk(e) construct(en) de test beoogt te meten?

Het te ontwikkelen instrument meet de psychische gesteldheid van de aanvrager van een wapenverlof, waarbij het de bedoeling is om specifieke combinaties van risicofactoren of cumulatie van specifieke risicofactoren voor legaal wapenbezit op te sporen (document 5, p.5). Welke risicofactoren en daderkarakteristieken zijn opgenomen in de E-screener komt voort uit literatuuronderzoek en de bevraging van experts. In hoeverre deze keuze is verantwoord wordt beantwoord bij vraag 1.2. Hoewel bij de opzet van het instrument wordt aangegeven dat het gaat om combinaties of cumulatie van risicofactoren, wordt dat principe in later onderzoek losgelaten. Uiteindelijk (document 13, p. 18-19) kan een negatief advies immers behalve op de gewogen somscore ook gebaseerd zijn op de scores op afzonderlijke risicofactoren.
Vanwege deze onduidelijkheid wordt vraag 1.1.a met 'voldoende' beantwoord.

Vraag 1.1.b: Is er aangegeven wat de doelgroep(en) is (zijn) van de test?

De E-screener is bedoeld voor de aanvrager van een wapenverlof of jachtakte. Hierbij worden geen verdere condities/beperkingen genoemd.
Vraag 1.1.b wordt met 'goed' beantwoord.

Vraag 1.1.c: Is er aangegeven wat de functie is van de test?

De E-screener is een hulpmiddel voor de politie om een inschatting te maken van mogelijke risico's bij het verlenen van een wapenverlof (document 5, p.6). Hoewel in de documenten 3 en 5 herhaaldelijk wordt benadrukt dat de E-screener een hulpmiddel is en dat de uiteindelijke beslissing

niet louter van de score op de E-screener zou moeten afhangen, is de tekst in document 13 minder terughoudend. Wanneer een van de scores zich in het rode gebied bevindt krijgt de gebruiker (in casu de politiefunctionaris die de aanvraag behandelt) de volgende tekst te zien: "Op basis van de uitkomsten van deze persoon lijkt er een te groot risico te zijn op misbruik van het wapen op basis van zijn of haar psychische gesteldheid. Om deze reden wordt afgeraden om een wapenverlof te verlenen." In het rapport voor de aanvrager zelf verschijnt in dat geval de tekst: "De einduitslag wijst op de mogelijke aanwezigheid van risicofactoren in uw geestelijke gesteldheid die het verantwoord omgaan met wapens of munitie in de weg kunnen staan". Weliswaar wordt de aanvrager gewezen op de mogelijkheid via contra-expertise deze beslissing aan te vechten, maar dat kan (ook financieel) een flinke drempel betekenen. Dit betekent dat de E-screener beschouwd moet worden als een test voor 'belangrijke beslissingen op individueel niveau'. Deze vaststelling is van belang omdat ten aanzien van dit type tests strengere eisen worden gesteld met betrekking tot Normen en Betrouwbaarheid (zie aldaar) dan ten aanzien van tests voor 'minder belangrijke beslissingen op individueel niveau'.
Omdat duidelijker had moeten zijn aangegeven wat het gewicht is van de E-screener bij het nemen van de uiteindelijke beslissing wordt vraag 1.1.c met 'voldoende' beantwoord.

Vraag 1.2: Is de herkomst van het constructie-idee beschreven en/of worden de te meten constructen gedefinieerd?

De inhoud van de E-screener is gebaseerd op bestaande risicotaxatie-instrumenten, literatuurstudie en expertbeoordelingen. De samenstelling van de E-screener is als volgt tot stand gekomen:

- Een klankbordgroep bestaande uit 12 personen die was samengesteld uit deelnemers van de politie die zich routinematig bezighouden met het verstrekken van vuurwapenverloven, deelnemers van de politieacademie die hierin trainingen verzorgen en enkele vertegenwoordigers van justitie en andere organisaties hebben een lijst van risicofactoren opgesteld waarbij zij aan de factoren een score toekenden. Factoren met een score boven een bepaalde kritieke grens werden voorlopig geselecteerd. Uit de hieruit ontstane lijst werden niet-psychische factoren (zoals lichamelijke factoren en factoren betreffende justitiële voorgeschiedenis) geschrapt.
- Aan een expertgroep, bestaande uit 14 deskundigen met verschillende kennis en expertise bestaande uit o.a. trainers van de politieacademie en hoogleraren criminologie, forensische psychologie en forensische psychiatrie, werd een lijst met risicofactoren voorgelegd die op basis van literatuuronderzoek was samengesteld. Ook deze personen moesten de risicofactoren op hun relevantie beoordelen. Wanneer risicofactoren op basis van deze expertoordelen boven een zekere kritische grens scoorden werden ze in een voorlopige selectie opgenomen. Echter, ook enkele risicofactoren die beneden de gestelde grens scoorden (de zgn. pro memorie of PM-factoren) werden in deze voorlopige selectie opgenomen. Tussen de klankbordgroep en deze expertgroep bestond overlap in de vorm van drie personen.
- Vervolgens werd naar de overlap tussen de twee voorlopige selecties gekeken en werd op basis hiervan een nieuwe lijst samengesteld aangevuld met enkele PM-factoren. Hoe de selectie op grond van de twee lijsten (de ene op basis van de klankbordgroepen en de andere op basis van literatuuronderzoek en expertgroep) tot stand is gekomen is niet duidelijk. Zo is bijvoorbeeld niet bekend hoe de twee voorlopige lijsten met hun scores eruitzien, welke de PM-factoren zijn en waarom sommige factoren wel en andere niet als PM-factor werden meegenomen.
- In een volgende fase werd gekeken naar de meetbaarheid van de geselecteerde risicofactoren, dat wil zeggen risicofactoren waarvoor geen gevalideerde vragenlijsten beschikbaar waren werden niet opgenomen. Er wordt geen informatie gegeven over de factoren die in deze fase zijn afgevallen, zodat niet duidelijk is of het hier om mogelijk cruciale variabelen gaat die alsnog meetbaar gemaakt hadden kunnen worden.
- Ten slotte werd in een bijeenkomst bestaande uit negen experts (dit was dezelfde expertgroep als de bovenvermelde expertgroep van 14 personen, waarbij vijf personen niet aanwezig waren) de definitieve lijst van te meten risicofactoren samengesteld. Ook over dit selectieproces wordt

verder niets vermeld, waardoor bijvoorbeeld niet duidelijk is of en welke factoren in deze fase nog zijn afgevallen. Ook [redacted] van het Trimbos instituut kon hier in het interview geen duidelijkheid over verschaffen (zie document 16), de lijsten van mogelijke risicofactoren werden regelmatig aangepast tijdens de ontwikkelingsperiode van de E-screener (die liep van circa februari 2012 tot september 2013) en zijn niet beschikbaar.

De uiteindelijke lijst met te meten risicofactoren bestond uit 12 risicofactoren, waaronder psychiatrische opnamegeschiedenis, suïcidaliteit, alcohol- en drugsgebruik en de combinatie van psychose en medicijngebruik.
In het valideringsonderzoek uitgevoerd door TNO (zie document 12) werden enkele andere vragenlijsten afgenomen. In dit onderzoek bleek de schaal Emotionele Stabiliteit een laag- en hoog-risicogroep goed te kunnen onderscheiden. Daarom is deze schaal, bestaande uit 12 items, later aan de E-screener toegevoegd. Met deze opzet lijkt de E-screener zoals hierboven vermeld goed aan te sluiten bij uit de literatuur bekende gegevens met uitzondering van de rol van cognitieve beperkingen en de combinatie Psychose/Medicijngebruik (zie ook het literatuuroverzicht dat is gegeven bij de beantwoording van vraag 1 uit de vraagstelling van de rechtbank).
De schalen zijn overgenomen uit bestaande (meestal uit het Engels vertaalde) vragenlijsten. Van de te meten constructen worden geen definities gegeven. Wel worden vermeld: de naam van de vragenlijst waaraan de schaal is ontleend, een voorbeelditem en enkele technische gegevens, zoals de gebruikte antwoordschaal en de betrouwbaarheid zoals deze in de literatuur wordt gerapporteerd. Mogelijk worden deze definities in de bronliteratuur wel gegeven, maar dat is niet gecheckt, want de definities horen in principe in de handleiding (of in dit geval de documenten 3, 5, 12, 13 of 14) te staan. Voor zover Nederlands validatie-onderzoek met deze vragenlijsten heeft plaatsgevonden wordt hiervan geen beschrijving gegeven.
De beoordeling van vraag 1.2 wordt 'voldoende'. De E-screener is tot stand gekomen via een degelijk constructieproces, maar verschillende fasen van het proces zijn niet helder en volledig beschreven, met name waar het gaat om de keuze van de te meten constructen. Een manco is ook het ontbreken van uitgebreide definities van deze constructen.
NB. In de eerste fase van de ontwikkeling van de E-screener werden de expertprofielen Gevaar, Braaf en Logisch ontwikkeld. Deze profielen worden in de uiteindelijke versie van de E-screener niet meer gebruikt. Aan deze profielen zal in dit beoordelingsrapport daarom verder geen aandacht worden besteed.

Vraag 1.3: Wordt de relevantie van de testinhoud voor de te meten construct(en) aannemelijk gemaakt?

De vragen met betrekking tot psychiatrische opnamegeschiedenis en medicijn- en drugsgebruik komen uit het inlichtingenformulier aanvraag wapenverlof (IFW) (de tekst van de vraag met betrekking tot de psychiatrische opnamegeschiedenis is in een latere fase gewijzigd). De vragen van alle andere schalen zijn afkomstig uit andere, reeds bestaande, vragenlijsten waarbij de complete schalen werden overgenomen. Dat geldt niet voor de twee vragen met betrekking tot suïcidaliteit, die een selectie vormen uit een bestaande vragenlijst (ook de tekst van een van deze twee items is in een latere fase gewijzigd).
In de beschrijving van het constructieproces van de E-screener wordt niet vermeld hoe men tot de keuze van items in de oorspronkelijke reeds bestaande schalen is gekomen. Het kan zijn dat deze beschrijving wel in de bronpublicaties wordt gegeven, maar dat is door de deskundigen niet gecheckt. Hierbij dient te worden aangetekend dat voldoende/goede betrouwbaarheids- en validiteitsgegevens in deze publicaties niet hoeft te beteken dat de items die worden gebruikt om het betreffende construct te meten ook werkelijk representatief zijn voor dat construct.
De beoordeling van vraag 1.3 wordt 'voldoende' vanwege het ontbreken van gegevens over de constructie van de oorspronkelijke schalen.

Het oordeel over de *Uitgangspunten van de testconstructie* wordt 'voldoende'.

Kwaliteit van het testmateriaal (Afname via de computer)

De E-screener wordt via de computer afgenomen, daarom zijn de vragen 2.9 t/m 2.16 van het COTAN-beoordelingssysteem van toepassing.

Vraag 2.9: Zijn de testopgaven gestandaardiseerd of worden bij adaptieve tests beslisregels geëxpliciteerd?

Er is geen sprake van een adaptieve test. De testopgaven zijn gestandaardiseerd.
De beoordeling van vraag 2.9 wordt 'goed'.

Vraag 2.10: Is er sprake van een geautomatiseerd of objectief scoringssysteem?

Er is sprake van een geautomatiseerd scoringssysteem, waarbij vooraf is bepaald welke score aan welk antwoord wordt toegekend.
Vraag 2.10 wordt met 'goed' beoordeeld.

Vraag 2.11: Zijn de items vrij van racistische, kwetsende inhoud?

Er is geen racistische of anderszins kwetsende inhoud in de items aangetroffen.
Vraag 2.11 wordt met 'goed' beoordeeld.

Vraag 2.12: Is de software zo ontworpen dat fouten door onjuist gebruik kunnen worden vermeden?

De software werkt probleemloos. Het invullen is simpel. Na het aanklikken van een antwoord, komt automatisch meteen de volgende vraag in beeld. Dat werkt efficiënt (zo weinig mogelijk klikken), echter herstel van een eventueel foute klik is niet mogelijk. Er kunnen geen vragen worden overgeslagen. Teruggaan naar een vorige vraag is niet mogelijk.
Vraag 2.12 wordt met 'goed' beoordeeld.

Vraag 2.13: Is de instructie voor de geteste volledig en duidelijk?

In de instructie wordt kort uitleg gegeven over het doel van de vragenlijst en ook wordt vermeld dat het invullen niet aan tijd is gebonden. Er wordt geen voorbeeldvraag gegeven. Er wordt aangegeven dat men het lettertype kan vergroten en/of de vragen kan laten voorlezen. Ook wordt vermeld dat men de aanvragen eerlijk moet invullen, omdat de E-screener anders onbruikbaar wordt en een vergunning kan worden geweigerd. Tenslotte wordt vermeld dat de korpschef over de aanvraag beslist. Wanneer men de voorleesoptie gebruikt moet men overigens wel naar de volgende vraag doorklikken, dat gaat dan niet automatisch. Dit betekent dat men nog tijd heeft om over het aangeklikte antwoord na te denken en dit kan wijzigen (teruggaan naar de vorige vraag is ook in de voorleesmodus niet mogelijk). Het zou nuttig zijn om te weten hoe vaak dit gebeurt, of dit effect heeft op de scores (relatie met impulsiviteit?) en hoe vaak de voorleesoptie wordt gekozen. Overigens dient men bij elk item de voorleesoptie opnieuw aan te klikken.
Vraag 2.13 wordt met 'goed' beoordeeld.

Vraag 2.14: Zijn de items correct geformuleerd?

De items zijn geformuleerd in correct Nederlands en in het algemeen niet te lang of ingewikkeld. Er komen wel enkele mogelijk moeilijke woorden voor. In een item wordt gevraagd of men wel eens misbruik van iemand heeft gemaakt; deze formulering maakt dat het item op verschillende manieren zou kunnen worden geïnterpreteerd, namelijk ook als seksueel misbruik wat niet de bedoeling lijkt.

Een indicatie van het taalniveau dat vereist is voor een goed begrip van de items zou wenselijk zijn. Veel items bevatten een ontkenning. In combinatie met de antwoordschaal (waarbij de negatieve pool uiteraard ook ontkennend is geformuleerd: "oneens") maakt dit het soms lastig om de vragen te beantwoorden. Dergelijke formuleringen zouden moeten worden vermeden. Soms worden er voorbeelden gegeven in de vraag, dit kan sturend werken bij de beantwoording van de vraag. Het voorlezen is kunstmatig met soms merkwaardige klemtonen en pauzes in de tekst, waardoor de tekst aan begrijpelijkheid verliest. Vaak wordt boven een item dezelfde zin herhaald, bijvoorbeeld "Kunt u aangeven of u het eens bent met de volgende uitspraken" of "Hieronder volgt een aantal stellingen ...". Dit is kennelijk overgenomen uit de papieren versie, maar deze teksten zijn in feite overbodig. Verder zijn ze in het meervoud geformuleerd, terwijl op elk scherm maar één vraag wordt gepresenteerd.
Vraag 2.14 wordt met 'voldoende' beoordeeld.

Vraag 2.15: Hoe is de kwaliteit van de vormgeving van de gebruikersinterface?

De schermen zijn rustig, er wordt één vraag gegeven per scherm en de vragen zijn goed leesbaar (en bovendien te vergroten).
Vraag 2.15 wordt met 'goed' beoordeeld.

Vraag 2.16: Is de test voldoende beveiligd?

In document 5 worden de beveiligingsmaatregelen die zijn genomen beschreven. Deze lijken afdoende. Bij het invullen van de vragenlijst bleek dat de vragen (per stuk) met ctrl-c zijn te kopiëren naar een ander bestand. Hoewel de vragen hierdoor bekend zouden kunnen worden, is dit in mindere mate een probleem omdat de E-screener alleen on-site wordt afgenomen (mits voldoende surveillance aanwezig is).
Vraag 2.16 wordt met 'goed' beoordeeld.

Het oordeel over de *Kwaliteit van het testmateriaal* wordt 'goed'.

Kwaliteit van de handleiding

Vraag 3.1: Is er een handleiding beschikbaar?

Er is niet echt sprake van een handleiding. Het enige document waar handleiding op staat (document 5, "Wapenverlof: E-screener voor risicotaxatie. Een handleiding.") is nog slechts deels van toepassing vanwege latere wijzigingen in de E-screener. De documenten 3, 5 en 12 tot en met 19 worden ten behoeve van deze beoordeling gezamenlijk als handleiding beschouwd.
Vraag 3.1 wordt met 'ja' beantwoord.

Vraag 3.2: Zijn de aanwijzingen voor de testleider volledig en duidelijk?

Er worden geen aanwijzingen gegeven voor de testleider hoe de vragenlijst bij een respondent te introduceren. Omdat de wijze waarop de testleider de vragenlijst introduceert invloed kan hebben op de motivatie en de gemoedsgesteldheid van de respondent dient letterlijk te zijn voorgeschreven wat de testleider wel en niet mag en moet zeggen.
Vraag 3.2 wordt met 'onvoldoende' beoordeeld.

Vraag 3.3: Wordt er informatie gegeven over de gebruiksmogelijkheden en beperkingen van de test?

De gebruiksmogelijkheden zijn duidelijk: op basis van de scores op de E-screener kan een beslissing worden genomen over aanvraag of verlenging van wapenverlof, er zijn geen andere

gebruiksmogelijkheden. Er worden geen beperkingen van de E-screener genoemd (bijvoorbeeld met betrekking tot het vereiste taalniveau).
Vraag 3.3 wordt met 'voldoende' beantwoord.

Vraag 3.4: Wordt er in de handleiding een samenvatting van de onderzoeksresultaten gegeven?

In de documenten 12, 13, 14, 18 en 19 worden de onderzoeksresultaten beschreven.
Vraag 3.4 wordt met 'goed' beoordeeld.

Vraag 3.5: Wordt er met behulp van voorbeelden aangegeven hoe testscores kunnen worden geïnterpreteerd?

Er worden geen gevalsbeschrijvingen gegeven.
Vraag 3.5 wordt met 'onvoldoende' beoordeeld.

Vraag 3.6: Wordt er gewezen op soorten informatie die bij de interpretatie van belang kunnen zijn?

In document 15 (p. 16) wordt gesteld dat ook rekening moet worden gehouden met risicofactoren zoals bijvoorbeeld radicalisering, extremisme of een crimineel verleden. Deze risicofactoren zouden vastgesteld kunnen worden door middel van een antecedentenonderzoek, een inlichtingenformulier, het nagaan van opgegeven referenten en huisbezoek in de procesaanvraag voor een wapenverlof. Hoe deze informatie vervolgens een rol kan spelen in combinatie met de scores op de E-screener wordt niet vermeld.
Vraag 3.6 wordt met 'voldoende' beoordeeld.

Vraag 3.7: Wordt de mate van deskundigheid vermeld die vereist is voor afname en interpretatie van de test?

Ten aanzien van het feitelijk gebruik van de E-screener worden twee rollen onderscheiden, die van de Lage en de Hoge Ambtenaar, praktisch gesproken is dit het onderscheid tussen degene die de test afneemt en degene die de scores te zien krijgt en interpreteert. Wat een Lage of een Hoge Ambtenaar is wordt niet omschreven, noch welke (test-technische of psychodiagnostische) kennis de personen in deze rollen zouden moeten bezitten.
Vraag 3.7 wordt beoordeeld met 'onvoldoende'.

Vraag 3.8: Wordt er informatie gegeven over de installatie van de computersoftware?

De installatie van de software, het backoffice en de administratieve systemen worden beschreven in document 5.
Vraag 3.8 wordt beoordeeld met 'goed'.

Vraag 3.9: Wordt er informatie gegeven over de bediening en mogelijkheden van de software?

Zie antwoord vraag 3.8.
Vraag 3.9 wordt beoordeeld met 'goed'.

Vraag 3.10: Zijn er voldoende mogelijkheden voor technische ondersteuning?

Er is geen sprake van een helpdesk. In hoeverre bijvoorbeeld TNO of IPSOS (IPSOS is de beheerder van de online vragenlijst) bereikbaar is of snel kan inspringen wanneer er problemen zijn met de E-screener wordt niet vermeld.
Vraag 3.10 wordt beoordeeld met 'onvoldoende'.

Het oordeel over de *Kwaliteit van de handleiding* wordt 'voldoende'.

Normen

Er is sprake van drie soorten normen:

1. Voor vier variabelen wordt gebruik gemaakt van *domeingerichte of absolute normen*. Een (absoluut) bepaalde score op elk van deze variabelen afzonderlijk leidt tot een negatief (Rood) advies met betrekking tot de aanvraag.
2. Voor de gewogen somscore wordt gebruik gemaakt van *criteriumgerichte normen*. De samenstelling van de gewogen somscore is verschillende keren aangepast maar is uiteindelijk gebaseerd op (de items van) zeven schalen. De grootte van het gewicht dat elke variabele krijgt wordt bepaald door de hoogte van de correlatie met de risicofactor (gecorrigeerd voor overlap met de andere variabelen). De scoregrens voor afwijzing van de aanvraag wordt empirisch bepaald (zie hieronder bij criteriumgerichte interpretatie). Een score boven deze grens leidt tot een negatief (Rood) advies met betrekking tot de aanvraag.
3. Zeven schalen worden ook *als afzonderlijke variabelen* gebruikt voor het bepalen van het advies. Hiertoe worden de scores op deze schalen gedichotomiseerd in de 95% laagste scores en de 5% hoogste scores. Voor het bepalen van deze percentages is een normgroep nodig. Daarom zijn voor de op deze wijze gebruikte variabelen de eisen die gelden voor een *normgerichte normering* van toepassing. Voor elke schaal afzonderlijk, behalve voor de schaal Sociale wenselijkheid, geldt dat personen met de 5% hoogste scores een negatief (Rood) advies krijgen. De hoogst scorende 5% op Sociale wenselijkheid krijgt de uitslag Oranje = onbetrouwbaar (zie document 17, punt 10). De feitelijke keuze voor de 95%/5% grens is gemaakt in een stuurgroepbijeenkomst van TNO en politie (zie document 17, punt 3) en is niet gebaseerd op onderzoek naar de voorspellende waarde of op basis van prevalentiegegevens. Daarmee lijkt de keuze voor de 95%/5%-grens arbitrair (waarom bijvoorbeeld niet 96%/4% of 90%/10%?). Ook spelen vier van de zeven schalen een rol *als onderdeel van de gewogen somscore*. Het gebruik van deze vier schalen in deze rol is niet consequent, want de gewogen somscore is er juist op gericht om een beslissing te nemen op grond van de samenhang tussen alle variabelen waarbij scores op de variabelen elkaar onderling kunnen compenseren, terwijl elke variabele afzonderlijk dus ook tot een negatieve beslissing kan leiden zoals uit het onderhavige punt 3 blijkt (waarvan de juistheid bovendien niet via het berekenen van ROC-curves wordt gecheckt zoals bij de gewogen somscore wel gebeurt).
   Voor deze zeven schalen en voor de gewogen somscore worden in het rapport voor de behandelend politiemedewerker, dat ook naar de aanvrager zelf wordt gestuurd, de percentielscores vermeld waarmee de ruwe scores van de aanvrager overeenkomen. Deze percentielscores zijn echter gebaseerd op een andere normgroep dan die waarop de 95%/5%-dichotomie van de schalen is bepaald. Beide normgroepen worden beschreven bij vraag 4.3. Het feit dat de percentielscores afkomstig zijn uit een andere normgroep heeft als consequentie dat een percentielscore boven de 95% op een van de zeven schalen in het rapport niet noodzakelijk een negatief advies tot gevolg heeft, omdat de bijbehorende ruwe score beneden de eerder bepaalde 95%-grens (in de andere normgroep) kan vallen (mededeling via e-mail van [redacted] van TNO op 17 september2021, document 20). De percentielscore voor de gewogen somscore is ook louter informatief bedoeld en heeft geen relatie met het via criteriumgerichte normering bepaalde afkappunt.

De kwaliteit van de verschillende soorten normen zullen achtereenvolgens worden besproken. De eerste twee vragen zijn voor elk type normen hetzelfde.

Vraag 4.1: Worden normen verstrekt?

Voor de verschillende typen scores worden verschillende typen normen verstrekt.
Vraag 4.1 wordt met 'ja' beantwoord.

Vraag 4.2: Zijn de normen actueel?

De gegevens waarop de normgerichte en criteriumgerichte normen zijn gebaseerd zijn verzameld in 2017 en zijn derhalve actueel te noemen.
Vraag 4.2 wordt met 'goed' beoordeeld.

Normgerichte interpretatie

Deze normen hebben betrekking op de zeven schalen waaraan werd gerefereerd bij punt 3 hierboven.

Vraag 4.3a: Zijn de normgroepen groot genoeg?

De speciaal voor dit onderzoek via een onderzoeksbureau geworven laag-risicogroep (aangeduid als 'representatieve groep', n = 303) is ten behoeve van de normering samengevoegd met de groep wapenverlofhouders, n = 102) die de E-screener ook heeft ingevuld (document 17, punt 5). Voor de 'representatieve groep' zijn door het onderzoeksbureau 2067 personen uitgenodigd, 318 personen (15.4%) hebben de E-screener ingevuld. Er zijn 15 personen van de laag-risicogroep naar de hoog-risicogroep verplaatst omdat zij voldeden aan de kenmerken die golden voor opname in de hoog-risicogroep. De 'representatieve' groep bestaat derhalve uiteindelijk uit 303 personen en de groep die is gebruikt voor het bepalen van de 95%/5%-grens bestaat uit 405 personen (303 + 102). Dit aantal kan als 'goed' worden gekwalificeerd.
De groep die gebruikt wordt voor de normering die op het rapport wordt vermeld bestaat uit de wapenverlofhouders uit het pilot-onderzoek (n = 102) en de aanvragers van een wapenverlof in de periode november 2019 tot maart 2020 (n = 731). Ook de omvang van deze groep (n = 833) is 'goed' te noemen.
NB. Voor de hoog-risicogroep zijn 718 personen uitgenodigd, van hen hebben 92 personen de E-screener ingevuld, wat neerkomt op een responspercentage van 12.8%. Inclusief de 15 personen die zijn overgeheveld uit de laag-risicogroep bestaat de hoog-risicogroep uit 107 personen. De omvang van de hoog-risicogroep speelt hier echter geen rol, omdat de normgerichte interpretatie is gebaseerd op de 'representatieve groep' plus de groep wapenverlofhouders.

Vraag 4.3.b: Zijn de normgroepen representatief?

Zoals hierboven vermeld bestaat de normgroep die is gebruikt voor het bepalen van de 95%/5%-dichotomie uit de zogenoemde 'representatieve' groep plus de groep vergunninghouders. Volgens document 12 (p. 8) bestaat de 'representatieve' groep uit "95% mannen evenredig verdeeld over algemene kenmerken als leeftijd en opleiding, in overeenstemming met aanvragers van een wapenvergunning". De verdeling wat betreft leeftijd en opleiding van zowel de normgroep (de steekproef) als de groep aanvragers van een wapenvergunning (de populatie) wordt echter niet gegeven, dus deze bewering kan niet worden gecheckt. Evenmin wordt vermeld wat de andere algemene kenmerken zijn waarop deze groepen zouden zijn gematcht (regio, migratieachtergrond, sociaaleconomische status, enz.?). Ook wordt niet vermeld wat het percentage mannen is in de populatie aanvragers (het percentage mannen in de normgroep is overigens niet 95% maar 91%). Ook van de groep wapenverlofhouders als onderdeel van de normgroep wordt geen beschrijving in termen van achtergrondvariabelen gegeven (behalve dat deze groep voor 98% uit mannen bestaat). Door de onderzoekers werd representativiteit nagestreefd ten opzichte van de populatie van aanvragers, en niet ten opzichte van een doorsnee van de Nederlandse bevolking. Op zich is dit een verdedigbare keuze, alleen kan de representativiteit door het ontbreken van gegevens over de samenstelling van de steekproef en de populatie met betrekking tot de achtergrondvariabelen niet worden vastgesteld. De representativiteit wordt daarom beoordeeld als 'onvoldoende'.

Van de groep die is gebruikt voor de normering ten behoeve van het rapport wordt geen beschrijving gegeven. Aangezien deze groep voor 88% uit werkelijke aanvragers van een wapenvergunning bestaat (en voor 12% uit wapenverlofhouders) en het alle aanvragers van een wapenvergunning in de betreffende periode betreft is het wel aannemelijk dat deze groep representatief zal zijn voor de hele populatie van mogelijke aanvragers van een wapenvergunning. Ondanks het ontbreken van een beschrijving van deze groep wordt de representativiteit daarom als 'voldoende' beoordeeld.

De beoordeling voor *Normgerichte interpretatie* wordt 'onvoldoende' voor de normgroep in het pilot-onderzoek aangezien vraag 4.3.b met 'onvoldoende' is beoordeeld. Volgens de systematiek van het COTAN-beoordelingssysteem kunnen de vragen 4.4 t/m 4.7 voor deze groep dan worden overgeslagen. Voor de volledigheid worden deze vragen hier toch beantwoord.

Vraag 4.4: Worden de betekenis en beperkingen van de normschaal duidelijk gemaakt? En is het type normschaal in overeenstemming met doel van test?

Voor beide groepen wordt er gebruik gemaakt van percentielscores; de betekenis van dit type schaal mag bekend worden verondersteld. Percentielscores zijn op zich te gedifferentieerd in verhouding tot de range van ruwe scores in de diverse schalen. Dit is vooral een bezwaar voor de percentielscores die in het rapport worden gebruikt, omdat er daarbij grote sprongen in de percentielscores zullen voorkomen en er vele percentielscoreklassen leeg zullen blijven hetgeen een foutieve interpretatie van de betekenis van scoreverschillen kan veroorzaken. Voor de dichotomie 95%/5% is dit echter geen bezwaar.
Voor de normgroep in het pilot-onderzoek wordt vraag 4.4 met 'goed' beoordeeld en voor de groep ten behoeve van het rapport met 'voldoende'.

Vraag 4.5: Worden gemiddelden, standaardafwijkingen en gegevens over de scoreverdeling vermeld?

In de oorspronkelijk toegestuurde documenten worden geen gemiddelden en standaardafwijkingen vermeld. In het nagestuurde document 18 (p. 1-2) worden per schaal het gemiddelde, de standaardafwijking, de minimumscore en de maximumscore vermeld voor de drie subgroepen en voor de totale groep in de pilotstudie. Voor de normgroep ten behoeve van het rapport worden deze gegevens niet verstrekt. Gegevens over scheefheid en kurtosis worden voor geen van beide groepen vermeld.
Vraag 4.5 wordt beoordeeld met 'voldoende' voor de normgroep in het pilot-onderzoek en met 'onvoldoende' voor de normgroep ten behoeve van het rapport.

Vraag 4.6: Worden gegevens verstrekt over mogelijke verschillen tussen subgroepen (wel of geen migratieachtergrond, vrouwen-mannen)?

Voor geen van beide normgroepen worden gegevens over mogelijke verschillen tussen personen met en zonder migratieachtergrond verstrekt. In Bijlage G van document 12 staan wel gegevens over man-vrouw verschillen in de drie subgroepen van het pilot-onderzoek. Op één schaal scoren mannen in de 'representatieve' groep hoger dan vrouwen en op twee schalen scoren vrouwen in de hoog-risicogroep hoger dan mannen. Op één andere schaal scoren vrouwen in beide groepen hoger dan mannen. In de groep vergunninghouders is het aantal vrouwen te gering om de verschillen te onderzoeken.
Uit de nagestuurde gegevens in document 18 blijkt verder dat de vergunninghouders op zes van de zeven schalen de laagste gemiddelde scores behalen en tevens de kleinste standaarddeviaties vertonen; de hoog-risicogroep behaalt op deze schalen juist de hoogste gemiddelde scores en vertoont de grootste standaarddeviaties. Op twee schalen zijn de verschillen in gemiddelden en standaardafwijkingen tussen de drie groepen minimaal. De grootste gestandaardiseerde verschillen

die op de schalen tussen de uiterste groepen (vergunninghouders en hoog-risicogroep) worden gevonden bedragen circa 1.5 standaarddeviatie.
Vraag 4.6 wordt beoordeeld met 'voldoende' voor de normgroep in het pilot-onderzoek en met 'onvoldoende' voor de normgroep ten behoeve van het rapport.

Vraag 4.7: Worden er gegevens verstrekt over de standaardmeetfout, de standaardschattingsfout en/of de testinformatiefunctie/standaardfout?

Er worden geen gegevens over de standaardmeet fout en/of de standaardschattingsfout verstrekt. Vraag 4.7 wordt beoordeeld met 'onvoldoende'.

Vanwege het ontbreken van verdere gegevens wordt de beoordeling voor *Normgerichte interpretatie* ook voor de normgroep die is gebruikt ten behoeve van het rapport 'onvoldoende'.

Domeingerichte interpretatie
Deze normen hebben betrekking op de vier variabelen waaraan wordt gerefereerd bij punt 1 hierboven. Hierbij worden de afkappunten bepaald door experts (zie vraag 4.8).

Vraag 4.8: Is er voldoende overeenstemming tussen de beoordelaars?

Met betrekking tot Suïcidaliteit en Psychiatrische opnamegeschiedenis geldt dat beide variabelen uit slechts een of twee vragen bestaan en dat experts en politie aan deze variabelen een hoge score hebben toegekend als zijnde een risicofactor (tabel 1, document 3). Hieruit zou men een hoge mate van overeenstemming kunnen afleiden. De kritische scores voor Psychose (in combinatie met Medicijngebruik) en één andere schaal zijn gebaseerd op de score die oorspronkelijk zijn vastgesteld door het Trimbos instituut (zie document 3) en zijn gebaseerd op "normscores uit de literatuur", zonder dat wordt aangegeven waarop deze waarden zijn gebaseerd (overeenstemming tussen experts, prevalentie, een grens van X% in een normverdeling?). Er werd daarbij geen rekening gehouden met mogelijke verschillen tussen de Verenigde Staten en Nederland (zie document 16, punt 4). Door TNO (zie document 13, p. 12-13) is vervolgens onderzoek gedaan naar de optimale afkappunten. Bij Psychose (in combinatie met Medicijngebruik) leverde het optimale afkappunt een onrealistisch hoog percentage 'rode' uitslagen op en bij de andere schaal kon geen geschikt en eenduidig optimaal afkappunt worden gevonden. Om deze redenen is er uiteindelijk voor gekozen om de oorspronkelijke (door het Trimbos-instituut gesuggereerde) waarden te handhaven. Hierbij moet worden aangetekend dat er bij de gekozen kritische score van de betreffende schaal sprake is van een zeer lage sensitiviteit (7.5%) en een minimaal contrast (3.8%) tussen hoog- en laag-risicogroep[3].
De beoordeling voor vraag 4.8 wordt 'onvoldoende', omdat niet duidelijk is hoe de kritische grenzen tot stand zijn gekomen, of daarbij experts zijn betrokken en wat de mate van overeenstemming was tussen experts. Dit geldt niet voor Suïcidaliteit en Psychiatrische opnamegeschiedenis omdat het begrip kritische score op deze variabelen eigenlijk niet van toepassing is vanwege het dichotome karakter van deze variabelen en omdat er grote overeenstemming was tussen experts dat deze variabelen als risicofactor moesten worden opgenomen.

---

[3] In dit rapport wordt een aantal termen gebruikt die hier worden toegelicht. De **sensitiviteit** van de E-screener is het percentage terecht 'positieve' uitslagen, dat wil zeggen het terecht afwijzen van de aanvraag van een wapenvergunning als risicovol. De **specificiteit** van de E-screener is het percentage terecht 'negatieve' uitslagen, dat wil zeggen het terecht honoreren van de aanvraag van een wapenvergunning. Positief en negatief staan hier tussen aanhalingstekens omdat positief in dit geval negatieve consequenties heeft, nl. het afwijzen van de aanvraag van een wapenvergunning en omgekeerd. Het **contrast** is het verschil tussen het percentage dat terecht als 'positief' wordt aangemerkt en het percentage dat onterecht als 'positief' wordt aangemerkt (dat wil zeggen de sensitiviteit minus 1 min de specificiteit). Deze laatste groep wordt ook wel de **"vals positieven"** genoemd, in dit geval dus degenen waarvoor de aanvraag onterecht wordt afgewezen.

Vraag 4.9: Zijn de procedures op grond waarvan de grensscores zijn bepaald correct?

Deze vraag wordt beoordeeld met 'onvoldoende' (zie toelichting bij vraag 4.8).

Vraag 4.10: Zijn de beoordelaars naar behoren geselecteerd en getraind?

Met betrekking tot Suïcidaliteit en Psychiatrische opnamegeschiedenis was er grote overeenstemming tussen de experts dat deze variabelen moesten worden aangemerkt als risicofactor; de deskundigheid van deze groep is bij vraag 1.2 beschreven en kan als 'goed' worden beoordeeld. Over de kwalificaties van de beoordelaars/experts die de kritische grens op Psychose (in combinatie met Medicijngebruik) en de andere schaal hebben vastgesteld is niets bekend. Deze vraag wordt daarom beoordeeld als 'onvoldoende' voor Psychose en de andere schaal, en als 'goed' voor Suïcidaliteit en Psychiatrische opnamegeschiedenis.

De beoordeling voor *Domeingerichte interpretatie* wordt 'onvoldoende' met uitzondering van Suïcidaliteit en Psychiatrische opnamegeschiedenis; voor deze variabelen wordt de beoordeling 'voldoende'.

Criteriumgerichte interpretatie

Criteriumgerichte normering is van toepassing op de gewogen somscore. Zoals in het beoordelingssysteem wordt gesteld is dit type normen altijd gebaseerd op onderzoek naar de criteriumvaliditeit. De kwaliteit van het onderzoek naar de criteriumvaliditeit zelf wordt beoordeeld bij *Criteriumvaliditeit*. Hier wordt alleen de kwaliteit van de uit dit onderzoek afgeleide normen beoordeeld.

Vraag 4.11: Rechtvaardigen de onderzoeksresultaten het gebruik van grensscores?

In document 12 wordt een gewogen somscore berekend op grond van *alle* variabelen. De sensitiviteit van deze gewogen somscore is 81%, de specificiteit 80% en het contrast tussen laag- en hoog-risicogroep is 62.
In document 13 wordt een benadering gepresenteerd waarin de aanvragers zowel op grond van de score op een enkele schaal als op grond van een afkapscore op de gewogen somscore een negatief advies kunnen krijgen. Deze benadering levert een sensitiviteit op van 86%, een specificiteit van 73% en een contrast van 59%. Deze benadering levert dus een hogere sensitiviteit op, ten koste van een iets lagere specificiteit en een iets lager contrast dan wanneer beslissingen zouden worden genomen op basis van uitsluitend de gewogen somscore.
Uit document 18 blijkt dat bovenvermelde waarden inmiddels zijn achterhaald. Nadat de berekening van de gewogen somscore nogmaals is aangepast (omdat enkele schalen uit de gewogen somscore zijn verwijderd) wordt een sensitiviteit van 85%, een specificiteit van 75% en een contrast van 60% van de *gehele procedure* vermeld. Ter illustratie: de waarden voor de sensitiviteit, specificiteit en het contrast van *alléén de aangepaste gewogen somscore* zijn respectievelijk 78%, 84% en 62%.
In het beoordelingssysteem van de COTAN worden slechts globale indicaties gegeven met betrekking tot de waarden die als 'voldoende' of 'goed' kunnen worden beschouwd. Bovendien worden deze indicaties gegeven voor de waarden van de ROC-curves en niet voor de sensitiviteit en specificiteit afzonderlijk. Bij de beoordeling dient bovendien het doel van de test te worden betrokken. In het kader van een discussie over wat goed genoeg is dient de kanttekening te worden geplaatst dat de keuze voor een hogere sensitiviteit bij de E-screener (maar dit geldt in feite voor elke test, omdat een test nooit perfect kan voorspellen) onvermijdelijk leidt tot een lagere specificiteit en daardoor tot een hoger percentage vals positieven. Het Trimbos-instituut heeft oorspronkelijk gekozen voor strengere grenzen om het risico van het onterecht verlenen van een vergunning te verkleinen, maar dit leidde tot een lage specificiteit van circa 50% (en daarmee tot circa 50% vals positieven). Voor

gebruik in de klinische praktijk lijken de sensitiviteits- en specificiteitswaarden van het uiteindelijk gekozen model 'voldoende', waarbij moet worden aangetekend dat een percentage van 25% vals positieven bij het uiteindelijk gekozen afkappunt (het onterecht afwijzen van een wapenvergunning) als 'hoog' ervaren kan worden.
De beoordeling van vraag 4.11 is 'voldoende'.

Vraag 4.12: Is de onderzoeksgroep in overeenstemming met het bedoelde gebruik?

Het gaat hierbij om twee groepen, waarbij de laag-risicogroep in overeenstemming zou moeten zijn met de aanvragers van een wapenvergunning en de hoog-risicogroep in overeenstemming met de groep die een potentieel gevaar vormt indien aan hen een wapenvergunning zou worden verleend. Bij vraag 4.3 werd vastgesteld dat de representativiteit van de laag-risicogroep ten opzichte van de aanvragers van een wapenvergunning niet kan worden vastgesteld. Echter, met 'in overeenstemming met het bedoelde gebruik' wordt geen strikte representativiteit bedoeld (die bij normgerichte interpretatie wel essentieel is). Cruciaal is dat gemiddelden en spreidingen van de scores op de variabelen in de laag-risicogroep overeenkomen met die in de groep aanvragers van een wapenvergunning. Hiertoe zijn de gemiddelden en spreidingen op alle schalen en de gewogen somscore in de laag-risicogroep vergeleken met die van de groep aanvragers die de E-screener tussen november 2019 en februari 2021 hebben ingevuld (N = 2201, zie document 19). Op alle schalen alsmede de gewogen somscore blijken zowel de gemiddelden als de spreidingen van de laag-risicogroep hoger te zijn dan in de groep aanvragers. Dit betekent dat de scores van de laag-risicogroep dichter bij die van de hoog-risicogroep liggen dan die van de werkelijke aanvragers van een wapenvergunning. Dit leidt tot een *onderschatting* van de sensitiviteit en specificiteit van de E-screener. Dit effect wordt nog iets versterkt door de grotere spreiding op de schalen in de laag-risicogroep. De groep aanvragers blijkt de E-screener niet sociaal-wenselijker in te vullen, dus dat kan niet de oorzaak zijn van dit effect.
Voor de hoog-risicogroep kan niet worden vastgesteld of gemiddelden en spreidingen op de gemeten variabelen vergelijkbaar zijn met die van een groep die werkelijk in het bezit van een wapen een gevaar zou zijn, omdat deze gegevens niet bekend zijn. Wel kan worden vastgesteld dat de hoog-risicogroep aanmerkelijk hoger scoort dan de laag-risicogroep op twee na alle schalen en op de gewogen somscore (zie vraag 4.6), hetgeen naar verwachting ook zal gelden voor degenen aan wie geen wapenvergunning zou moeten worden verleend.
De beoordeling van vraag 4.12 is 'onvoldoende' wegens het ontbreken van gegevens.

Vraag 4.13: Is de onderzoeksgroep groot genoeg?

Als minimale eis voor de groepsgrootte wordt in het COTAN-beoordelingssysteem een omvang van 200 personen vermeld. De laag-risicogroep voldoet ruimschoots aan deze eis, maar de omvang van de hoog-risicogroep is te klein.
De beoordeling van vraag 4.13 is 'onvoldoende'.

De beoordeling voor *Criteriumgerichte interpretatie* wordt 'onvoldoende'. Weliswaar zijn de sensitiviteits- en specificiteitswaarden als 'voldoende' beoordeeld, maar er is niet vast te stellen of de onderzoeksgroepen in overeenstemming zijn met de doelgroepen; ook is een van de onderzoeksgroepen te klein.

Betrouwbaarheid

Vraag 5.1: Worden er gegevens over de betrouwbaarheid verstrekt?

Er worden gegevens over de betrouwbaarheid verstrekt in de vorm van coëfficiënt alfa.
Vraag 5.1 wordt met 'ja' beantwoord.

Vraag 5.2.b: Betrouwbaarheid op basis van inter-itemrelaties. Zijn de resultaten voldoende, gelet op het beoogde type beslissingen?

Voor Suïcidaliteit, Psychiatrische opnamegeschiedenis en Alcohol-, Drugs- en Medicijngebruik kan coëfficiënt alfa niet worden berekend: Psychiatrische opnamegeschiedenis bestaat uit slechts één item; Suïcidaliteit bestaat uit twee items, maar het antwoord op elk van de items apart wordt als risico beschouwd; Alcoholgebruik bestaat uit drie items, maar daarbij kunnen het tweede en derde item worden overgeslagen afhankelijk van het antwoord op het eerste item. Wanneer men ervan uitgaat dat deze (vooral feitelijke) vragen naar waarheid zullen worden ingevuld is een hoge betrouwbaarheid aannemelijk.
Voor acht schalen worden in document 12 alfa's gerapporteerd, maar deze zijn niet van toepassing omdat deze op de laag- en hoog-risicogroep tezamen zijn berekend. Omdat de normen zijn gebaseerd op de laag-risicogroep (en daarop in principe de beslissingen omtrent het toekennen van een wapenvergunning worden gebaseerd) dienen de alfa's ook op deze groep berekend te zijn. Zoals vermeld bij vraag 4.6 verschillen de gemiddelden en spreidingen in de laag- en hoog-risicogroep; hierdoor zullen de op de gezamenlijke groep berekende alfa's een positief vertekend beeld opleveren. Op verzoek van de deskundigen zijn de alfa's die berekend zijn op de aparte groepen ter beschikking gesteld (vertrouwelijke bijlage bij document 17). De alfa's voor de acht schalen berekend op de laag-risicogroep zijn (tussen haakjes de alfa's zoals gerapporteerd in document 12): .75 (.74), .63 (.62), .79 (.82), .70 (.77), .65 (.75), .96 (.96), .88 (.92) en .68 (.69). Voor tests voor 'belangrijke beslissingen op individueel niveau' is de grens voor 'voldoende' een betrouwbaarheid van minimaal .80 en voor 'goed' van minimaal .90. De betrouwbaarheid van zes schalen wordt derhalve gekwalificeerd als 'onvoldoende' (waarbij twee schalen zelfs een betrouwbaarheid lager dan .70 hebben), van één schaal is de betrouwbaarheid 'voldoende' en van eveneens één schaal 'goed'. Wel moet hierbij worden aangetekend dat zich onder de groep aanvragers van een wapenvergunning normaliter ook altijd enkele hoog-risico personen zullen bevinden (waardoor de spreiding en als gevolg daarvan ook de betrouwbaarheid wat hoger wordt); deze zijn er bij de laag-risicogroep uitgefilterd. De alfa's zullen in de praktijk derhalve tussen de hier gebruikte waarden en de waarden die tussen haakjes zijn vermeld liggen (bij deze waarden hebben echter ook nog vijf schalen een 'onvoldoende' betrouwbaarheid). In de groep vergunninghouders kunnen de alfa's van zeven schalen als 'onvoldoende' worden aangemerkt en van één schaal als 'goed'.
De betrouwbaarheid van de gewogen somscore wordt niet in de oorspronkelijk toegestuurde informatie vermeld, maar wel in de aanvullende gegevens (Document 17, punt 13). Deze is .96 wat leidt tot de kwalificatie 'goed'.
De betrouwbaarheid op basis van inter-itemrelaties van de schalen in de E-screener wordt voor zes schalen als 'onvoldoende' beoordeeld, voor één schaal als 'goed', voor één schaal als 'voldoende' en voor de gewogen somscore als 'goed'.

5.2.c: Test-hertestbetrouwbaarheid. Zijn de resultaten voldoende, gelet op het beoogde type beslissingen?

Wanneer een test geacht wordt over langere tijd zijn voorspellende waarde te behouden dient ook de test-hertestbetrouwbaarheid te worden vastgesteld. In document 16 (punt 5) wordt gesteld dat de E-screener jaarlijks zou moeten worden afgenomen. De geldigheidsduur van de E-screener is derhalve maximaal 1 jaar en test-hertestonderzoek over deze periode is dan ook vereist.
Wegens het ontbreken van onderzoek wordt de test-hertestbetrouwbaarheid van de E-screener als 'onvoldoende' beoordeeld.

NB. In de vragen 5.2.a en 5.2.d t/m 5.2.f worden andere vormen van betrouwbaarheid beoordeeld; deze zijn hier niet van toepassing.

Vraag 5.3.a: Wat is de kwaliteit van het onderzoek naar de betrouwbaarheid: Zijn de procedures voor berekening correct?

Men mag ervan uitgaan dat de alfa's correct zijn berekend. De methode van Qingping He voor het berekenen van de betrouwbaarheid van de gewogen somscore is adequaat.
Vraag 5.3.a wordt met 'goed' beoordeeld.

Vraag 5.3.b: Wat is de kwaliteit van het onderzoek naar de betrouwbaarheid: Zijn de steekproeven overeenkomstig het beoogde gebruik?

Zie de opmerkingen hierover bij vraag 5.2.b.
Vraag 5.3.b wordt voor de nieuw verstrekte gegevens met 'goed' beoordeeld.

Vraag 5.3.c: Wat is de kwaliteit van het onderzoek naar de betrouwbaarheid: Maken de gegevens die worden verstrekt een gefundeerd oordeel over de betrouwbaarheid mogelijk?

Voor de uiteindelijk verstrekte gegevens is het antwoord 'ja'.
Vraag 5.3.c wordt met 'goed' beoordeeld.

De beoordeling voor *Betrouwbaarheid* wordt 'onvoldoende' vanwege te lage betrouwbaarheden van de meerderheid van de schalen. De betrouwbaarheid van de gewogen somscore is 'goed'.

Begripsvaliditeit

Vraag 6.1: Worden er gegevens over begripsvaliditeit verstrekt?

Vraag 6.1 wordt met 'ja' beantwoord.

Vraag 6.2: Maken de resultaten voldoende aannemelijk dat het begrip zoals bedoeld wordt gemeten (of: maken de resultaten voldoende duidelijk wat wordt gemeten) op basis van gegevens over:

Vraag 6.2.a: De dimensionaliteit van de scores?

Er worden correlaties tussen de onderscheiden schalen berekend (zie de tabellen 8 en 9 in document 12 en de tabel op de laatste pagina van document 18; hier wordt uitgegaan van document 18 omdat dit de meest recente gegevens betreft). De later toegevoegde schaal Emotionele stabiliteit vertoont hoge correlaties met veel andere schalen, namelijk (.67, .64, .63, .60, .53 en .50). Deze hoge correlaties zijn niet direct verklaarbaar, tenzij men aanneemt dat Emotionele stabiliteit een soort overkoepelende klinische factor meet. De andere schalen vertonen minder hoge intercorrelaties. Slechts vier correlaties net boven .50 worden nog gevonden, alle andere correlaties zijn lager. Op zich zijn dit gematigde en vrij gebruikelijke correlaties bij dit type vragenlijsten. De onderzoekers zien de correlaties echter als een sterke samenhang, die zij ook verwachtten. Er lijkt echter geen reden om op voorhand een sterke samenhang te verwachten. Sterk correlerende schalen zouden bovendien zelfs contraproductief zijn: dat zou betekenen dat men met de verschillende schalen voor een groot deel hetzelfde meet. Dat is niet alleen inefficiënt (eigenlijk kan men dan met het meten van een enkele variabele volstaan), maar ook ongunstig bij de voorspelling van een externe variabele zoals risicogedrag, omdat men daarbij juist gebaat is bij variabelen die elk een uniek deel van de variantie met het criterium gemeen hebben. Vanuit dit oogpunt is de toevoeging van de schaal Emotionele stabiliteit die met veel andere schalen hoog correleert geen logische keuze (zie hiervoor verder bij *Criteriumvaliditeit*). De schaal Sociale wenselijkheid correleert licht negatief met alle overige schalen (correlaties tussen -.10 en -.30), wat erop lijkt te wijzen dat het geven van sociale wenselijke antwoorden een licht dempend effect heeft op de overige scores op de E-screener. Er

wordt geen verder onderzoek naar de dimensionaliteit van de E-screener als totaal (bijvoorbeeld factoranalyse over de schaalscores) of naar de schaalscores afzonderlijk (onderzoek naar unidimensionaliteit) vermeld.
Vraag 6.2.a wordt beoordeeld met 'voldoende'.

Vraag 6.2.b: De psychometrische kwaliteit van de items?

Met betrekking tot de kwaliteit van de items wordt als enige informatie op p. 19 van document 12 vermeld dat verwijdering van items uit schalen niet tot hogere alfa's zou leiden. Er worden geen item-restcorrelaties en p-waarden vermeld.
Vraag 6.2.b wordt beoordeeld met 'voldoende'.

Vraag 6.2.c: De invariantie van de factorstructuur en mogelijke itembias bij verschillende groepen?

De invariantie van de factorstructuur en mogelijk itembias bij subgroepen (bijvoorbeeld mannen-vrouwen) wordt niet onderzocht.
Vraag 6.2.c wordt beoordeeld met 'onvoldoende'.

Vraag 6.2.d: De convergentie en de discriminante validiteit?

Er worden zes vragenlijsten afgenomen om de convergente en discriminante validiteit van de schalen in de E-screener te onderzoeken. Hierbij is in de literatuur gezocht naar goed gevalideerde schalen met dezelfde betekenis als de schalen in de E-screener. De onderzoekers verwachtten hoge correlaties met de schalen van vragenlijsten die hetzelfde begrip zouden moeten meten en lagere correlaties met schalen die andere begrippen meten. Het betreft zes schalen waaronder de eerdergenoemde schaal Emotionele stabiliteit.
Voor de vragen met betrekking tot Suïcidaliteit, voor de schaal Sociale wenselijkheid en voor een van de andere schalen is de begripsvaliditeit bevestigd blijkens correlaties met een soortgenoot van boven .60. Voor één schaal zijn de resultaten niet overtuigend blijkens een correlatie van .42 met een soortgenoot (deze schaal correleerde bovendien hoger met enkele andere schalen). Voor vier schalen van de E-screener zijn in het geheel geen specifieke instrumenten ter bepaling van de convergente validiteit afgenomen, hoewel soms wel hoge correlaties met andere instrumenten worden gevonden. De vermelde correlaties met de andere vragenlijsten maakt voor deze vier schalen echter niet duidelijk of de schalen nu werkelijk de begrippen zoals bedoeld meten. De schaal Emotionele stabiliteit correleert zoals gezegd tamelijk hoog met de meeste schalen van de E-screener (correlaties tussen .51 en .67).
Vraag 6.2.d wordt met 'onvoldoende' beoordeeld omdat voor slechts twee schalen in de E-screener de convergente en discriminante validiteit eenduidig is vastgesteld

Vraag 6.2.e: Verschillen tussen relevante groepen?

Relevante groepen zijn de laag-risicogroep (bestaande uit de 'representatieve groep en de groep vergunninghouders) versus de hoog-risicogroep. Zoals bij vraag 4.6 vermeld vertoont de hoog-risicogroep de hoogste gemiddelden op alle schalen met uitzondering van twee schalen, waarop geen verschillen tussen de drie groepen worden gevonden. In het algemeen is dit ook wat men zou mogen verwachten. Het feit dat de groep vergunninghouders op alle schalen de laagste gemiddelden behaalt (eveneens met uitzondering van genoemde twee schalen), ook lager dan de 'representatieve groep', kan betekenen dat deze groep zich duidelijk bewust is van wat er op het spel stond en mogelijk heeft dat de invulling van de E-screener beïnvloed. Overigens is dergelijk gedrag niet ongebruikelijk in 'high-stake' situaties. Vraag 6.2.e wordt met 'voldoende' beoordeeld.

Vraag 6.2.f: Op basis van overige gegevens?

Er worden geen overige gegevens met betrekking tot de begripsvaliditeit vermeld. De gegevens met betrekking tot de criteriumvaliditeit zijn hier niet relevant, omdat deze in dit geval geen informatie verstrekken over de begripsvaliditeit van de schalen in de E-screener.
Vraag 6.2.f wordt met 'onvoldoende' beoordeeld.

Op grond van het bovenstaande wordt de beoordeling van vraag 6.2 in zijn totaliteit 'onvoldoende'.

Vraag 6.3.a: Zijn de procedures op basis waarvan de begripsvaliditeitsgegevens worden berekend correct?

Vooraf zijn er geen duidelijke verwachtingen ten aanzien van te verwachten correlaties geformuleerd, hoewel de keuze van de instrumenten bij de te verwachten correlaties natuurlijk wel een rol hebben gespeeld. De interpretatie heeft nu enigszins een post hoc karakter.
Vraag 6.3.a wordt met 'voldoende' beoordeeld.

Vraag 6.3.b: Komen de steekproeven in het begripsvalideringsonderzoek overeen met de groepen waarvoor de test is bedoeld?

In de oorspronkelijk ter beschikking gestelde documenten wordt niet vermeld op welke steekproef de correlaties zijn berekend. Uit het interview (document 17, punt 14) blijkt dat het gaat om het totaal van de drie groepen in de pilotstudie (n = 512). Bij vraag 4.6 is geconstateerd dat de drie groepen op de meeste schalen behoorlijk grote verschillen te zien geven in gemiddelden en spreidingen. Het is moeilijk in te schatten wat het effect is van het berekenen van de validiteitscoëfficiënten op de samengevoegde groep op de hoogte van deze coëfficiënten. Wellicht was het beter geweest de correlatiematrix per groep te berekenen. Wanneer deze geen verschillen te zien zouden geven zou in ieder geval zijn aangetoond dat de constructen binnen elke groep dezelfde betekenis hebben.
Vraag 6.3.c wordt met 'voldoende' beoordeeld.

Vraag 6.3.c: Wat is de kwaliteit van de andere maten die in het begripsvalideringsonderzoek zijn gebruikt?

In tabel 5 van document 12 wordt een overzicht gegeven van de psychometrische eigenschappen van de andere vragenlijsten. Voor de meeste schalen wordt door de onderzoekers vermeld dat ze een goede validiteit hebben. De alfa's van deze schalen zijn in het onderzoek van TNO zelf vastgesteld en liggen tussen .78 en .92, voor onderzoeksdoeleinden is dit hoog genoeg.
Vraag 6.3.c wordt met 'goed' beoordeeld.

Vraag 6.3.d: Is de kwaliteit van het onderzoek, zoals beoordeeld in de vragen 6.3a t/m 6.3c, zodanig dat de beoordeling van de begripsvaliditeit, zoals gegeven in vraag 6.2, kan worden bevestigd?

Deze samenvattend vraag wordt met 'voldoende' beoordeeld.

De beoordeling voor de *Begripsvaliditeit* wordt 'onvoldoende' op grond van de beoordeling van vraag 6.2. Er worden te weinig gegevens aangedragen die de begripsvaliditeit van de schalen in de E-screener ondersteunen en de wel vermelde gegevens leveren geen eenduidige ondersteuning.

Criteriumvaliditeit

Vraag 7.1: Worden er gegevens verstrekt over het verband test-criterium?

Er worden gegevens verstrekt over het onderscheid dat met de E-screener kan worden gemaakt tussen een laag- en hoog-risicogroep, dit is het criterium.
Vraag 7.1 wordt beantwoord met 'ja'.

Vraag 7.2: Zijn de resultaten voldoende, gelet op het type beslissingen dat met de test moet worden genomen?

Zoals bij Normen beschreven kan een negatief advies tot stand komen op grond van de score op een van de schalen afzonderlijk en/of op grond van de gewogen somscore. De gewichten voor de gewogen somscore werden bepaald met behulp van regressieanalyse waarbij werd 'voorspeld' of iemand tot de laag- dan wel hoog-risicogroep behoort. De sensitiviteit, specificiteit en het contrast van de totale procedure waren respectievelijk 85%, 75% en 60%. Deze benadering levert een iets hogere sensitiviteit op (85% versus 78%), ten koste van een iets lagere specificiteit (75% versus 84%) en een iets lager contrast (60% versus 62%), dan wanneer uitsluitend de gewogen somscore als selectie-instrument zou zijn gebruikt. Uit figuur 2 in document 13 is af te leiden dat de zeven schalen waarbij de kritische grens via de dichotomie 95/5% wordt bepaald nog slechts 2% aan de sensitiviteit van de E-screener toevoegen. Mogelijk kan dit worden verklaard doordat vier van de betreffende zeven schalen ook al een rol spelen in de gewogen somscore. Blijkens de resultaten van de regressieanalyse met betrekking tot de somscore (document 18, p. 3) zijn de gewichten van vier variabelen significant. Drie van deze variabelen spelen een ondergeschikte rol; de bijdragen van deze variabelen aan de regressie zijn niet significant.
Wanneer daar aanleiding toe is (bijvoorbeeld als subgroepen verschillen in gemiddelde scores), dient de auteur van een vragenlijst onderzoek uit te voeren naar mogelijke predictiebias voor de betreffende groepen. In dit geval zou via dergelijk onderzoek bijvoorbeeld kunnen zijn nagegaan of de sensitiviteits- en specificiteitswaarden voor personen met en zonder migratieachtergrond verschillen, zodat misschien andere kritieke grenzen in de onderscheiden groepen moeten worden gehanteerd. De gemiddelde scores van mannen en vrouwen bleken op enkele schalen te verschillen (zie vraag 4.6). In bijlage G, document 12 worden de sensitiviteits- en specificiteitswaarden voor de einduitslag op de E-screener bij mannen en vrouwen apart voor de laag- en hoog-risicogroep vermeld. Deze gegevens laten iets betere validiteitsresultaten voor vrouwen zien. Hierbij is uitgegaan van beslissingsregels gebaseerd op de totale groep. Dit resultaat had aanleiding kunnen zijn om te onderzoeken of aparte beslissingsregels per sekse tot een hogere validiteit had geleid (met name bij mannen, in de praktijk immers verreweg de grootste groep bij de aanvragers voor een wapenvergunning). Onderzoek naar predictiebias bij andere subgroepen is niet uitgevoerd.
Vraag 7.2 wordt beoordeeld met 'voldoende'.

Vraag 7.3.a: Zijn de procedures op grond waarvan de criteriumvaliditeitsgegevens zijn berekend correct?

Het belangrijkste punt van kritiek is dat voor het bepalen van de waarde van de gewogen somscore geen kruisvalidatie is toegepast. Dit klemt temeer omdat de gewogen somscore in het beslissingsproces de belangrijkste bijdrage levert. Technieken als regressieanalyse kapitaliseren op toeval: in de onderzoeksgroep wordt gezocht naar de meest gunstige combinatie van variabelen, maar bij gebruik van de gewichten op een andere groep is in het algemeen de voorspellende waarde lager, met name wanneer de groepen niet al te groot zijn, zoals hier bij de hoog-risicogroep. In dit verband speelt tevens een rol dat in de somscore ook variabelen worden meegeteld die geen significante bijdrage leveren, al zal hun bijdrage als gevolg van hun lage regressiegewicht beperkt zijn. Merkwaardig is dat dezelfde variabelen in verschillende stappen van het beslissingsproces een rol spelen; het effect daarvan is moeilijk te overzien, maar het kan zijn dat aanvragers die op een van deze variabelen hoog scoren hierop dubbel worden afgerekend.
De schaal Emotionele stabiliteit werd opgenomen in de procedure vanwege de hoge correlaties met de andere variabelen en het feit dat met deze schaal de hoog- en laag-risicogroep goed konden

worden onderscheiden. Echter, door de hoge correlaties met de andere schalen is de unieke bijdrage van de schaal Emotionele stabiliteit gering, zoals blijkt uit de regressieanalyse ten behoeve van de gewogen somscore. Er had beter kunnen worden gezocht naar relevante variabelen die laag met de andere variabelen correleerden zodat ze een specifieke bijdrage hadden kunnen leveren in de voorspelling.
Vraag 7.3.a wordt met 'onvoldoende' beoordeeld.

Vraag 7.3.b: Zijn de steekproeven op grond waarvan de criteriumvaliditeitsgegevens zijn berekend in overeenstemming met het beoogde testgebruik?

Duidelijk wordt aangegeven dat in het onderzoek naar de criteriumvaliditeit voor de laag-risicogroep zowel de personen in de 'representatieve' groep als in de groep wapenverlofhouders zijn meegenomen. Zoals bij vraag 4.3.b aangegeven is op grond van de verstrekte gegevens niet vast te stellen of deze groep representatief geacht kan worden voor de aanvragers van een wapenvergunning. Echter, als 'normale' contrastgroep voor de hoog-risicogroep kan deze groep waarschijnlijk goed worden gebruikt.
De hoog-risicogroep is samengesteld op basis van de Pre-screener (document 12, p. 9). De Pre-screener bestaat uit 8 vragen die voor een belangrijk deel overlappen met vragen die in de E-screener worden gesteld. Dus: Eerst worden op grond van de antwoorden op de Pre-screener personen voor een groep geselecteerd en vervolgens wordt met de E-screener (mede) op grond van de antwoorden op enkele vergelijkbare vragen 'voorspeld' of personen tot deze groep behoren. Deze contaminatie heeft een ongewis maar in ieder geval inflatoir effect op de uitkomsten van het validiteitsonderzoek. Deze ongewenste invloed wordt overigens ook door de onderzoekers erkend (document 12, p. 5). Een ander punt betreft de samenstelling van de hoog-risicogroep wat betreft sekse. De hoog-risicogroep bestaat voor circa de helft uit vrouwen, terwijl de groep aanvragers van een wapenvergunning voornamelijk uit mannen bestaat. Deze observatie is van belang omdat de E-screener zoals bij vraag 7.2 reeds is geconstateerd bij vrouwen iets betere validiteitsresultaten laat zien. In een groep die voornamelijk uit mannen zou hebben bestaan, overeenkomstig de samenstelling in de groep aanvragers van een wapenvergunning, zou de validiteit dus wat lager zijn geweest. Een ander punt van kritiek op de samenstelling van de hoog-risicogroep wordt door het Trimbos-instituut genoemd in document 11. Van degenen die zijn uitgenodigd voor deelname aan het onderzoek en die werden aangemerkt als hoog-riskant heeft slechts een laag percentage (12.8%) aan het onderzoek deelgenomen. Dit zou een selectie van respondenten kunnen zijn, dat wil zeggen respondenten die wat minder extreem scoren op bepaalde schalen. Het effect hiervan is echter vermoedelijk dat de hoog-risicogroep minder afwijkend is samengesteld dan eigenlijk zou moeten. Het gevolg hiervan is dat het contrast met de laag-risicogroep wordt verkleind; dit zal dus geen opwaarts maar eerder een dempend effect op de validiteit tot gevolg hebben. Dit lage responspercentage leidt wel tot de vraag of de hoog-risicogroep een valide representatie vormt van de 'werkelijke' groep die met een wapen in handen een risico voor zichzelf of anderen zou vormen.
Vraag 7.3.b wordt met 'onvoldoende' beoordeeld.

Vraag 7.3.c: Wat is de kwaliteit van de criteriummaten?

De criteriummaat wordt in dit geval gevormd door de laag- en hoog-risicogroep. De samenstelling van deze groepen is bij vraag 7.3.b reeds besproken.
Vraag 7.3.c wordt met 'onvoldoende' beoordeeld.

Vraag 7.3.d: Is de kwaliteit van het onderzoek, zoals beoordeeld in de vragen 7.3.a t/m 7.3.c, zodanig dat de beoordeling van de criteriumvaliditeit, zoals gegeven in vraag 7.2, kan worden bevestigd?

Op grond van de beoordelingen op de vragen 7.3.a t/m 7.3.c wordt vraag 7.3.d eveneens met 'onvoldoende' beantwoord.

De beoordeling van de *Criteriumvaliditeit* wordt 'onvoldoende'. Weliswaar konden de resultaten op zich als 'voldoende' worden beoordeeld, maar de kwaliteit van het onderzoek waarop deze resultaten zijn gebaseerd is 'onvoldoende' waardoor er gerede twijfel is over de houdbaarheid van deze resultaten.

Addendum Fairness

Fairness kan een rol spelen bij de volgende vragen in het beoordelingssysteem:

Vraag 2.11: Zijn de items vrij van racistische, kwetsende inhoud?

Er is geen racistische of anderszins kwetsende inhoud in de items aangetroffen.

Vraag 4.6: Worden gegevens verstrekt over mogelijke verschillen tussen subgroepen (wel of geen migratieachtergrond, vrouwen-mannen)?

Deze gegevens worden voor etnische subgroepen niet verstrekt, maar wel voor vrouwen-mannen. Op enkele schalen blijken de gemiddelde scores van vrouwen en mannen te verschillen. De mogelijke consequenties hiervan worden niet onderzocht (zie vraag 6.2.c en 7.2).

Vraag 6.2.c: De invariantie van de factorstructuur en mogelijke itembias bij verschillende groepen?

De invariantie van de factorstructuur en mogelijk itembias bij subgroepen (bijvoorbeeld mannen-vrouwen) wordt niet onderzocht.

Vraag 7.2: Zijn de resultaten voldoende, gelet op het type beslissingen dat met de test moet worden genomen?

Onderzoek naar mogelijke predictiebias bij subgroepen is niet uitgevoerd.

Samenvattend kan worden geconstateerd dat aan fairness van de E-screener voor diverse subgroepen middels onderzoek geen aandacht is besteed.

Samenvatting van de beoordelingen

Uitgangspunten van de testconstructie	voldoende
Kwaliteit van het testmateriaal	goed
Kwaliteit van de handleiding	voldoende
Normen	
- *Normgerichte interpretatie*	onvoldoende
- *Domeingerichte interpretatie*	voldoende
- *Criteriumgerichte interpretatie*	onvoldoende
Betrouwbaarheid	onvoldoende
Begripsvaliditeit	onvoldoende
Criteriumvaliditeit	onvoldoende

Conclusie

In de Inleiding van het hier gebruikte COTAN-beoordelingssysteem voor de kwaliteit van tests is de waarschuwing opgenomen dat een minder deskundige testgebruiker een test c.q. vragenlijst vooral niet zou moeten gebruiken wanneer er bij een test meerdere onvoldoendes voorkomen. Tot deskundige testgebruikers kan men in dit geval rekenen klinische psychologen met de basisaantekening psychodiagnostiek (BAPD), forensisch psychiaters, e.d. In het algemeen zijn dit deskundigen die voldoen aan het EFPA Level 3 zoals beschreven in de European Federation of

Psychologists' Associations Standards for Test Use (2012) en de Algemene Standaard Testgebruik van het Nederlands Instituut van Psychologen (2018). Tenzij de betrokken politiefunctionarissen kwalificaties bezitten die hiermee overeenkomen zou om deze reden het gebruik van de E-screener door deze functionarissen moeten worden afgeraden.

De onvoldoendes die aan de E-screener worden toegekend hebben voor een deel als oorzaak dat sommige gegevens niet zijn berekend, voor een deel aan psychometrische tekortkomingen (bijvoorbeeld bij de betrouwbaarheid) of aan problemen met de inrichting van het onderzoek (zoals bij de begripsvaliditeit en criteriumvaliditeit).

In eerste instantie was er op vele punten ook een tekort aan informatie, maar dit kon grotendeels worden opgelost doordat deze via de interviews met de vertegenwoordigers van het Trimbos-instituut en TNO alsnog kon worden verschaft. Omdat de E-screener, niet alleen inhoudelijk maar vooral in de wijze waarop deze wordt gebruikt, in de loop der tijd een aantal wijzigingen heeft ondergaan was tevens niet altijd duidelijk welke gegevens wel en welke niet meer van toepassing waren. Hier wordt dan ook sterk aanbevolen om alle informatie uit de vele documenten die de deskundigen toegestuurd hebben gekregen te ordenen en samen te voegen tot een document dat zich laat lezen als een echte testhandleiding die een up-to-date beschrijving geeft van het instrument.

Met betrekking tot de samenstelling van de E-screener moet worden opgemerkt dat de later toegevoegde variabele Emotionele stabiliteit weinig bijdraagt en dat actiever gezocht had kunnen worden naar variabelen die juist wel specifieke variantie hadden kunnen bijdragen. Sommige informatie met betrekking tot de psychische gesteldheid, zoals cognitieve beperkingen, radicalisering e.d. wordt niet met de E-screener verkregen en zal dus langs andere weg moeten worden verzameld.

Het feit dat betrouwbaarheid voor de meeste schalen te laag is hoeft geen bezwaar te zijn als ervoor wordt gezorgd dat deze schalen afzonderlijk geen rol meer spelen in de beslissingsprocedure; als afzonderlijke beslisgrond voegen deze schalen bovendien nauwelijks iets toe aan de voorspellende waarde van de E-screener. De gewogen somscore is de belangrijkste voorspeller, maar niettemin konden ook enkele schalen die door de experts als essentieel worden gezien niet worden gemist en deze zullen dus moeten worden gehandhaafd, ook al zijn de betrouwbaarheden aan de magere kant. In alle gevallen is wel nog onderzoek naar de test-hertestbetrouwbaarheid gewenst.

Ten behoeve van de bepaling van de begripsvaliditeit zou aanvullend onderzoek moeten worden gedaan, waarbij voor alle schalen waarvoor dat nu niet is gebeurd specifieke soortgenoten worden afgenomen (waarbij hopelijk voldoende hoge correlaties blijken en lagere met niet verwante begrippen). Aangezien het allemaal bestaande schalen betreft zou een alternatief kunnen zijn dat wordt nagegaan of in de bronpublicaties waaraan deze instrumenten zijn ontleend deze gegevens wel voorhanden zijn en deze in de handleiding van de E-screener over te nemen. Voorwaarde is wel dat het Nederlands onderzoek betreft.

De criteriumvaliditeit werd vastgesteld door te onderzoeken of de E-screener juist kan voorspellen of een persoon tot een laag- of een hoog-risicogroep voor het verlenen van een wapenvergunning behoort. De resultaten waren hoopgevend maar het ontbreken van kruisvalidatie op een andere groep, contaminatie bij de samenstelling van de hoog-risicogroep en twijfels over de samenstelling van deze groep maakten dat deze resultaten niet als zodanig kunnen worden geaccepteerd. Hier zal dus aanvullend onderzoek moeten plaatsvinden, waarbij wordt erkend dat het uitermate lastig zal zijn, wellicht ondoenlijk, om een geschikte hoog-risicogroep samen te stellen.

In het beoordelingssysteem voor de kwaliteit van tests wordt niet geverifieerd of de testauteur/testuitgever/testonderzoeker aangeeft hoelang testresultaten geldig blijven. Bij de E-screener is dat zeker een relevant issue, omdat persoonlijke omstandigheden kunnen veranderen en de uitkomst 'niet voor het leven geldt'. In de handleiding zou dan ook duidelijk moeten worden aangegeven, zowel bij een negatieve als bij een positieve uitslag, na welke periode een herhaalde

afname zou kunnen of moeten plaatsvinden. Vervolgens kan de lengte van het test-hertestinterval bij het uit te voeren onderzoek hierop worden afgestemd.

Beantwoording van de 15 vragen uit de vraagstelling van de rechtbank

1. Wat zijn de theoretische achtergronden van de E-screener (wat wordt er gemeten)?

Bij de ontwikkeling van de E-screener heeft een inventarisatie van risico-indicatoren voor geweld plaatsgevonden. Bestaande risicotaxatie-instrumenten zijn ontwikkeld voor gebruik in forensische (psychiatrische) settings of als klinisch interview en daarmee ongeschikt als instrument voor zelfrapportage in de algemene bevolking (Trimbos-instituut 13-1-2014, p. 15 e.v.). Wel suggereren deze instrumenten welke factoren van belang zouden kunnen zijn (onder andere een gewelddadig/crimineel verleden, psychische problematiek, middelengebruik).
Uit het literatuuronderzoek verricht door het Trimbos-instituut kwamen een aantal psychologische risico-indicatoren naar voren, t.w. de aanwezigheid van psychotische stoornissen (schizofrenie en bipolaire stoornis), depressie en sommige persoonlijkheidsstoornissen (met name de antisociale persoonlijkheidsstoornis). Geconstateerd werd dat een psychiatrische stoornis *an sich* niet hoeft te leiden tot gewelddadig gedrag, maar mogelijk wel in combinatie met behandelstatus (b.v. therapie-ontrouw), alcohol- en drugsmisbruik, een forensisch verleden enz. Ten slotte is bij de ontwikkeling van de E-screener een consensusmeeting georganiseerd waarbij risico-indicatoren uit verschillende categorieën (ziekte-/stoornisgerelateerd, stressvolle omstandigheden, kenmerken aanvrager) werden geïnventariseerd (Trimbos-instituut 13-1-2014, p. 37 e.v.).

In het kader van het huidige onderzoek is door de deskundigen eveneens een (oriënterende) literatuurstudie uitgevoerd, waarbij op PubMed (https://pubmed.ncbi.nlm.nih.gov/) gezocht werd naar recente wetenschappelijke artikelen met betrekking tot risicofactoren bij vuurwapengeweld (zoektermen 'gun' 'firearm' 'violence' 'mental' 'psychopathology'), met waar nodig *snowball sampling*, bij voorkeur (systematische) reviews. Enkele algemene conclusies hieruit:

- Het overgrote deel van de gepubliceerde artikelen heeft betrekking op de Verenigde Staten, met zoals bekend een veel minder strikte wapenwetgeving dan Nederland; bevindingen van dergelijke studies zijn daarom maar beperkt toepasbaar voor de Nederlandse situatie. Studies uit andere landen zijn schaars en hebben veelal betrekking op kleine aantallen (daders resp. slachtoffers) wat statistische benaderingen bemoeilijkt. In een recent overzicht werd geconcludeerd dat voor de Europese situatie onvoldoende gegevens beschikbaar zijn (Krüsselman et al., 2021).
- Resultaten van eerder onderzoek zijn niet eenduidig, diagnostische omschrijvingen ontbreken nogal eens ('serious mental illness' als verzamelterm) of er wordt geen duidelijk onderscheid gemaakt tussen verschillende typen indicatoren: psychiatrische stoornissen, persoonskenmerken en omgevingsfactoren.
- De aanwezigheid van een psychotische stoornis (zoals schizofrenie) of depressie is een risicofactor, maar lijkt substantieel kleiner dan bij middelenmisbruik of psychopathie. Ook psychosociale factoren (leeftijd, socio-economische status, voorgeschiedenis van betrokkenheid bij geweld) zijn relevante predictoren (Pinals & Anacker, 2016; Rozel & Mulvey, 2017). In een recente meta-analyse kwamen vergelijkbare risicofactoren naar voren, t.w. een voorgeschiedenis van geweldsincidenten (met name huiselijk geweld), impulsief gedrag, middelenmisbruik en lage socio-economische status. Psychiatrische aandoeningen waren geassocieerd met verhoogde suïcidaliteit, niet zozeer een verhoogd geweldsrisico (Sanchez et al., 2020). Dit betreft vooral epidemiologische studies; bij daderstudies is wel een verband met psychosen aangetoond (Swinson et al., 2011), in een studie bij 'mass shootings' is gebleken dat twee-derde (23:35) van de (overlevende) daders een onbehandelde psychotische stoornis had, meestal schizofrenie (Glick et al., 2021). Er is discussie in de literatuur of psychiatrische stoornissen als schizofrenie en bipolaire stoornis op zichzelf een risicofactor zijn of alleen in combinatie met andere indicatoren zoals alcohol- en middelengebruik (Swanson et al., 2015; Baumann & Teasdale, 2018); wel is er consensus over het verhoogde suïciderisico bij psychiatrische stoornissen (McGinty et al., 2014). Niet verrassend is dat het suïciderisico substantieel afneemt als vuurwapens afgesloten en

ongeladen worden bewaard (Shenassa, 2004). Van belang is nog dat het geweldsrisico bij psychiatrische stoornissen mede afhangt van de fase waarin patiënten verkeren (relatief laag bij patiënten in ambulante behandeling, hoog bij eerste-episode psychose voorafgaand aan behandeling; Swanson et al., 2015).
- Behalve bij psychosen, depressie en psychopathie zijn er risico's van vuurwapenbezit bij verstandelijke beperking en niet-aangeboren hersenletsel (Byers et al., 2020; Swanson et al., 2015) en verder bij neurodegeneratieve aandoeningen zoals Alzheimer (Betz, 2020).
- In de psychologische literatuur is recent aandacht gegeven aan de relatie tussen 'negatieve' persoonskenmerken als subklinisch narcisme en psychopathie, sadisme en Machiavellisme, de z.g. 'dark tetrad', en het gebruik van (o.a. seksueel) geweld (b.v. Pineda et al., 2021), maar studies gerelateerd aan incidenten met vuurwapens ontbreken.
- Tenslotte wordt in de literatuur nog gewezen op het gegeven dat incidenten met vuurwapens vaak uiting zijn van een crisis (Auxemery, 2015), meestal in de interpersoonlijke sfeer, in het gezin/de familie of tussen (seksuele) partners (De Moore et al., 1994; Dudley, 1996).

De opzet van de E-screener lijkt goed aan te sluiten bij uit de literatuur bekende gegevens met uitzondering van de rol van cognitieve beperkingen: in het interview met [redacted] (Trimbos) kwam naar voren dat deze laatste informatie langs andere weg beschikbaar zou dienen te zijn. Wel is van belang erop te wijzen dat de combinatie Psychose en Medicatiegebruik als exclusiecriterium problematisch kan zijn omdat de literatuur suggereert dat met name *onbehandelde* psychose een risicofactor vormt. Zo was bij Tristan van der V. schizofrenie geconstateerd, waarbij hij ook bekend was als 'zorgwekkende zorgmijder' (https://tinyurl.com/wfmmavz2) en voor zover bekend geen medicatie gebruikte.

2. In hoeverre is de E-screener psychometrisch betrouwbaar?

Het antwoord op deze vraag is een samenvatting van de beoordeling van het criterium *Betrouwbaarheid* volgens het beoordelingssysteem van de COTAN. Zie voor de volledige beoordeling en een toelichting daarop het vorige hoofdstuk.

De betrouwbaarheid van de meerderheid van de afzonderlijke schalen is 'onvoldoende', de betrouwbaarheid van één schaal is 'voldoende' en de betrouwbaarheid van één schaal is 'goed'. De betrouwbaarheid van de gewogen somscore is 'goed'.

Vanwege te lage betrouwbaarheden van de meerderheid van de schalen en omdat gegevens over de test-hertestbetrouwbaarheid van de E-screener ontbreken is de beoordeling voor de betrouwbaarheid van de E-screener als geheel 'onvoldoende'.

3. In hoeverre is de E-screener psychometrisch gevalideerd?

Het antwoord op deze vraag is een samenvatting van de beoordeling van de criteria *Begripsvaliditeit* en *Criteriumvaliditeit* volgens het beoordelingssysteem van de COTAN. Zie voor de volledige beoordeling en een toelichting daarop het vorige hoofdstuk.

De begripsvaliditeit van de E-screener is 'onvoldoende', omdat te weinig gegevens worden aangedragen die de begripsvaliditeit van de schalen in de E-screener ondersteunen. De wel vermelde gegevens leveren geen afdoende ondersteuning.

De criteriumvaliditeit van de E-screener is 'onvoldoende'. Weliswaar konden de resultaten op zich als 'voldoende' worden beoordeeld, maar de kwaliteit van het onderzoek waarop deze resultaten zijn gebaseerd is 'onvoldoende' waardoor er gerede twijfel is ontstaan aan de houdbaarheid van deze resultaten.

4a. Hoe is de cesuur bepaald (dat wil zeggen hoe is bepaald wanneer iemand voldoet of niet voldoet)?

De cesuur wordt op drie verschillende manieren bepaald:

- Voor vier variabelen is de kritische score bepaald door experts en/of door gegevens uit de literatuur.
- Voor de gewogen somscore is de kritische score empirisch bepaald met behulp van de sensitiviteit, de specificiteit en het contrast.
- Voor zeven schalen is de kritische score bepaald op grond van de gegevens van een normgroep.

4b. Hoe werkt de cesuur?

Wanneer op een van de hierboven genoemde variabelen een score boven de kritische grens wordt gehaald leidt dit tot een negatief advies voor het verlenen van een wapenvergunning (uitslag Rood). Alleen voor de schaal Sociale wenselijkheid leidt een score boven de kritische grens tot de score Oranje (= scores onbetrouwbaar).

4c. Hoe is men tot de vaststelling van deze cesuur gekomen?

Zie de beantwoording van vraag 4a en ook de beoordeling van *Normen* en *Criteriumvaliditeit* in het vorige hoofdstuk.

5 Voldoet deze vorm van afname (in hoeverre is er een kans op 'faken')?

Bij zelfrapportage met betrekking tot de eigen persoonlijkheid is er altijd kans op faken. Daarop is de E-screener geen uitzondering. Oplossingen voor faking, zoals het gebruik van een sociale wenselijkheidsschaal, het berekenen van een sociale wenselijkheidsindexwaarde per item, blijven vaak onbevredigend. In tegenstelling tot situaties waarin er beslissingen vergelijkenderwijs zouden moeten worden genomen op basis van rangordescores van aanvragers – zoals doorgaans wel het geval is bij personeelsselectie – of op basis van een bepaald quotum van de te verlenen vergunningen, speelt faking bij de E-screener vooral een rol in verband met de aftestgrens. Het faken van antwoorden kan overigens ook een onderdeel zijn van iemands psychopathologie (bijv. zelfmisleiding; zie Paulhus, 1998). Het valt te overwegen om het effect van faking op de E-screener nader te onderzoeken bijvoorbeeld door ipsatieve in plaats van Likert-schalen te gebruiken, waarbij de aanvrager bijvoorbeeld gedwongen moet kiezen tussen twee antwoordalternatieven die beide even sociaal (on)wenselijk zijn. Een andere onderzoeksoptie is om twee groepen respondenten de E-screener te laten invullen, waarbij de ene groep de opdracht krijgt om de antwoorden zo in te vullen dat de kans op het verkrijgen van een vergunning zo groot mogelijk is, en de andere groep de E-screener zo eerlijk mogelijk moet invullen. Het verschil in scores tussen beide groepen vormt een indicatie voor de fake-baarheid van de E-screener. De fake-baarheid van instrumenten zoals de E-screener pleit ervoor om een hoge cut-off grens te hanteren, met andere woorden om de E-screener vooral te gebruiken 'to exclude the unacceptable'. Dit impliceert dat de E-screener niet direct kan differentiëren tussen diegenen die daadwerkelijk acceptabel zijn en diegenen die gefaked acceptabel zijn. Zoals gezegd is dit geen uniek probleem van de E-screener.

6 In hoeverre is de vragenlijst mogelijk niet fair voor bepaalde groepen (bijvoorbeeld ouderen – jongeren)?

De E-screener is een schriftelijke vragenlijst, die indien men dat wenst ook wordt voorgelezen. Sommige woorden of zinsdelen (lastige woorden, uitdrukkingen) kunnen mogelijk wat minder leesbaar zijn voor bepaalde groepen die minder verbaal onderlegd zijn. Ook valt het op dat er items (uitspraken) voorkomen die het woord 'niet' bevatten, waarna de aanvrager moet aangeven in hoeverre zij het ermee oneens of eens zijn. Deze constructie kan lastig zijn en in het nadeel van groepen die minder cognitief vaardig zijn. Omdat de E-screener geen meting dient te zijn van IQ/ verbale vaardigheid, valt het te overwegen het leesbaarheidsniveau van de vragenlijst in te schatten,

bijvoorbeeld met de Flesch-Kincaid leesbaarheidsindex (Flesch Kincaid Readability Index; ook voor de Nederlandse taal), die lengte van zinnen en woorden en woordfrequentie in de index betrekt. Fairness kan voorts betrekking hebben op de item-inhoud (bijvoorbeeld een seksistische formulering), op scoreverschillen, itempartijdigheid en factorstructuurverschillen tussen subgroepen (bijvoorbeeld mensen met en mensen zonder een migratieachtergrond), en op predictiebias. Er is geen racistische of anderszins kwetsende inhoud in de items aangetroffen. Er zijn geen gegevens verschaft over eventuele scoreverschillen tussen etnische subgroepen maar wel is gerapporteerd dat op enkele schalen de gemiddelde scores van vrouwen en mannen verschillen. De mogelijke gevolgen hiervan zijn niet onderzocht. Factorstructuurverschillen noch mogelijke itempartijdigheid zijn onderzocht voor subgroepen, evenmin als potentiële predictieve bias. Al met al is er middels empirisch onderzoek nauwelijks aandacht besteed aan de fairness van de E-screener voor diverse subgroepen (zie verder de sectie beoordeling van de E-screener volgens het beoordelingssysteem van de COTAN). De grootte en de scheve samenstelling van de beschikbare steekproeven bood voor dergelijk empirisch onderzoek ook maar weinig gelegenheid.

7. Hoe beoordeelt u alle aspecten van de wijze van samenstelling en verzameling van gegevens van de 'hoogrisicogroep', al dan niet middels de 'pre-screener'? Welke invloed hebben samenstelling en verzamelde gegevens van de 'hoogrisicogroep' op de resultaten van de E-screenertest?

De deelnemers aan het pilot-onderzoek zijn in de hoog-risicogroep respectievelijk laag-risicogroep ingedeeld op grond van de antwoorden op de pre-screener. Vervolgens is met de E-screener van alle deelnemers berekend of zij een hoog risico of een laag risico vormen voor het verlenen van een wapenvergunning. Met andere woorden, er is als het ware 'voorspeld' of zij tot de hoog- of laag-risicogroep behoren zoals vastgesteld met de pre-screener. De pre-screener bestond uit acht vragen. Voor een klein deel bevat de E-screener dezelfde of vergelijkbare vragen als de pre-screener. Dit zal zeker hebben bijgedragen aan het verhogen van de validiteit van de voorspelling. De grootte van dit 'spurieuse' effect is echter niet na te gaan; waarschijnlijk is het beperkt, gelet op het relatief gering aandeel van deze vragen in het totaal aantal vragen van de E-screener. Wat omgekeerd mogelijk validiteitsverlagend heeft gewerkt is het relatief lage responspercentage: mogelijk hebben juist de relatief 'brave' personen binnen de hoog-risicogroep die zijn uitgenodigd voor deelname hun medewerking toegezegd. Dit betreft dan personen die op de diverse onderdelen van de E-screener wat minder extreem scoren, waardoor de verschillen met personen in de laag-risicogroep kleiner zijn.
Een ander kritisch punt is dat de hoog-risicogroep voor circa de helft bestaat uit vrouwen, hetgeen niet in overeenstemming is met de samenstelling van de groep aanvragers van een wapenvergunning. Hiervoor werd gekozen omdat het zeer lastig bleek om voldoende deelnemers te werven die aan de vereisten voor de hoog-risicogroep voldeden (terzijde: in de laag-risicogroep was het percentage vrouwen 9% en in de groep vergunninghouders 3%). Wanneer werd uitgegaan van gemeenschappelijke beslisregels bleek bij vrouwen de specificiteit iets hoger en bij mannen de sensitiviteit iets beter, maar er werden geen grote verschillen geconstateerd. Concluderend kan worden gesteld dat de wijze waarop de groepen in het pilot-onderzoek zijn samengesteld niet ideaal is, maar dat dit wel het best haalbare scenario lijkt gelet op de complexiteit van dit type onderzoek.

8. Welke wijzigingen zijn er na 13 februari 2018 precies aan de E-screener toegevoegd, door wie en wanneer? Welke wijzigingen aan de E-screener heeft de Staat zelfstandig toegevoegd en wanneer zijn die doorgevoerd? Hoe verhouden deze wijzigingen zich met de conclusies en aanbevelingen van de TNO-managementrapportage van 13 februari 2018? Welke invloed heeft ieder van deze wijzigingen (per wijziging) gehad op de resultaten die de E-screener vanaf 1 oktober 2019 heeft gegenereerd?

De wijzigingen die in de E-screener na 13 februari 2018 zijn doorgevoerd staan beschreven in het rapport *TNO 2019 R11285 Inhoudelijke aanpassingen aan de E-screener: 'Psychische Gesteldheid van Wapenverlofaanvragers'*. De wijzigingen zijn doorgevoerd na overleg met vertegenwoordigers van het Ministerie van Justitie en Veiligheid (DGP&V, JUSTIS), de Nationale Politie, IPSOS, en TNO,

gehouden op 1 juli en 12 augustus 2019 en geïmplementeerd met ingang van 1 oktober 2019. De wijzigingen betreffen:

a. Inhoudelijke wijzigingen:
   - De schaal 'Emotionele stabiliteit' is toegevoegd.
   - 'Middelengebruik' is gesplitst in drie losse schalen, die elk een aparte risicofactor meten, in lijn met het 'Inlichtingenformulier' van de politie.
   - De formulering van het item met betrekking tot de risicofactor Psychiatrische opname is aangepast.
   - De formulering van het eerste item met betrekking tot de risicofactor Suïcidaliteit is aangepast.
b. Volgorde: In verband met de 'heftigheid' van enkele vragen zijn drie items verplaatst van het begin van de vragenlijst naar het eind van de vragenlijst.
c. Instructie: De aanvrager krijgt extra uitleg over praktische zaken en over de consequenties als hij of zij de test niet naar waarheid invult, waarbij een disclaimer is toegevoegd.
d. De reken- en beslisregels zijn aangepast op grond van de statistische analyses zoals beschreven in het rapport *TNO 2018 R10219 Validatie E-screening "Psychische Gesteldheid van Wapenverlofaanvragers"*.

De doorgevoerde wijzigingen zijn geheel in overeenstemming met de *Managementsamenvatting TNO-RAPPORT: TNO 2018 R10219.* Het effect van elk van de bovengenoemde wijzigingen apart op de resultaten sinds 1 oktober 2019 is niet onderzocht. Het percentage afwijzingen sinds 1 oktober 2019 is 7%. Volgens de afstelling in de oorspronkelijke versie zou circa 50% zijn afgewezen.
Op 13 januari 2020 zijn nog weer nieuwere rekenregels ingevoerd (zie het rapport *E-screener - technische toelichting op de afleiding van de einduitslag*) die vooral consequenties hebben voor de gewogen somscore. Uit de somscore werden een aantal variabelen verwijderd hetgeen mede consequenties had voor de gewichten van de wel opgenomen variabelen. Deze nieuwe afstelling leidt tot een percentage afwijzingen van 9% bij nieuwe aanvragers (per maart 2021, N = 2201; gegevens verstrekt door TNO per e-mail, document 20).
Voor de wijze waarop de kritische grenzen zijn vastgesteld zie de beschrijving en evaluatie bij het criterium Normen in de beoordeling in het vorige hoofdstuk.

9. Voldoet de E-screener aan het COTAN beoordelingssysteem voor de kwaliteit van tests (2010) en eventuele vergelijkbare normen?

Deze vraag is beantwoord in het vorige hoofdstuk.

10. Na verloop van welke termijn kan een E-screenerresultaat redelijkerwijs zijn werking verloren hebben?

Deze vraag hangt samen met 1) de psychometrische kwaliteit van de E-screener zelf, en 2) met de stabiliteit/veranderbaarheid van de kenmerken (constructen) die met de E-screener worden gemeten.
Met betrekking tot 1) geldt dat de normgroepen altijd aan slijtage onderhevig zijn. De normen zijn echter in 2017 verzameld en daarmee 'actueel' volgens het beoordelingssysteem van de COTAN (zie vraag 4.2 van de beoordeling).
Met betrekking tot 2) de stabiliteit/veranderbaarheid van de kenmerken (constructen) die tot doel hebben iemands psychische gesteldheid te meten in relatie tot wapenbezit, kan worden gesteld dat sommige van deze kenmerken duurzaam zijn (bijvoorbeeld een persoonlijkheidsstoornis, een voorgeschiedenis van eerdere suïcidepogingen), maar andere niet of veel minder (bijvoorbeeld alcohol- en drugsgebruik). Bij een negatief oordeel (afwijzing van de verlofaanvraag) hangt het antwoord op deze vraag af van de criteria waarop de afwijzing gebaseerd was (een stabiel versus een veranderbaar kenmerk). Bij een positief oordeel is de geldigheidsduur ook beperkt (hertesten blijft nodig).

11. Welke gegevens bevat de rapportage van een E-screener resultaat, ofwel de 'output' van de gemaakte test die (naar verzoeksters begrijpen) aan het ministerie van Justitie en Veiligheid wordt gezonden?

Het rapport wordt naar de politie verstuurd. Aan het ministerie van Justitie en Veiligheid wordt door IPSOS (dit is het bedrijf dat de afnames en de rapportages verzorgt) geen afschrift of een andere versie verstuurd. Het rapport dat aan de politie wordt gestuurd is gelijk aan de versie die de aanvrager zelf te zien krijgt. De rapportage van de E-screener bevat de volgende gegevens:

- De einduitslag weergegeven in de vorm van een rood of groen stoplicht. Als de uitslag rood is wordt een overzicht gegeven van de (combinatie van) risicofactor(en) die tot deze conclusie hebben geleid.
- De mate van betrouwbaarheid van de ingevulde vragenlijst weergegeven in een percentage.
- De uitslag wel of geen risico op vijf dichotome variabelen.
- De uitslag op zes schalen weergegeven in percentages.
- De uitslag op de (gewogen) totaalscore weergegeven in een percentage.

Blijkens de toelichting die door TNO per e-mail (document 20) is gegeven is er geen een-op-een relatie tussen de resultaten zoals in deze in de rapportage worden weergegeven en de beslissing rood of groen. Zo wordt de score op Medicijngebruik niet als afzonderlijke variabele gebruikt (maar wel als zodanig in de rapportage vermeld), maar in combinatie met de score op Psychose. Ook is de normgroep waarop de percentages zijn gebaseerd die worden weergegeven een andere dan de normgroep (in het pilot-onderzoek) waarin de 95%/5%-grenzen zijn vastgesteld. Uit een score boven de 95% op een van de genoemde variabelen in de rapportage hoeft dus geen negatief advies te volgen, wanneer deze score lager is dan de 95%-grens in het pilot-onderzoek. De percentielscores in de rapportage zijn bedoeld om aan deskundigen toch enig inzicht te geven in de wijze waarop een kandidaat scoort ten opzichte van andere aanvragers.

12. Is de e-escreener, gemeten naar professionele maatstaven, een deugdelijk hulpmiddel om de mogelijke aanwezigheid van risicofactoren te detecteren?

Vanwege (1) de in onze ogen vooralsnog onvoldoende onderbouwde psychometrische kwaliteit van de E-screener (zie het vorige hoofdstuk), en (2) de onduidelijkheid die er in de literatuur bestaat met betrekking tot de vraag of de in de E-screener opgenomen risicofactoren zonder andere factoren in ogenschouw te nemen voldoende basis kan vormen voor het nemen van een beslissing of de gescreende al dan niet een wapen mag dragen, zijn wij van mening dat de E-screener niet als 'stand alone' meetinstrument dient te worden gebruikt door een niet-psychologisch of niet-psychiatrisch geschoolde expert maar als hulpmiddel door psychologisch of psychiatrisch geschoolde experts. Wetenschappelijk onderzoek naar de mate waarin wapengeweld kan worden voorspeld door aan het individu gerelateerde psychische factoren is nog in volle ontwikkeling en kent allerlei beperkingen (zie ook de beantwoording van vraag 1 en de beoordeling van de theoretische uitgangspunten in het vorige hoofdstuk).

13 Is het in uw visie mogelijk verschillen in relevantie aan te geven tussen de risicofactoren die de E-screener onderzoekt? Verschillen deze risicofactoren in mate van voorspelbaarheid van incidenten bij wapengebruik?

Het is onzes inziens maar beperkt mogelijk hier wat over te zeggen. Uit consensus tussen experts over de psychische gesteldheid van wapenverlofaanvragers blijkt dat het vrijwel altijd een combinatie van soms drie of meer factoren betreft die relevant is, en gaat het niet om de afzonderlijke relevantie van de risicofactoren. Uit consensus tussen experts komt wel duidelijk naar voren dat bijvoorbeeld suïcidaliteit afzonderlijk een belangrijke voorspeller vormt. Empirisch onderzoek tot op heden met de E-screener kent nog te veel methodologische beperkingen (kleine steekproef in combinatie met een zeer lage *base rate* (proportie ongeschikte aanvragers) om op grond daarvan solide uitspraken over relevantieverschillen tussen de risicofactoren te kunnen doen.

Ook in de (internationale) wetenschappelijke literatuur zijn nauwelijks kwantitatieve gegevens (b.v. odds ratio's) beschikbaar met betrekking tot het onderlinge gewicht van verschillende risicofactoren, en als er kwantitatieve gegevens beschikbaar zijn, zijn deze gebaseerd op te onnauwkeurige en instabiele informatie (te kleine en niet representatieve steekproeven).

14. Is een negatieve score op de met behulp van de E-screener afgenomen test, bezien tegen de achtergrond van het antwoord op de vorige vraag, een deugdelijke reden om te zeggen dat er (geringe) twijfel mogelijk is aan het verantwoord zijn van vuurwapenbezit van de aanvrager?

Om deze vraag te kunnen beantwoorden is het nodig om de voorspellende kracht te kennen van de E-screener of om hier een inschatting van te kunnen geven. Immers, voor de kwaliteit van de E-screener speelt de voorspellende kracht (predictieve validiteit) een centrale rol. Omdat deze predictieve validiteit als onvoldoende is beoordeeld (zie vraag 7, criteriumvaliditeit, van de beoordeling volgens de COTAN-criteria), is een negatieve score niet per se een deugdelijker reden dan een positieve score om te zeggen dat er (geringe) twijfel mogelijk is aan het verantwoord zijn van vuurwapenbezit van de aanvrager.
Tegelijkertijd moet de betekenis van de voorspellend kracht van de E-screener op een realistische manier in de context worden geplaatst waarin de test wordt gebruikt, namelijk een context waarin de base rate (proportie ongeschikte aanvragers) extreem laag is, en het percentage aanvragers dat een vergunning krijgt (de 'selectieratio') bijzonder hoog. Een extreem lage base rate en een extreem hoge selectieratio impliceren dat inspanningen om de voorspellende kracht van de E-screener te verbeteren niet per se tot sterk verschillende uitkomsten hoeven te leiden. Dit kan gedemonstreerd worden aan de hand van een hypothetisch scenario uit de klassieke Taylor-Russell (1939) verwachtingstabellen.
Bij een base rate van 0.05 (5% van de aanvragers is ongeschikt) en een selectieratio van 95%, levert een test (bijvoorbeeld de E-screener) met een voorspellende kracht van r=.30 ten opzichte van het niet gebruiken van een test, een verbetering van 0.5% op van de kans dat het toewijzen van de vergunning aan een aanvrager terecht is (een verbetering van 0.5% in de succes ratio). In een hypothetisch getallenvoorbeeld waarbij er 1000 aanvragers zijn, neemt dan het aantal aanvragers dat terecht een vergunning krijgt toe van 903 naar 907, het aantal aanvragers dat ten onrechte een vergunning krijgt af van 47 naar 43, het aantal aanvragers dat ten onrechte wordt afgewezen af van 48 naar 43, en het aantal aanvragers dat terecht wordt afgewezen toe van 2 naar 7. Zonder E-screener worden er in dit geval in totaal 905 correcte beslissingen genomen (en 95 incorrecte) en met E-screener 914 correcte beslissingen (en 86 incorrecte). Dit is een toename van 9 correcte beslissingen.
Echter, met betrekking tot de E-screener is het wellicht realistischer dat er wordt uitgegaan van een base rate van 0.001 (0,1% van de aanvragers is ongeschikt, dat wil zeggen 1 op de 1000) en een selectieratio van 93% (7% van de aanvragers wordt afgewezen). Deze base rate is extreem laag, terwijl de selectieratio tegelijkertijd hoog is. Een scenario waarin de E-screener een behoorlijk goede voorspellende kracht zou hebben volgens de maatstaven van het onderzoeksdomein naar iemands psychische gesteldheid - bijvoorbeeld r=.35 - levert ten opzichte van het niet gebruiken van een test (dit staat gelijk aan een validiteit van .00) een verbetering van 0.1% op van de kans dat het toewijzen van de vergunning aan een aanvrager terecht is (een verbetering van 0.1% in de succes ratio). In een hypothetisch getallenvoorbeeld waarbij er 1000 aanvragers zijn, neemt in dit geval het aantal aanvragers dat terecht een vergunning krijgt toe van 929 naar 930, het aantal aanvragers dat ten onrechte een vergunning krijgt af van 1 naar 0, het aantal aanvragers dat ten onrechte wordt afgewezen af van 70 naar 69, en dat terecht wordt afgewezen toe van 0 naar 1. Zonder E-screener worden er in dit geval 929 correcte beslissingen genomen (en 71 incorrecte) en met E-screener 931 correcte beslissingen (en 69 incorrecte). Dit is een toename van 2 correcte beslissingen.
Uit deze getallenvoorbeelden wordt duidelijk hoe moeilijk het is om bij een meer extreme, maar waarschijnlijk meer realistische, base rate van 0.001 (in plaats van 0.05) een betekenisvolle bijdrage te leveren aan het beslissingsproces ook al neemt de validiteit van het selectie-instrument toe van

.00 naar .35. In het tweede voorbeeld is de winst in het aantal correcte voorspellingen dat wordt bereikt miniem. Het is een sociaal-maatschappelijke vraag en geen technische vraag welke waarde er wordt gehecht aan dergelijke kleine uitkomstverschillen.
Het zal overigens, gezien het feit dat het bijzonder lastig zal zijn om gedegen empirisch onderzoek te doen aan de hand van voldoende grote en representatieve steekproeven, op de korte en wellicht middellange termijn niet goed mogelijk zijn om na te gaan of de voorspellende kracht van de E-screener daadwerkelijk kan verbeteren/ is verbeterd.

15. *Heeft u verder nog aan- of opmerkingen met betrekking tot de E-screener?*

De procedures die door het Trimbos Instituut en door TNO zijn gevolgd voor het ontwikkelen en valideren van de E-screener zijn naar ons oordeel zeer grondig. Tegelijkertijd is het in onze ogen vrijwel ondoenlijk om een E-screener te ontwikkelen die volgens professionele maatstaven zoals worden gehanteerd door de COTAN geheel deugdelijk is. Dit heeft te maken met een combinatie van de volgende zaken: 1) de zeer lage base rate (het percentage 'hoog risico') binnen de groep aanvragers, 2) de relatief kleine steekproef die beschikbaar is en op de middellange termijn beschikbaar zal zijn om voorspellende statistische analyses uit te voeren, en 3) het uitgangspunt dat risicofactoren veelal niet afzonderlijk, maar in onderlinge samenhang bestudeerd dienen te worden bij het doen van een voorspelling. Toch is het hanteren van dit meetinstrument, dat gebreken vertoont, naar onze mening beter dan het niet gebruiken van dit instrument, mits de E-screener wordt gebruikt door psychologische of psychiatrische experts die ook de beperkingen van de E-screener goed op waarde weten te beoordelen. We komen mede tot deze conclusie omdat de geconstateerde gebreken niet uniek zijn voor de E-screener en in meer of mindere mate voor elke andere procedure zullen gelden.
Een fundamentele vraag die niet met psychometrisch-technische of wetenschappelijke beslissingen te maken heeft is het toekennen van *waarden* aan het al dan niet terecht toewijzen of afwijzen van een wapenverlofaanvraag. Dit is een politiek-maatschappelijk vraagstuk dat discussie behoeft tussen de belanghebbenden. In onderstaande tabel worden ter illustratie twee fictieve voorbeelden gegeven met mogelijke waarden die twee belanghebbende instanties respectievelijk toekennen aan de uitkomsten op een fictieve schaal lopend van -3 (zeer onwenselijk) via 0 (neutraal) tot +3 (zeer wenselijk), en waarin de waardentoekenning deels verschilt tussen beide partijen. In het eerste voorbeeld wordt er vooral veel belang aan gehecht dat in de groep die eigenlijk geen wapenvergunning zou mogen krijgen de juiste beslissingen worden genomen: aan het terecht afwijzen of onterecht verlenen van een wapenvergunning wordt een hoge waarde toegekend. Aan de juistheid van de beslissingen die genomen worden in de groep die eigenlijk zonder risico wel een wapenvergunning zouden moeten krijgen wordt minder belang gehecht. Een dergelijke waardentoekenning kan ertoe leiden dat aan relatief weinig personen een wapenvergunning wordt toegekend. In het tweede voorbeeld wordt aan de juistheid van beslissingen in beide groepen evenveel waarde toegekend. Dit kan leiden tot een soepeler beleid ten aanzien van het verlenen van een vergunning. Het kan zinvol zijn om in politiek-maatschappelijke discussies deze waardentoekenningen van verschillende partijen te expliciteren.

	Geen/ laag risico	Hoog risico
Vergunning verlenen	+1	-3
Vergunning afwijzen	-1	+3

	Geen/ laag risico	Hoog risico
Vergunning verlenen	+3	-3
Vergunning afwijzen	-3	+3

Los van de bovenstaande problematiek verdient het aanbeveling om na te gaan of de betrouwbaarheid en begripsvaliditeit van de E-screener naar aanleiding van de beoordeling volgens de COTAN-richtlijnen kunnen verbeteren en of er meer informatie kan worden verkregen over de fairness van de E-screener. Bovendien kan het de moeite waard zijn de acceptatie van de E-screener te peilen door een survey af te nemen waarin 'applicant perceptions' onder vergunningaanvragers worden gemeten volgens het model van Gilliland (1993). Dit model maakt een onderscheid tussen oordelen over onder andere de relevantie, voorspellende kracht, eerlijkheid, en geloofwaardigheid, en het gemak en in hoeverre het aangenaam is om de E-screener te maken.

Literatuur

Auxemery, Y. (2015). The mass murderer history: modern classifications, sociodemographic and psychopathological characteristics, suicidal dimensions, and media contagion of mass murders. *Comprehensive Psychiatry, 56,* 149-54.

Baumann, M. L., & Teasdale, B. (2018). Severe mental illness and firearm access: Is violence really the danger? *International Journal of Law and Psychiatry, 56,* 44-49.

Betz, M. E., Azrael, D., Johnson, R. L., Knoepke, C. E., Ranney, M. L., Wintemute, G. J., Matlock, D., Suresh, K., & Miller, M. (2020). Views on Firearm Safety Among Caregivers of People with Alzheimer Disease and Related Dementias. *JAMA Network Open, July 1;3(7):e207756.*

Byers, A. L., Li, Y., Barnes, D. E., Seal, K. H., Boscardin, W. J., & Yaffe, K. (2020). A national study of TBI and risk of suicide and unintended death by overdose and firearms. *Brain Injury, 34*(3), 328-334.

Carter, P. M., Zimmerman, M. A., & Cunningham, R. M. (2021). Addressing key gaps in existing longitudinal research and establishing a pathway forward for firearm violence prevention research. *Journal of Clinical Child & Adolescent Psychology, 50*(3), 367-384.

De Moore, G. M., Plew, J. D., Bray, K. M., & Snars, J. N. (1994). Survivors of self-inflicted firearm injury. A liaison psychiatry perspective. *Medical Journal of Australia, 160*(7), 421-425.

Dudley, M., Cantor, C., & de Moore, G. (1996). Jumping the gun: firearms and the mental health of Australians. *Australian & New Zealand Journal of Psychiatry, 30*(3), 370-381.

EFPA (2012). *EFPA Standards for Test Use.* Brussel: EFPA.

EFPA (2013). *EFPA Review Model for the description and evaluation of psychological and educational tests.* Brussel: EFPA.

Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2010). *COTAN beoordelingssysteem voor de kwaliteit van tests (gewijzigde herdruk mei 2010).* Utrecht: NIP.

Gilliland, S. W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *Academy of Management Review, 18*(4), 694-734.

Glick, I. D., Cerfolio, N. E., Kamis, D., & Laurence, M. (2021). Domestic Mass Shooters: The Association with Unmedicated and Untreated Psychiatric Illness. *Journal of Clinical Psychopharmacology, 41*(4),366-369.

Krüsselman, K., Aarten, P., & Liem M (2021). Firearms and violence in Europe – A systematic review. *PloS ONE 16(4): e0248955.*

McGinty, E. E., Webster, D. W., & Barry, C. L. (2014). Gun policy and serious mental illness: priorities for future research and policy. *Psychiatriac Services, 65*(1), 50-58.

Nederlands Instituut van Psychologen (2018). *Algemene Standaard Testgebruik NIP 2017.* Amsterdam: NIP/COTAN.

Paulhus, D. L. (1998). Interpersonal and intrapsychic adaptiveness of trait self-enhancement: A mixed blessing? *Journal of Personality and Social Psychology, 74,* 1197-1208.

Pinals, D. A., & Anacker, L. (2016). Mental Illness and Firearms: Legal Context and Clinical Approaches. *Psychiatry Clinical North America Journal, 39*(4), 611-621.

Pineda, D., Galan, M., Martinez-Martinez, A., Campagne, D. M., & Piqueras, J. A. (2021). Same personality, new ways to abuse: how dark tetrad personalities are connected with cyber intimate partner violence. *Journal of Interpersonal Violence, 886260521991307.*

Rozel, J. S., & Mulvey, E. P. (2017). The Link Between Mental Illness and Firearm Violence: Implications for Social Policy and Clinical Practice. *Annual Review of Clinical Psychology, 13,* 445-469.

Sanchez, C., Jaguan, D., Shaikh, S., McKenney, M., & Elkbuli, A. (2020). A systematic review of the causes and prevention strategies in reducing gun violence in the United States. *American Journal of Emergency Medicine, 38*(10), 2169-2178.

Shenassa, E. D., Rogers, M. L., Spalding, K. L., & Roberts, M. B. (2004). Safer storage of firearms at home and risk of suicide: a study of protective factors in a nationally representative sample. *Journal of Epidemiology and Community Health, 58*(10), 841-8.

Swanson, J. W., McGinty, E. E., Fazel, S., & Mays, V. M. (2015). Mental illness and reduction of gun violence and suicide: bringing epidemiologic research to policy. *Annals of Epidemiology, 25*(5), 366-376.

Swinson, N., Flynn, S. M., While, D., Roscoe, A., Kapur, N., Appleby, L., & Shaw, J. (2011). Trends in rates of mental illness in homicide perpetrators. *British Journal of Psychiatry, 198*(6), 485-489.

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: discussion and tables. *Journal of Applied Psychology, 23*(5), 565.

Van der Knaap, L. M., Alberda, D. L., Oosterveld, P., & Born, M. P. (2012). The predictive validity of criminogenic needs for male and female offenders: Comparing the relative impact of needs in predicting recidivism. *Law and Human Behavior, 36*(5), 413-422.

Bijlage 1 Vragen interview met a b van het Trimbos instituut

Vragen aan [redacted] (Trimbos) op dinsdag 10 augustus 2021.

1. Bij de ontwikkeling van de E-Screener is er sprake van twee klankbordgroepen, een groep van 12 (14) experts en in de laatste fase een groep van 9 experts. Kan er iets meer gezegd worden over de deskundigheid van deze experts? Was er overlap tussen leden van de klankbordgroepen, de 1e groep experts en de 2e groep experts?

2. Is er informatie over de consensus versus de spreiding van de experts binnen en tussen de klankbordgroepen beschikbaar/ Kan hier iets over worden gezegd? Valt er iets te zeggen over het besluitvormingsproces binnen de klankbordgroepen en de groep deskundigen bij de totstandkoming van de lijsten met risicofactoren (zijn onafhankelijke oordelen over het belang van de factoren geïntegreerd, bijv., of verliep de procedure anders?)?

3. Hoe zagen de lijsten met voorlopige selecties van risicofactoren van enerzijds de klankbordgroepen en anderzijds de groep deskundigen eruit? Hoe is de men tot de selectie uit deze twee lijsten gekomen? Waarom werden sommige (typen van) factoren wel (en welke waren dat) en andere niet (en welke waren dat) als PM-factor meegenomen? Welke (soort) factoren zijn afgevallen omdat ze niet meetbaar werden geacht ook al werden ze wel relevant bevonden? En welke factoren zijn in de laatste fase (de bijeenkomst met 9 experts) afgevallen en waarom? Welke inclusie- versus exclusiecriteria zijn er gehanteerd voor het al dan niet meenemen van de factoren?

4. De kritische score voor variabele A[4] is gebaseerd 'op normscores uit de literatuur'. Is daar iets meer over te zeggen, hoe zijn deze tot stand gekomen? Op welke wijze is hierbij rekening gehouden met verschillen tussen de (voornamelijk Amerikaanse) literatuur en de Nederlandse/ Europese context?

5. Zijn er overwegingen geweest met betrekking tot het tijdelijk vs. permanent zijn van de risicofactoren?

6. In hoeverre is er sprake van een weging van risicofactoren?

7. Klopt het dat (lichte) verstandelijke beperking, niet-aangeboren hersenletsel en neurodegeneratieve aandoeningen niet zijn meegenomen als risicofactoren?

[4] Ten behoeve van de geheimhouding van de inhoud van de E-screener is de werkelijke naam van genoemde variabele vervangen door variabele A.

Bijlage 2 Vragen interview met [redacted] van TNO

Vragen aan TNO mbt constructie, gebruik en validering E-Screener.

1. Wat is een Lage ambtenaar en wat een Hoge ambtenaar? Welke kennis of kwalificaties dienen zij te bezitten?

2. Hoe wordt nu feitelijk een beslissing genomen met de E-screener?

3. Waarom is er gekozen voor de verdeling 95%/5% als grens die gebruikt wordt bij sommige variabelen in het beslissingsproces mbv de E-Screener?

4. Waarom wordt er tav deze variabelen een andere grens gehanteerd dan in de rapportage?

5. Zijn tbv de normering de 'representatieve' groep (n=303) samengevoegd met de groep wapenverlofhouders (n=103), maw zijn de normen op deze 2 groepen gezamenlijk berekend?

6. Zijn gemiddelden en sd's van alle groepen (representatieve groep, wapenverlofhouders en hoogrisicogroep) op alle variabelen beschikbaar?

7. Welke grensscore wordt feitelijk gebruikt als grensscore voor schaal A[5]? (op verschillende plekken in de documentatie wordt daar tegenstrijdige info over gegeven).

8. De kritische score voor schaal B is kennelijk overgenomen uit het Trimbos-rapport en is gebaseerd 'op normscores uit de literatuur'. Is daar iets meer over te zeggen, hoe zijn deze tot stand gekomen?

9. Ten behoeve van het rapport worden voor schaal C en schaal D absolute kritische scores vastgesteld. Voor de kritische score bij schaal C wordt verwezen naar normscores in de literatuur zonder verdere onderbouwing. Voor schaal D wordt de kritische score (X) vastgesteld zonder nadere motivatie. Is hier iets meer over te zeggen, hoe zijn deze tot stand gekomen?

10. Zijn er ROC-curves voor de totaalscore beschikbaar? Zijn er gegevens beschikbaar voor de sensitiviteit en specificiteit bij alternatieve kritische grenzen voor de hele procedure?

11. Is de betrouwbaarheid van de totaalscore berekend?

12. Op welke steekproef zijn de alfa's berekend? Zijn hiertoe laag- en hoog-risicogroep samengevoegd?

13. Op welke steekproef zijn de correlaties met andere instrumenten ter bepaling van de begripsvaliditeit berekend? Zijn hiertoe laag- en hoog-risicogroep samengevoegd?

14. Hoe moeilijk was het doen van het empirische onderzoek (was het goed te doen of was het extreem lastig bijv de dataverzameling etc)?

15. Is er overwogen om configureel denken (dus effect van evt toxische combinaties) te toetsen?

[5] Ten behoeve van de geheimhouding van de inhoud van de E-screener zijn de werkelijke namen van de schalen vervangen door schaal A, schaal B, enz.