Betrouwbaarheid schattingsmethodiek PTSS-protocol

De intra- en interdoktervariatie van de schattingsmethodiek ter beoordeling van de mate van invaliditeit bij psychische aandoeningen bij post-actieve militairen.

Onderzoeksteam TGO UMCG
Dr. Andrea Fokkens
Prof. dr. Johan Groothoff
Prof. dr. Jac van der Klink
Dr. Jolanda Tuinstra

Met dank aan: drs. Nienke Verheij, drs. Lex van de Ven, dr. Roel Popping, dr. Roy Stewart, dr. Sandra Brouwer en dr. Roel Bakker

Toegepast GezondheidsOnderzoek (TGO)
Afdeling Gezondheidswetenschappen
Universitair Medisch Centrum Groningen
Groningen, december 2011

©TGO / UMCG

Belangrijkste conclusies van dit onderzoek
• **De *intra*dokterbetrouwbaarheid van de schattingsmethodiek is goed tot zeer goed.** • **De *inter*dokterbetrouwbaarheid van de schattingsmethodiek is goed.** • **De uitkomst van het schattingsinstrument bij de beoordeling van de ernst van beperkingen bij psychische aandoeningen van een post-actieve militair is onafhankelijk van de - getrainde - verzekeringsarts die het instrument toepast.**
Over het onderzoek
• **Alle 24 gefilmde beoordelingsgesprekken bevatten voldoende informatie voor inclusie in deze studie.** • **Alle 24 films zijn door 5 onafhankelijke verzekeringsartsen 2 keer (na periode van 2 maanden) bekeken, zij hebben met behulp van de schattingsmethodiek de aard van de beperkingen gescoord.** • **De non-respons analyse geeft geen aanleiding om te veronderstellen dat er selectie heeft plaatsgevonden in het onderzoek. Resultaten en conclusies geven een representatief beeld voor de groep waarbij het schattingsinstrument in de praktijk wordt toegepast.**

Samenvatting

Achtergrond van het onderzoek en de nieuwe schattingsmethodiek

Sinds 1 juli 2008 gebruikt het Zorgloket Militaire Oorlogs-en Dienstslachtoffers (MOD) in Utrecht een nieuwe schattingsmethodiek voor het bepalen van beperkingen en van het percentage invaliditeit bij militairen met psychische aandoeningen. De nieuwe schattingsmethodiek maakt deel uit van een protocol dat in stappen is ontwikkeld door verschillende commissies[1].

De schattingsmethodiek bestaat uit een scoringslijst die na een verzekeringsgeneeskundige beoordeling wordt ingevuld door een verzekeringsarts van het Zorgloket MOD. De schattingsmethodiek is onderverdeeld in vier rubrieken, die vervolgens weer onderverdeeld zijn in totaal tien subrubrieken.

De tabel (volgende pagina) geeft alle rubrieken en de bijbehorende subrubrieken weer.

[1] Commissie WPC-schaal, 2004; Commissie WPC-PIM, 2005; Medische Werkgroep Project WIA ZORG wet Defensie, 2007.

De vier rubrieken met de tien subrubrieken van de schattingsmethodiek

Rubriek	*Subrubriek*
1. Activiteiten van het dagelijkse leven	1. Persoonlijke hygiëne en zelfzorg 2. Mobiliteit 3. Slapen 4. Seksuele functie
2. Sociaal functioneren	5. Basale communicatie 6. Communicatief sociaal functioneren 7. Communicatieve (emotionele) vaardigheid
3. Concentratie, doorzettingsvermogen en tempo	8. Structuur aanbrengen 9. Huishoudelijke activiteiten
4. Adaptatie aan stressvolle omstandigheden	10. Omgaan met stressvolle gebeurtenissen

De ernst van de beperkingen zijn per subrubriek beschreven en geclassificeerd, variërend van klasse 0 'normaal', tot en met klasse 5 'extreme beperkingen'. Op basis van de gescoorde klassen op de tien subrubrieken wordt het invaliditeitspercentage berekend.

Onderzoeksvragen van de betrouwbaarheidsstudie

Bij de invoering van het protocol en de schattingsmethodiek in de praktijk, is besloten om een betrouwbaarheidsstudie uit te voeren naar de schattingsmethodiek. Een onderzoeksteam van Toegepast GezondheidsOnderzoek (TGO) van het UMCG heeft het onderzoek uitgevoerd.
De belangrijkste vraagstelling van het onderzoek is:

Is de nieuwe schattingsmethodiek ter beoordeling van de mate van invaliditeit bij psychische aandoeningen betrouwbaar en zo ja in welke mate?

Methode van het onderzoek

Voor het onderzoek zijn in totaal 24 beoordelingsgesprekken tussen een post-actieve militair en een verzekeringsarts van het Zorgloket MOD –na goedkeuring van de post-actieve militair- gefilmd. Iedere film is 2 keer bekeken en gescoord door vijf andere, onafhankelijke verzekeringsartsen die getraind zijn in het toepassen van de schattingsmethodiek. Een geringe variatie in het oordeel binnen een arts (na 2 maanden) en tussen artsen zijn goede indicaties voor een betrouwbaar instrument.

Resultaten en conclusies van het onderzoek

Overall conclusies

Er blijkt weinig variatie te bestaan tussen twee beoordelingen die één arts van hetzelfde gefilmde beoordelingsgesprek heeft gedaan (*intra*doktervariatie). Dit geldt ook wanneer verschillende artsen hetzelfde gefilmde beoordelingsgesprek beoordelen (*inter*doktervariatie). De eindconclusie van dit onderzoek is dat de *intra*dokterbetrouwbaarheid van het nieuwe schattingsinstrument goed tot zeer goed en de *inter*dokterbetrouwbaarheid goed is.
Analyse van de gegevens van post-actieve militairen die geen medewerking aan het onderzoek hebben verleend geeft geen aanleiding om te veronderstellen dat er een vertekenende selectie heeft plaatsgevonden in het onderzoek. De resultaten en de conclusies geven daarom een

representatief beeld van de groep waarbij het schattingsinstrument in de praktijk wordt toegepast.
De bovengenoemde conclusies zijn gebaseerd op de volgende resultaten.

Verdeling van de gescoorde klassen per subrubriek
Niet alle klassen waarbinnen kon worden gescoord op het instrument om de ernst van de beperkingen aan te geven, zijn evenredig vaak van toepassing, zodat sprake is van een scheve verdeling van deze klassen. Zo is er binnen klasse 0, welke staat voor "normaal", relatief veel gescoord. Klasse 5, welke staat voor "extreme beperkingen", is echter nooit gescoord. De tussenliggende klassen 1 tot en met 4 zijn wel gescoord, met een duidelijk accent op de klassen 1 en 2.
Bij de subrubrieken 3 'Slapen', 10 'Omgaan met stressvolle gebeurtenissen', 6 'Sociale communicatie' en 7 'Communicatieve vaardigheden' zijn relatief vaker hogere klassen gescoord. Dit geeft aan dat er binnen deze subrubrieken vaker beperkingen worden gescoord door de onafhankelijke verzekeringsartsen.

Intradoktervariatie
Over het geheel genomen is er sprake van een geringe variatie in het oordeel van verzekeringsartsen die twee keer dezelfde film hebben beoordeeld. De *intra*doktervariatie is gering. Met andere woorden, de overeenstemming binnen een arts tussen de gescoorde klasse op de 1^{e} meting in vergelijking met de gescoorde klasse –voor dezelfde film- op de 2^{e} meting is hoog.
Indien alle films en alle artsen tezamen worden beschouwd, blijkt het percentage overeenstemming (kwadratisch gewogen) binnen een arts hoog te zijn en varieert dit voor alle subrubrieken van 84% - 100%.
Wanneer een arts verschillend scoort op de 2^{e} meting in vergelijking met de 1^{e} meting dan is dit verschil, per subrubriek, in minder dan 6% 2 klassen of meer.

Interdoktervariatie
Over het geheel genomen is de variatie in gescoorde klassen *tussen* vijf verschillende artsen hoger dan de variatie in gescoorde klassen indien één arts dezelfde film twee keer beoordeelt. Dit geldt vooral voor de subrubrieken 6 'Sociale communicatie' en 7 'Communicatieve vaardigheden'. Bij deze subrubrieken wordt respectievelijk 25% en 41.7% overeenstemming gevonden tussen vier of vijf artsen. Wanneer er een verschil wordt gevonden tussen de vijf artsen, is de afwijking van de vijfde arts meestal, 72% van de gevallen maar 1 klasse.
Het percentage overeenstemming (gepaard kwadratisch gewogen) in gescoorde klassen voor alle films en alle artsen is hoog, voor alle subrubrieken variërend van 92% - 100%. Overall is de *inter*doktervariatie klein.

Invaliditeitspercentage, de uitkomstmaat IP
De variatie tussen vijf artsen bij de bepaling van het afgeronde invaliditeitspercentage (IP) per film is relatief klein: bij 17 van de 24 films is deze spreiding maximaal 10%. De spreiding in het IP voor de 7 films waar de variatie groter is dan 10% wordt voor het grootste deel verklaard door spreiding op de subrubriek 6 'Sociale communicatie', 7 'Communicatieve vaardigheden' en 8 'Structuur aanbrengen'.

Wanneer voor deze 7 films de 5e meest afwijkende score wordt weggelaten, wordt bij 22 van de 24 films een spreiding van maximaal 10% gevonden.

Algemene discussiepunten

Het huidige onderzoek betreft een betrouwbaarheidstudie van een schattingsinstrument. Deze studie is uitgevoerd door de *intra*doktervariatie en de *inter*doktervariatie te bestuderen. Weinig variatie binnen een verzekeringsarts en tussen verschillende verzekeringsartsen is een goede indicatie voor een betrouwbaar instrument. Immers, wanneer verschillende artsen hetzelfde instrument toepassen bij het beoordelen van eenzelfde situatie (lees: gefilmd beoordelingsgesprek) dan veronderstelt een betrouwbaar instrument dat zij bij gelijke bevindingen tot dezelfde scores komen. Een belangrijke aanvulling hierbij is, dat een goede betrouwbaarheid van een instrument een beperkte betekenis heeft voor een uitspraak over de validiteit van dat instrument. Bij de validiteit gaat het over de vraag of het instrument ook daadwerkelijk meet wat het beoogt te meten. Een instrument dat niet betrouwbaar is en bij herhaald invullen onder exact dezelfde omstandigheden tot andere resultaten leidt kan nooit valide zijn. Een goede betrouwbaarheid is in die zin een van de onderdelen van het bepalen van de validiteit van het instrument.

Slotconclusie

Op basis van de resultaten van het huidige onderzoek kan worden geconcludeerd dat de nieuwe schattingsmethodiek een betrouwbaar instrument is. Met andere woorden, de uitkomst van het schattingsinstrument bij de beoordeling van de ernst van beperkingen bij militairen met psychische aandoeningen is onafhankelijk van de –getrainde- verzekeringsarts die het instrument toepast.

Inhoudsopgave *p.*

Hoofdstuk 1 Achtergrond van het onderzoek…… 1
1.1 Inleiding 1
1.2 Probleemstelling en onderzoeksvragen 4
1.3 Organisatie en betrokken partijen 4
Hoofdstuk 2 De schattingsmethodiek…… 5
2.1 Beschrijving van de schattingsmethodiek 5
2.2 Van beperkingen naar invaliditeitspercentage 5
Hoofdstuk 3 Methoden…… 7
3.1 Opzet 7
3.2 Het gefilmde beoordelingsgesprek 7
3.2.1 De in- en exclusie criteria 7
3.2.2 Het inclusieproces 8
3.2.3 De verzekeringsartsen van het Zorgloket (VA-U) 9
3.2.4. De techniek waarmee het gesprek wordt gefilmd 10
3.3 De onafhankelijke verzekeringsartsen (VA-G) 10
3.4 Het onderzoeksdesign en de uitvoering daarvan in de vorm van O&S dagen 10
3.4.1 Een pilot 11
3.4.2 Wat vullen de VA-G in? 11
3.5 De respons 12
3.5.1. Non-respons analyse 12
3.6 Analyses 14
3.6.1 Beschrijvende statistieken en Ag09 14
Hoofdstuk 4 Resultaten…… 15
4.1 Beschrijving van de onderzoeksdata 15
4.1.1 Kenmerken van de onderzoeksgroepen 15
4.1.2 “Niet aan de orde gekomen” subrubrieken 16
4.1.3 Verdeling van klassen per subrubriek 17
4.2 Intradoktervariatie 18
4.2.1. Intradoktervariatie en percentage overeenstemming 20
4.3 Interdoktervariatie 22
4.3.1 Interdoktervariatie en percentage overeenstemming 25
4.4 Invaliditeitspercentage, IP 27
4.4.1 Totstandkoming IP 31
Hoofdstuk 5 Conclusies en discussie…… 35
5.1 Conclusies 35
5.2 Discussie 37
Bijlagen…… 41
1 Achtergrond van de opdracht van het Ministerie van Defensie 43
2 Organisatie en betrokken partijen 45
3 Folder met toelichting voor de cliënt 46
4 Toestemmingsformulier cliënt 48
5 Niet-deelname formulier cliënt 49
6 Formulier achtergrond informatie 50
7 Opzet onderzoek 51
8 Informatie over analyse techniek 52
9 Tabellen 2e meting 55
10 IP, hypothetische berekeningen 59

Hoofdstuk 1 Achtergrond van het onderzoek

1.1 Inleiding

Sinds 1 juli 2008 gebruikt het Zorgloket MOD (Militaire Oorlogs- en Dienstslachtoffers) een nieuwe schattingsmethodiek voor het bepalen van het percentage invaliditeit bij militairen met beperkingen door psychische aandoeningen. Deze methodiek is ontwikkeld op basis van het PIM-rapport[2] (juni 2005) dat richtlijnen bevat voor de beoordeling van de invaliditeit van militairen. Direct gerelateerd aan het PIM-rapport is het PTSS-protocol[3]. Het PTSS-protocol benut de richtlijnen en aanbevelingen van het PIM-rapport volledig en kan worden toegepast bij de beoordeling van het Militair invaliditeitspensioen bij PTSS (Post Traumatische Stress Stoornis) en andere psychische aandoeningen. In het kader van de militaire pensioenwetgeving is het noodzakelijk dat beperkingen bij psychische aandoeningen worden vertaald in een invaliditeitspercentage.

Het Ministerie van Defensie heeft de militaire vakbonden en de Tweede Kamer toegezegd de betrouwbaarheid van de nieuwe schattingsmethodiek zoals verwoord in het PTSS-protocol en het PIM rapport te onderzoeken. Daarbij is het van belang dat de onderzoeksopzet, de uitvoering, de analyses en de rapportage op een wetenschappelijk verantwoorde wijze gestalte krijgen. Verder zijn onafhankelijkheid en professionele verantwoordelijkheid bij de uitvoering van belang. Het Ministerie van Defensie heeft daarom een extern wetenschappelijk centrum de opdracht gegeven om de betrouwbaarheidsstudie uit te voeren. Daarbij zijn vooraf door de opdrachtgever keuzes gemaakt om de onderzoeksopdracht af te bakenen en nader te definiëren (bijlage 1).

De betrouwbaarheidsstudie wordt uitgevoerd door een onderzoeksteam van Toegepast GezondheidsOnderzoek (TGO), onderdeel van het Universitair Medisch Centrum Groningen, waarbij tegemoet gekomen wordt aan de voorwaarden van de opdrachtgever. De centrale vraag van de betrouwbaarheidsstudie luidt: “Is de nieuwe schattingsmethodiek ter beoordeling van de mate van invaliditeit bij psychische aandoeningen betrouwbaar en zo ja in welke mate?”. Voor het bepalen van de betrouwbaarheid wordt gebruik gemaakt van filmopnames van het beoordelingsgesprek. De opnames van dit gesprek worden door meerdere onafhankelijke artsen bekeken, die afzonderlijk de beperkingen scoren met behulp van de nieuwe schattingsmethodiek. Het zijn deze scores die worden gebruikt om de *intra*doktervariatie en *inter*doktervariatie te analyseren en de betrouwbaarheid van de schattingsmethodiek te bepalen. Belangrijke voorwaarden van de onderzoeksopdracht zijn dat (1) de causaliteit en de diagnose buiten de focus van het huidige onderzoek liggen, en (2) er geen vergelijking gemaakt wordt van uitkomstmaten tussen de verzekeringsartsen van het Zorgloket MOD (VA-U[4]) en de onafhankelijke verzekeringsartsen (VA-G[5]).

[2] PIM staat voor Psychische Invaliditeit Militairen. Het betreft het rapport *Schade in Schalen – de psyche in beeld. Het vaststellen van de mate van psychische invaliditeit in het kader van de militaire pensioenvoorschriften*. Advies Commissie WPC-PIM, ’s Gravenhage, juni 2005

[3] PTSS protocol – PTSS WIA IP Protocol, Project WIA ZORG wet, Defensie 2007

[4] De verzekeringsartsen die werken in het Zorgloket MOD in Utrecht worden verder in deze rapportage aangeduid met VA-U.

[5] De verzekeringsartsen die door het onderzoeksteam in Groningen zijn ingezet om de gefilmde gesprekken te zien en te scoren, worden verder in deze rapportage aangeduid met VA-G.

Schets van de literatuur

In de literatuur wordt de behoefte aan goede betrouwbaarheidsstudies in de geneeskunde en het bijzonder de sociale geneeskunde en verzekeringsgeneeskunde regelmatig benadrukt. Over de betrouwbaarheid van de binnen het verzekeringsgeneeskundig proces toegepaste instrumenten[6,7] en interviews[8] worden verschillende gegevens gepresenteerd.

Onderzoek van Brouwer en e.a. (2003) is een voorbeeld van de *intra*- en *inter*-beoordelaarsbetrouwbaarheid van het 'Belastbaarheidspatroon' (FIS-BLP) en de 'Functionele mogelijkhedenlijst'(FML) voor de beoordeling van arbeidsgerelateerde functionele status bij patiënten met chronische aspecifieke lage-rugpijn. Zij concluderen dat op basis van de kappa-berekeningen en het percentage overeenstemming, beide instrumenten te weinig betrouwbare gegevens opleveren. In de uitvoering van deze studie waren n=30 patiënten betrokken, die door twee verzekeringsartsen afzonderlijk en op twee opeenvolgende dagen werden beoordeeld. Na twee weken werd dat herhaald. In totaal deden n=4 verzekeringsartsen mee.

Een ander voorbeeld is een onderzoek van Spanjer en e.a. (2001). Zij onderzochten de betrouwbaarheid van het 'Functie Informatie Systeem' (FIS) en de 'Psychische Mogelijkheden Lijst' (PML). Zij concludeerden dat zowel de *intra*- als de *inter*beoordelaarsbetrouwbaarheid redelijk tot goed was, met uitzondering van het specifieke item 'urenbeperking'. In deze studie werden 14 beoordelingsgesprekken gefilmd van vijf verzekeringsartsen. Elke film werd bekeken en beoordeeld door n=22 verzekeringsartsen.

In de literatuur[9] wordt een onderscheid gemaakt tussen twee stromingen bij de uitvoering van betrouwbaarheidsonderzoeken: (1) de ideografische, waarbij een of twee cases worden beoordeeld door veel beoordelaars, en (2) de nomothetische waarbij veel cases worden beoordeeld door twee beoordelaars. Een scan van de literatuur laat echter zien dat men zich niet tot deze stromingen beperkt, er wordt een variatie van methoden gevonden (aantallen en interval duur). Ter illustratie geven we voor 8 publicaties in tabel 1.1 een overzicht per studie van het aantal cases, het aantal 'beoordelaars', het materiaal dat de beoordelaars beoordelen, het tijdsinterval in geval van *intra*beoordelaarsvariatie.

[6] Brouwer S, Dijkstra PU, Gerrits EHJ, Schellekens JMH, Groothoff JW, Geertzen JHB, Göeken LNH, Intra- en inter-beoordelaarsbetrouwbaarheid 'FIS-Belastbaarheidspatroon' en 'Functionele mogelijkhedenlijst'. TBV 12, 2003: 360-367

[7] Spanjer J. De inter- en intrabeoordelaars betrouwbaarheid van WAP beoordelingen. TBV 9, 2001: 234-241

[8] Boer WEL de, Wind H, Dijk FJH van, Willems HBM. Interviews for the assessment of long-term incapacity for work: a study on adherence to protocols and principles. BMC Public Health 2009, 9:169

[9] Cicchetti D, Fontana A, Showalter D. Evaluating the reliability of multiple assessments of PTSD symptomatology: multiple examiners, one patient. Psychiatry Res 2009 Apr 30;166(2-3):269-80

Tabel 1.1: Publicaties met variatie in opzet betrouwbaarheidsstudie

auteurs	*N cases*	*N raters*	*'materiaal'*	*interval*	*instrumenten*
Brouwer et al (2003)[a]	30	4 (tweetallen, ook intra)	eigen gesprek	14 dagen	FIS[i], FML[j]
Cicchetti et al (2009)[b]	2	12	video	nvt	CAPS-1[k]
Hendriks et al (1997)[c]	33 81	5 1 voor intra 2*2 voor inter	eigen onderzoek	2 dagen	Impairments disability
Rootmensen et al (2007)[d]	40	3 2 voor intra	video	6 maanden	Inhalatietechniek, bij COPD
Spanjer (2001)[e]	14	22 (22 voor intra)	video	6 maanden	DASI[l]
Spanjer et al (2008)[f]	12	12	schriftelijk rapport	nvt	FIS[i], MAL[m]
Suijlenkom, Van et al (2000)[g]	24	2	eigen gesprek	nvt	Lichamelijk onderzoek bij hoofdpijn
Toro et al (2007)[h]	13	23 (17 voor intra)	video	14 dagen	SF-GT[n]

[a] Brouwer S, Dijkstra PU, Gerrits EHJ, Schellekens JMH, Groothoff JW, Geertzen JHB, Göeken LNH, Intra- en inter-beoordelaarsbetrouwbaarheid 'FIS-Belastbaarheidspatroon' en 'Functionele mogelijkhedenlijst'. TBV 12, 2003: 360-367

[b] Cicchetti D, Fontana A, Showalter D. Evaluating the reliability of multiple assessments of PTSD symptomatology: multiple examiners, one patient. Psychiatry Res 2009 Apr 30;166(2-3):269-80

[c] Hendriks EJM, Brandsma JW, Heerkens YF, Oostendorp RAB, Nelsoen RM, Intraobeserver and interobserver reliability of assessments of impairments and disabilities. Physical Therapy 77, 10:1097-1106

[d] Rootmensen GN, Keimpema ARJ, Looysen van, EE, Schaaf van der L, Jansen HM, Haan RJ, De, Reliability in the assessment of videotaped inhalation technique. J Aerosol Med 2007;20(4):429-33

[e] Spanjer J. De inter- en intrabeoordelaars betrouwbaarheid van WAP beoordelingen. TBV 9, 2001: 234-241

[f] Spanjer J, Krol B, Brouwer S, Groothoff JW. Inter-rater reliability in disability assessment based on a semi-structured interview report. Disabil Rehabil 2008;30(24):1885-90

[g] Suijlekom HA, van, Vet HC, De, Berg SG, Van den, Weber WE. Interobserver reliability in physical examination of the cervical spine in patients with headache. Headache 2000 Jul;40(7):581-6

[h] Toro B, Nester CJ, Farren PC. Inter- and intraobserver repeatability of the Salford Gait Tool: an observation-based clinical gait assessment tool. Arch Phys Med Rehabil 2007 Mar;88(3):328-32

[i] FIS staat voor 'Functioneel Informatie Systeem'.

[j] FML staat voor 'Functionele Mogelijkheden Lijst'.

[k] CAPS-1 staat voor 'Clinician Administered PTSD Scale-1'.

[l] DASI staat voor 'Disability Assessment Structured Interview'.

[m] MAL staat voor 'Mental Ability List'.

[n] SF-GT staat voor 'Salford Gait Tool'.

1.2 Probleemstelling en onderzoeksvragen

In deze studie wordt de betrouwbaarheid onderzocht van de nieuwe schattingsmethodiek. Een korte beschrijving van deze methodiek volgt in hoofdstuk 2.

De algemene probleemstelling van het onderzoek luidt: “Zijn de PIM-richtlijnen in het PTSS-protocol betrouwbaar bij de beoordeling van de mate van invaliditeit bij militairen met psychische aandoeningen?”

De betrouwbaarheid is de mate waarin een meetinstrument een uitkomst consistent meet. Deze consistentie kan in tijd -binnen een beoordelaar- en tussen beoordelaars bestudeerd worden. Wordt voor het bepalen van de betrouwbaarheid van de schattingsmethodiek de overeenkomst in bevindingen onderzocht van een beoordelaar, die herhaald in de tijd, met behulp van de schattingsmethodiek bij dezelfde persoon de invaliditeit bepaalt, dan spreekt men van *intra*beoordelaarsvariatie. Wordt echter de overeenstemming tussen twee of meer beoordelaars gehanteerd dan spreekt men van *inter*beoordelaarsvariatie.[10].

Meer specifiek worden in het huidige onderzoek de volgende onderzoeksvragen beantwoord:

1. Wat is de *intra*doktervariatie en de *inter*doktervariatie op verschillende niveaus van de nieuwe schattingsmethodiek: subrubriek, rubriek en invaliditeitspercentage?
2. Is de nieuwe schattingsmethodiek ter beoordeling van de mate van invaliditeit bij psychische aandoeningen betrouwbaar en zo ja in welke mate?

1.3 Organisatie en betrokken partijen

Bij het huidige onderzoek zijn verschillende partijen betrokken, zoals de opdrachtgever het Ministerie van Defensie, het onderzoeksteam TGO UMCG, een wetenschappelijke adviesgroep, een expertpanel, een klankbordgroep, het zorgloket MOD, (o.a. de VA-U) en de onafhankelijke verzekeringsartsen (VA-G). Bijlage 2 geeft een overzicht van alle betrokkenen.

[10] Epidemiologisch onderzoek, opzet en interpretatie. LM Bouter, MCJM van Dongen, Houten 1995

Hoofdstuk 2 De schattingsmethodiek

2.1 Beschrijving van de schattingsmethodiek

De schattingsmethodiek binnen het PTSS protocol[11] is ontwikkeld voor het bepalen van de beperkingen bij psychische aandoeningen in het kader van de invaliditeitsbeoordelingen. Het invullen van de scoringslijst is telkens het resultaat van een daaraan voorafgaande, in een rapportage vastgelegde verzekeringsgeneeskundige beoordeling.
De beperkingen worden onderverdeeld in 4 rubrieken met 10 subrubrieken, zoals weergegeven in tabel 2.1.

Tabel 2.1: De vier rubrieken met de tien subrubrieken van de schattingsmethodiek

Rubriek	Subrubriek
1. Activiteiten van het dagelijkse leven	11. Persoonlijke hygiëne en zelfzorg 12. Mobiliteit 13. Slapen 14. Seksuele functie
2. Sociaal functioneren	15. Basale communicatie 16. Communicatief sociaal functioneren 17. Communicatieve (emotionele) vaardigheid
3. Concentratie, doorzettingsvermogen en tempo	18. Structuur aanbrengen 19. Huishoudelijke activiteiten
4. Adaptatie aan stressvolle omstandigheden	20. Omgaan met stressvolle gebeurtenissen

De ernst van de beperkingen per subrubriek wordt geclassificeerd. De ernst van de beperkingen wordt onderverdeeld in: klasse 0 normaal; klasse 1 geringe beperkingen; klasse 2 milde beperkingen; klasse 3 matige beperkingen; klasse 4 ernstige beperkingen; klasse 5 extreme beperkingen. Voor iedere subrubriek wordt tevens iedere klasse uitgebreid omschreven.
Tot slot dienen de aldus gerubriceerde en geclassificeerde beperkingen te worden geconverteerd naar een invaliditeitspercentage.

2.2 Van beperkingen naar invaliditeitspercentage

Per subrubriek wordt de ernst van de beperking geduid en gescoord. De vier rubrieken, onafhankelijk van het aantal subrubrieken, tellen even zwaar mee. De scores van alle subrubrieken, die tot dezelfde rubriek behoren, worden bij elkaar opgeteld en vervolgens gedeeld door het aantal subrubrieken. Daarna worden de subscores van de 4 rubrieken bij elkaar opgeteld en gedeeld door vier. Dit quotiënt, zijnde een getal tussen de 0 en 5, geeft de gemiddelde rubriekscore aan. De waarden 0-5 vertegenwoordigen een percentage tussen de 0 en 100 procent. De vertaalslag van rubriekscore naar invaliditeitspercentage verloopt

[11] PTSS protocol – PTSS WIA IP Protocol, Project WIA ZORG wet, Defensie 2007

vervolgens lineair en wordt weergegeven in stappen van 5%. Zo komt een gemiddelde rubriekscore van 2,4 uit op een invaliditeitspercentage van 48%, zijnde een schatting van 45-50%. Het hiermee overeenkomende afgerond invaliditeitspercentage (IP) bedraagt 45%.
Als op iedere subrubriek klasse 1 wordt gescoord, bedraagt het IP 20%, als op iedere subrubriek klasse 5 wordt gescoord, bedraagt het IP 100%.

Hoofdstuk 3 Methoden

In dit hoofdstuk worden de methoden van het onderzoek beschreven. De opzet van het onderzoek wordt weergeven, de manier waarop de beoordelingsgesprekken zijn gefilmd wordt toegelicht en vervolgens worden de verschillende actoren van het onderzoek beschreven. Tot slot wordt een beschrijving gegeven van de analyses die zijn uitgevoerd.

3.1 Opzet

Om de onderzoeksvragen te beantwoorden is gebruik gemaakt van gefilmde beoordelingsgesprekken. De filmbeelden bestaan uit het beoordelingsgesprek met de post-actieve militairen met psychische klachten en de VA-U. Deze films worden door meerdere VA-G bekeken. De VA-G scoren de beperkingen van de betreffende post-actieve militair met behulp van de schattingsmethodiek.
De gescoorde beperkingen op rubriek- en subrubriek-niveau worden vergeleken tussen de VA-G. Zoals eerder aan de orde gekomen is, wordt er geen vergelijking gemaakt tussen de beoordeling van de VA-U en de beoordeling van de VA-G. Op basis van advies van een methodoloog en vanwege praktische haalbaarheid is gekozen om 24 filmopnames van beoordelingsgesprekken te verzamelen. Op basis van de eerste 8 films is een power berekening uitgevoerd[12]. Met 24 films is de power goed te noemen (>0.8). Elke film wordt 2 keer bekeken door vijf verschillende VA-G.

3.2 Het gefilmde beoordelingsgesprek

In deze paragraaf wordt ingegaan op het onderzoeksmateriaal: het gefilmde beoordelingsgesprek. Het gefilmde beoordelingsgesprek bestaat uit het beoordelingsgesprek van de post-actieve militairen met psychische klachten, die een aanvraag voor militair pensioen hebben ingediend, met de VA-U. Allereerst worden de in- en exclusie criteria beschreven: welke post-actieve militairen worden gefilmd? Vervolgens wordt ingegaan op de VA-U. Tenslotte volgt een korte uitleg over de techniek die is toegepast bij het filmen van de gesprekken.

3.2.1 De in- en exclusie criteria

Het onderzoeksmateriaal bestaat uit het gefilmde beoordelingsgesprek van een post-actieve militair met psychische klachten met een VA-U. De post-actieve militairen die gefilmd kunnen worden dienen aan een aantal voorwaarden te voldoen. De in- en exclusie criteria zoals die in het onderzoek zijn vastgesteld staan vermeld in tabel 3.1 en worden hieronder besproken.
De opdracht beperkt zich in het huidige onderzoek tot post-actieve militairen waarbij sprake is van een eerste beoordeling. Dit betekent dat herbeoordelingen worden geëxcludeerd. Dit omdat er bij herbeoordelingen al informatie in eerdere gesprekken aan de orde is geweest en daardoor de informatie voor het gefilmde gesprek waarschijnlijk onvolledig zou zijn.
Geïncludeerd worden post-actieve militairen in de leeftijd van 18 tot en met 64 jaar. Deze grens is bij 64 jaar gesteld, omdat bij ouderen (≥ 65) weinig eerste beoordelingen voorkomen;

[12] Flack VF, Afifi AA, Lachenbruch PA, Schouten HJA, Sample size determinations for the two rater kappa statistic. Psychometrika vol. 53, no. 3, 321-325, 1988

daarnaast kan volgens de huidige Algemene militaire pensioenwet militair invaliditeitspensioen tot 65 jaar aangevraagd worden.
De beoordelingen vinden plaats bij het Zorgloket MOD in Utrecht. Om pragmatische en opnametechnische redenen is gekozen om alleen de gesprekken die plaatsvinden op het Zorgloket MOD in Utrecht voor dit onderzoek te filmen. Post-actieve militairen die om welke reden dan ook (bv lichamelijke redenen, detentie) niet in staat zijn om naar het Zorgloket MOD in Utrecht te komen, worden geëxcludeerd voor dit onderzoek.

3.2.2 Het inclusieproces
De zorgcoördinatoren van het Zorgloket MOD includeren de post-actieve militairen voor dit onderzoek. Zij informeren alle post-actieve militairen die voldoen aan de vastgestelde criteria (zie tabel 3.1). Voor het informeren hebben de zorgcoördinatoren instructies gekregen van het onderzoeksteam en er is een informatiefolder beschikbaar voor de post-actieve militairen. De informatie wordt in het eerste gesprek door de zorgcoördinator gegeven (zie bijlage 3). Dit gesprek vindt plaats naar aanleiding van indiening van een rekest voor MIP.
Naast de informatie geeft de zorgcoördinator ook een toestemmingsformulier en een niet-deelname formulier. De post-actieve militair wordt gevraagd een van de twee formulieren terug te sturen naar het onderzoeksteam van het UMC Groningen. In het toestemmings-formulier geeft de post-actieve militair toestemming voor het laten filmen van het beoordelingsgesprek; er wordt vermeld dat de film opnames uitsluitend gebruikt zullen worden voor dit onderzoek, dat de beelden na afloop worden vernietigd, en dat de beelden niet opeisbaar zijn (zie bijlage 4). Het niet-deelname formulier wordt teruggestuurd indien de post-actieve militair geen toestemming geeft voor de filmopnames. Hierbij wordt ook gevraagd naar de reden voor niet-deelname (zie bijlage 5). Deze laatste informatie kan benut worden in de analyses wie wel en niet deelneemt aan het onderzoek.
Wanneer het onderzoeksteam het ondertekende deelname formulier ontvangen heeft van de post-actieve militair die mee wil werken aan het onderzoek, geeft het onderzoeksteam dit door aan de VA-U. De VA-U beoordeelt vervolgens met behulp van het dossier of de post-actieve militair inderdaad aan de in- en exclusiecriteria voldoet. Indien de VA-U denkt dat een post-actieve militair niet voldoet aan de criteria of niet in staat is om mee te werken aan het onderzoek wordt dit in overleg met de manager teruggekoppeld naar de onderzoekers. De post-actieve militair kan dan alsnog geëxcludeerd worden. De zorgcoördinatoren worden hiervan ook op de hoogte gehouden. Het beoordelingsgesprek wordt vervolgens ingepland en het filmproces gestart (paragraaf 3.2.3)
Wanneer het beoordelingsgesprek gefilmd is, vindt er nog één inclusiemoment plaats. De films worden namelijk bekeken en beoordeeld door een expertpanel[13]. Het expertpanel bekijkt of er minimaal zes van de tien subrubrieken in het gesprek aan bod komen en daarmee te scoren zijn door de VA-G. Het al dan niet aanwezig zijn van beperkingen is daarbij niet van belang. Bovendien wordt door het expertpanel bekeken of het minimum van zes subrubrieken binnen 2,5 uur aan de orde is gekomen. Tot slot geeft het expertpanel aan of er gedeeltes uit de film geknipt moeten worden, bijvoorbeeld gedeeltes waarin een conclusie over de beperkingen wordt getrokken.

[13] Dit panel, bestaande uit twee deskundigen, is ingesteld door de opdrachtgever. Beide deskundigen zijn nauw betrokken geweest bij de opzet van het schattingsinstrument. Ook hebben zij een belangrijke rol vervuld bij de scholing in het schattingsinstrument, zowel bij de VA-U als bij de VA-G.

Tabel 3.1 In- en exclusiecriteria per functie

Betreffend	*Door wie*	*Inclusie*	*Exclusie*
Post-actieve militair	Zorgcoördinator	• Post-actieve militair heeft rekest ingediend voor MIP vanwege psychische klachten *(de hoofdklacht is psychisch, er mogen wel bijkomende fysieke klachten zijn)* • Leeftijd 18 - 64 jaar	• Herbeoordeling • Beoordelingsgesprek vindt ergens anders plaats dan Zorgloket Utrecht
Post-actieve militair	Verzekeringsarts Zorgloket MOD		• Herbeoordeling • Beoordelingsgesprek vindt ergens anders plaats dan Zorgloket Utrecht • Op basis van dossier studie voorafgaand aan gesprek met post-actieve militair. Beleidsmatig overleg met manager. Aangeven: niet in staat mee te doen aan het onderzoek, vanwege…..
Gefilmd beoordelings-gesprek	Expertpanel	• Volledig beoordelingsgesprek gefilmd	• Minder dan 6 van de 10 subrubrieken te beoordelen

3.2.3 De verzekeringsartsen van het Zorgloket MOD (VA-U)

De VA-U die het beoordelingsgesprek met de post-actieve militairen voeren zijn werkzaam bij het Zorgloket Militaire Oorlogs- en Dienstslachtoffers (MOD) in Utrecht. Dit Zorgloket is ondergebracht bij het ABP. Vier[14] van de 8 verzekeringsartsen hebben toestemming gegeven voor het laten filmen van het beoordelingsgesprek. Het beoordelingsgesprek dient zo normaal mogelijk te verlopen en er mag geen extra persoon (filmer) in de kamer aanwezig zijn. Ook dient het gesprek in de eigen kamer van de arts plaats te vinden. Hieraan is voldaan door gebruik te maken van filmopnames met behulp van video-conferencing (VC).
De VA-U voert het beoordelingsgesprek zoveel mogelijk op reguliere wijze.
De VA-U heeft de mogelijkheid om na afloop, schriftelijk, eventuele bijzonderheden van het gesprek aan te geven. Een formulier met achtergrondinformatie van de post-actieve militair wordt na afloop van het beoordelingsgesprek door de VA-U ingevuld en naar het onderzoeksteam gestuurd.

[14] In de loop van het onderzoek is er een 5e arts bijgekomen die 1 filmopname voor dit onderzoek gedaan heeft.

3.2.4 De techniek waarmee het gesprek wordt gefilmd
De film wordt opgenomen door middel van video-conferencing. Dit betekent dat de filmopnames door middel van beveiligde internetlijnen op afstand, vanuit Groningen, gemaakt worden. Op de film is alleen de post-actieve militair in beeld te zien, eventueel een partner of andere persoon die mee is. De VA-U is alleen te horen, maar zelf niet in beeld.

3.3 De onafhankelijke verzekeringsartsen (VA-G)
De artsen die de beperkingen van de post-actieve militair op film schatten zijn onafhankelijke verzekeringsartsen (VA-G). Het gaat daarbij om geregistreerde verzekeringsartsen die niet werkzaam zijn bij het Zorgloket MOD. De verzekeringsartsen zijn geselecteerd en geworven bij een sociale zekerheidsinstelling.
Alle verzekeringsartsen zijn geïnformeerd over het onderzoek door middel van een informatiebrief. Geïnteresseerde artsen, die voldoen aan de inclusie criteria, hebben zich aangemeld voor een bijeenkomst. Hierin is het onderzoek nader gepresenteerd en konden de VA-G zich aanmelden voor deelname. De inclusie criteria voor de VA-G zijn:

(a) geregistreerd verzekeringsarts;
(b) onafhankelijk in termen van niet eerdere betrokkenheid bij medische beoordelingen in het kader van de militaire regelgeving; deze eis geldt ook gedurende de onderzoeksperiode indien het beoordelingen buiten het onderzoek betreft;
(c) gebonden aan geheimhouding gedurende en na het onderzoek
    a. medisch beroepsgeheim;
    b. onderzoekseis: geen onderlinge communicatie met collega's over het beoordelingsgesprek.

Er zijn minimaal tien artsen nodig. Uiteindelijk hebben 14 artsen zich aangemeld. De VA-G zijn getraind in het toepassen van de nieuwe schattingsmethodiek. Deze training is verzorgd door het expertpanel dat ook de VA-U heeft getraind. De training bestaat uit 1 dag uitleg en oefenen van en met de schattingsmethodiek. Een week voor de eerste dag dat de VA-G de films bekeken en scoorden, heeft nog een opfrisbijeenkomst van twee uur plaatsgevonden door het expertpanel.
De aandachtspunten, die naar voren zijn gekomen tijdens de trainingsdag in combinatie met aandachtspunten die uit de ervaringen uit de praktijk naar voren zijn gekomen, zijn door het expertpanel samengevat.

3.4 Het onderzoeksdesign en de uitvoering daarvan in de vorm van O&S dagen
Het scoren van de beperkingen van de post-actieve militair op film door VA-G gebeurt tijdens plenair georganiseerde onderzoeks- en scoringsdagen (O&S dagen).
In totaal zijn voor het onderzoek zes keer twee O&S dagen georganiseerd (zie ook het Onderzoeksdesign in bijlage 7: T1, T1', T2, T2', T3, T3').
Tijdens de eerste sessie van twee O&S dagen (T1) wordt de eerste serie van 8 films bekeken door de VA-G. Twee maanden daarna vindt T1' plaats (de tweede O&S dagen), waarbij dezelfde films worden gezien en gescoord om de *intra*doktervariatie te bepalen. De tweede serie van 8 films wordt vervolgens gescoord tijdens de tweede sessie van twee O&S dagen (T2), etc..
Vanuit de groep van 14 deelnemende VA-G worden er per O&S dag 10 beoordelaars ingepland: elke film wordt door vijf verschillende VA-G bekeken en gescoord. Per dagdeel krijgen twee

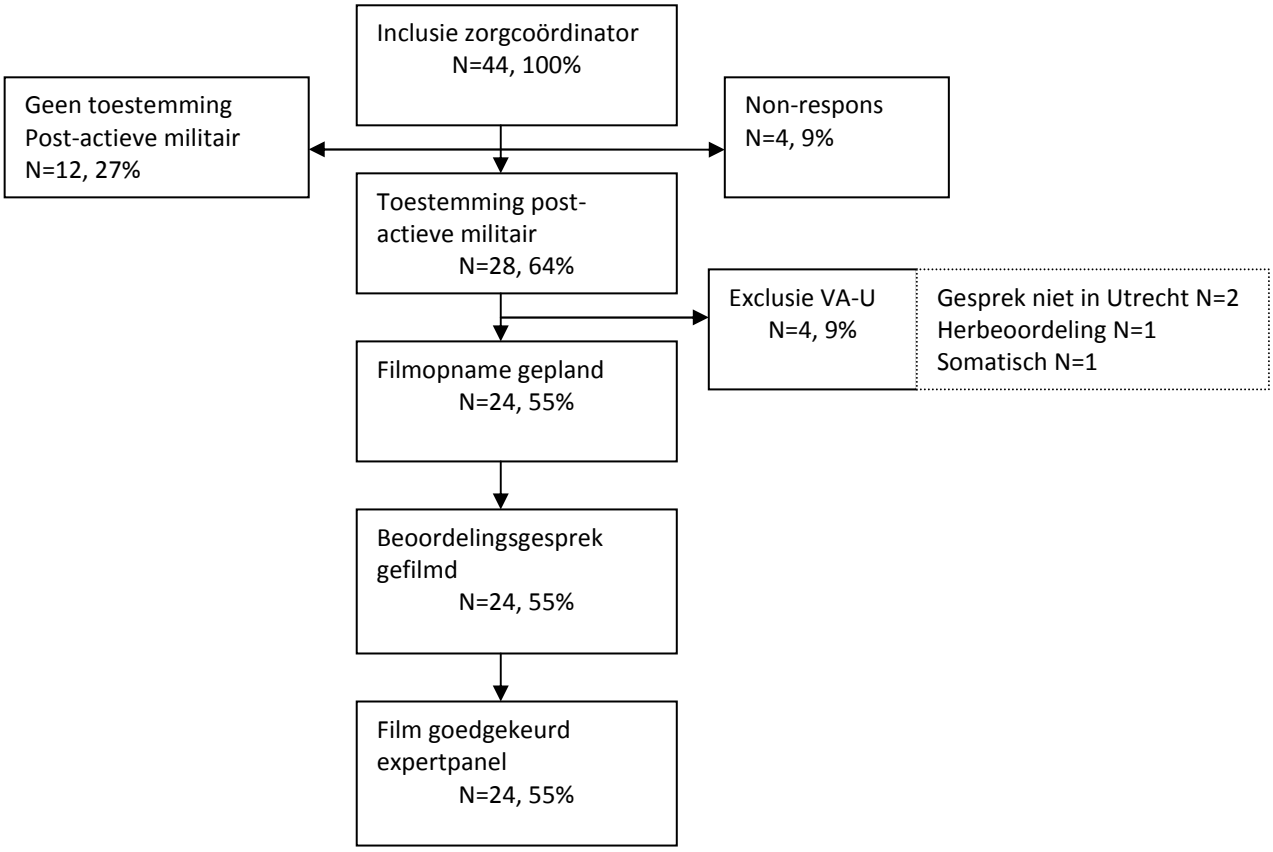
groepen van 5 VA-G een film te zien. Op die manier kunnen per dag vier films worden bekeken en gescoord. Om selectie te voorkomen wordt de groep van vijf VA-G per film verschillend samengesteld.
Elke O&S dag start met een plenair gedeelte met de VA-G en het onderzoeksteam. Tijdens het plenaire gedeelte wordt naast logistieke punten, het lijstje met aandachtspunten over de schattingsmethodiek, dat is samengesteld door het expertpanel, toegelicht.
Vervolgens bekijkt elke VA-G individueel per dagdeel 1 gesprek op film. Voor elke film is de achtergrondinformatie van de gefilmde post-actieve militair beschikbaar voor de VA-G, opgesteld door de VA-U die het gesprek voerde. Met behulp van de nieuwe schattingsmethodiek scoort de VA-G na het zien van de film de aard van de beperkingen. Na afloop van de film en het scoren van de beperkingen heeft de VA-G nog de gelegenheid om eventuele opmerkingen over het gesprek schriftelijk en anoniem door te geven.
De rubriecksscores en het invaliditeitspercentage (IP) wordt achteraf berekend door het onderzoeksteam.

3.4.1 Pilot
De eerste O&S dag is tevens gebruikt als pilot. Tijdens deze pilot staat de vraag centraal in hoeverre het mogelijk is voor de VA-G om twee films te bekijken en te scoren op een dag: een film per dagdeel. Tijdens de voorbereidingen van het onderzoek kwam naar voren dat twee films op een dag wellicht te veel is. Daarom heeft in de pilot een groep VA-G twee films per dag bekeken en gescoord en een groep VA-G een film bekeken en gescoord.
Het resultaat van de pilot was dat het goed haalbaar is om twee films op een dag te bekijken en te scoren. In het vervolg van het onderzoek wordt op die manier verder gewerkt.

3.4.2 Wat vullen de VA-G in?
De VA-G vullen een formulier in met hierop de tien subrubrieken, waarop ze 1 van de 5 klassen aankruisen. Het kan zijn dat een betreffende subrubriek volgens de VA-G niet te schatten is doordat deze, volgens de VA-G, onvoldoende aan de orde is gekomen in het beoordelingsgesprek. Er is op die manier onvoldoende informatie om de betreffende subrubriek te kunnen scoren. De subrubriek wordt dan leeg gelaten, na de eerste sessie O&S dagen is een kolom toegevoegd aan het formulier waarin “niet aan de orde” kan worden aangekruist.
Afgesproken is dat wanneer er getwijfeld wordt tussen twee klassen, de hogere klasse wordt gescoord. Op het formulier kan aangegeven worden wanneer van twijfel sprake is.

3.5 De respons

In figuur I is in een stroomschema weergegeven hoe de inclusie en respons van de post-actieve militairen is verlopen.



Figuur I Stroomschema inclusie en respons post-actieve militairen

Van de n=44 post-actieve militairen die zijn benaderd door de zorgcoördinatoren, hebben n=28 (64%) toestemming gegeven voor filmopname, n=12 (27%) hebben geen toestemming gegeven en n=4 (9%) hebben niet gereageerd (ook niet na reminders van de zorgcoördinator).
Van de n=28 (64%) toestemmingen zijn uiteindelijk nog n=4 (9%) geëxludeerd vanwege uiteenlopende redenen (bij 2 vond het gesprek niet plaats in Utrecht, bij 1 bleek sprake te zijn van een herbeoordeling en bij 1 bleek sprake te zijn van somatische problematiek). In totaal zijn n=24 (55%) gesprekken met post-actieve militairen gefilmd.

3.5.1 Non-respons analyse

Op verschillende manieren is een non-response analyse uitgevoerd (zie ook tabel 3.2) met als doel na te gaan of er sprake is van selectie: doen post-actieve militairen met ernstiger problematiek vaker wel of niet mee met het onderzoek?
Om te beginnen hebben alle (n=12) post-actieve militairen die geen toestemming hebben gegeven voor deelname een ‘niet-deelname’-formulier ingevuld en de reden daarvoor aangegeven.

Daarnaast zijn voor de non-response analyse drie vragen gesteld aan de zorgcoördinatoren die de post-actieve militairen geïncludeerd hebben.[15] Ten eerste of alle post-actieve militairen die voldeden aan de inclusiecriteria gevraagd zijn voor medewerking aan het onderzoek. Ten tweede is het uiteindelijke percentage IP gevraagd van zowel de deelnemende als de niet deelnemende post-actieve militairen. En ten slotte is gevraagd of de zorgcoördinator zelf van mening is dat er sprake kan zijn geweest van selectie bij het al dan niet meedoen aan het onderzoek.

Tabel 3.2: Overzicht non-response analyse

	Geen toestemming n=12	Wel toestemming N=28
Niet-deelname formulieren Reden niet deelname: - Ongemakkelijk gevoel bij camera - Emotioneel belastend	 n=10 n=2	
	Geen toestemming N=7	Gefilmde post-actieve militair N=24
Vragen aan zorgcoördinatoren		
Aanvraag militair invaliditeitspensioen afgewezen	3	4
Invaliditeitspercentage niet bekend	1	2
Gem. afgerond invaliditeitspercentage (sd)	11.7 (2.9)	12.2 (9.7)
Range afgerond invaliditeitspercentage	10-15	10-30

Alle post-actieve militairen die aangaven niet mee te willen doen aan het onderzoek hebben op een 'niet deelname'-formulier de reden daarvoor aangegeven. Bij de meerderheid, n=10, werd als reden genoemd samengevat als "Ongemakkelijk gevoel bij een camera". Als tweede reden werd door n=2 genoemd "Emotioneel belastend".
De zorgcoördinatoren gaven aan dat alle post-actieve militairen die aan de inclusiecriteria voldeden zijn gevraagd voor deelname. De zorgcoördinatoren schatten zelf in dat er geen sprake is geweest van selectie.
Verder is voor n=7 van de n=12 post-actieve militairen die geen toestemming hebben gegeven om deel te nemen informatie ontvangen voor de non-respons analyse en voor n=24 van de n=24 deelnemende post-actieve militairen.
Voor zowel post-actieve militairen die geen toestemming hebben gegeven als ook voor de deelnemers aan het onderzoek geldt dat voor enkelen de aanvraag voor een IP is afgewezen. Zowel het gemiddelde afgeronde invaliditeitspercentage en de range afgerond invaliditeitspercentage voor de deelnemende post-actieve militairen is iets hoger dan voor de 'niet-deelnemers'. De resultaten van de non-response analyse geven geen indicatie voor selectie in termen dat post-actieve militairen met ernstige problematiek vaker niet deelnemen aan het onderzoek.

[15] De zorgcoördinatoren zijn op de hoogte van deze informatie en hebben dit anoniem aan de onderzoekers doorgegeven.

3.6 Analyses

3.6.1 Beschrijvende statistieken en Ag09

Kenmerken van de post-actieve militairen, de VA-U, VA-G en de films worden beschreven in het resultaten hoofdstuk 4.

De verdeling van de geschatte klassen van de 10 subrubrieken worden met beschrijvende statistiek weergegeven. Ook worden de resultaten weergegeven wanneer een verschil van 1 klasse niet als praktijk relevant verschil wordt beschouwd.

De *intra*doktervariatie en *inter*doktervariatie worden weergegeven door middel van percentage overeenstemming. Doordat de verdeling van de scores op de subrubrieken erg scheef is, is overeenstemming berekening op basis van Kappa niet mogelijk[16].

Voor de berekening van het percentage overeenstemming wordt gebruik gemaakt van een programma op basis van Ag09. Ag09 is een statistisch programma dat wordt beschreven door Popping (2010)[17]. De volgende berekeningen worden daarmee uitgevoerd: de overeenstemming tussen de 1^{e} en de 2^{e} meting binnen een VA-G, de gepaarde overeenstemming tussen alle VA-G over alle films (apart voor de 1^{e} en de 2^{e} meting), en de overeenstemming per klasse wordt bepaald. De uitkomsten worden weergegeven in kwadratisch gewogen percentage overeenstemming.

In de literatuur wordt een overeenstemmingspercentage van 70% of hoger als goed beschouwd, en een overeenstemmingspercentage van hoger dan 90% als excellent.[18] [19]Het statistische programma Ag09 berekent een gemiddeld 'kwadratisch gewogen percentage overeenstemming' tussen alle beoordelingsparen voor alle films (zie ook bijlage 8). Met gewogen wordt bedoeld dat een verschil van 1 klasse minder zwaar weegt dan een verschil van 2 of meer klassen. De weging kan op twee manieren plaats vinden. Ten eerste, met een vaststaand theoretisch aantal klassen, in deze studie 5 klassen. Het theoretisch aantal klassen loopt van 0 tot en met 4 (is 5 klassen), omdat de hoogste klasse (klasse 5) in deze studie nooit gescoord is. Ten tweede met het daadwerkelijke aantal klassen, welke verschillend is per subrubriek. Bijvoorbeeld wanneer alleen de klassen 0 en 1 gescoord zijn, dan loopt het daadwerkelijk aantal klassen van 0 tot en met 1. In bijlage 8 is de analyse techniek nader toegelicht.

Als laatste analyse wordt de variatie op het niveau van de uitkomstmaat afgerond invaliditeitspercentage weergegeven en de totstandkoming daarvan nader geanalyseerd.

[16] High agreement but low kappa: II. Resolving the paradoxes. Cicchetti DV, Feinstein AR. *J Clin Epidemiol*. 1990;43(6):551-8

[17] Ag09 : A Computer Program for Interrater Agreement for Judgments. R. Popping. *Social Science Computer Review* 2010 28: 391

[18] Spanjer J, Krol B, Popping R, Groothoff JW, Brouwer S. Disability assessment interview: the role of detailed information on functioning in addition to medical history-taking. Journal of Rehabilitation Medicine 2009;41:267-72

[19] Landis JR, Koch GG. The measurement of Observer Agreement for Categorical Data. Biometics, Vol. 33, No1. 1977, pp 159-174

Hoofdstuk 4 Resultaten

In dit hoofdstuk worden de resultaten van het onderzoek gepresenteerd. Om te beginnen worden de onderzoeksdata beschreven. Vervolgens wordt verder ingegaan op de *intra*doktervariatie, daarna op de *inter*doktervariatie. Tot slot wordt nader ingegaan op het invaliditeitspercentage en de totstandkoming daarvan.

4.1 Beschrijving van de onderzoeksdata

4.1.1 Kenmerken van de onderzoeksgroepen

Tabel 4.1 geeft een overzicht van enkele kenmerken van de deelnemende post-actieve militairen, de verzekeringsartsen in Groningen (VA-G), de verzekeringsartsen in Utrecht (VA-U) en de films.

Tabel 4.1: Kenmerken post-actieve militairen, VA-G, VA-U en films

Post-actieve militairen	
Geslacht, N	
man	23
vrouw	1
Leeftijd, gem (sd)	43,3 (6,3) 1 missing
Missie, N (%)*	
Libanon	9 (37,5)
Joegoslavië	15 (62,5)
Irak	1 (4,2)
VA-G	
Geslacht, N (%)	
man	9 (64.3)
vrouw	5 (35.7)
Leeftijd, gem (sd)	50.5 (6.2)
Werkervaring, jaren (sd)	16.3 (7.3)
Range	4-31
Gemiddeld aantal films per VA-G, (sd)	8.6 (2.8)
Range	4-13[20]
VA-U	
Geslacht, N	
man	
vrouw**	
Gemiddeld aantal films per VA-U, N (sd)	5.8 (1.0)
Films	
Gemiddelde duur films***, (sd)	1:52 uur (26 min)
Range	1:08 – 2:38

* totaal meer dan 100% doordat 1 post-actieve militair verschillende uitzendingen had

** 1 gesprek door vervanger (vrouw) doordat een VA-U uitviel ivm ziekte

***duur film, gezien door VA-G

[20] Een VA-G heeft ingevallen voor een andere VA-G (wegens ziekte), vandaar dat deze VA-G 13 films gezien heeft.

4.1.2 "Niet aan de orde gekomen" subrubrieken

In tabel 4.2 is weergegeven hoe vaak het voorkomt dat een VA-G bij een bepaalde film het vakje "niet aan de orde gekomen" heeft aangevinkt. Uit deze tabel blijkt, dat dit zelden gebeurt, met als uitzondering de subrubriek 4 'Seksuele functie'. Daar wordt in de 1e meting 26 keer "niet aan de orde gekomen" gescoord en in de 2e meting, 25 keer. Het betreft in 23 van de 25 keren dezelfde VA-G en dezelfde films.

Tabel 4.2: Overzicht voorkomen "niet aan de orde gekomen" per subrubriek

Subrubriek		*NG, niet aan de orde gekomen*		*Overeenstemming (zelfde VA-G en zelfde film)*
		1e meting	2e meting	
1	Hygiëne zelfzorg	2	4	1
2	Mobiliteit	1	0	-
3	Slapen	1	1	0
4	Seksuele functie	26	25	23
5	Basale communicatie	1	2	0
6	Sociale communicatie	0	0	-
7	Communicatieve vaardigheden	0	0	-
8	Structuur aanbrengen	3	1	1
9	Huishoudelijke activiteiten	2	2	2
10	Omgaan stressvolle gebeurtenissen	4	0	-

Theoretisch kan maximaal n=120 voorkomen per subrubriek: N=5 VA-G x 24 films=120

De volgende tabel 4.3 gaat nader in op de scores "niet aan de orde gekomen" bij de subrubriek 4 'Seksuele functie'. Hierin wordt het aantal VA-G weergegeven die "niet aan de orde gekomen" gescoord hebben per film. Er zijn bijvoorbeeld bij de 1e meting 5 films waarbij 1 VA-G "niet aan de orde gekomen" gescoord heeft. De overige 4 VA-G hebben bij deze zelfde 5 films wel een klasse gescoord. En er zijn bijvoorbeeld 3 films waarbij, zowel bij de 1e meting als bij de 2e meting, alle 5 VA-G "niet aan de orde gekomen" gescoord hebben.

Tabel 4.3: Subrubriek 'Seksuele functie', nadere specificatie van "niet aan de orde gekomen"

4	*Seksuele functie*	*NG*	
	1e meting	26	5 films → 1 VA-G
			1 film → 2 VA-G
			1 film → 4 VA-G
			3 films → 5 VA-G
	2e meting	25	3 films → 1 VA-G
			1 film → 3 VA-G
			1 film → 4 VA-G
			3 films → 5 VA-G

NG="niet aan de orde gekomen"

Wat valt op:

- Er zijn bij beide metingen, de 1e en de 2e, 4 films waarbij 4 of 5 VA-G aangeven dat de subrubriek 4 'Seksuele functie' niet aan de orde is gekomen.

Vervolgens is nagegaan of er sprake is van een leereffect. Met andere woorden, of er een verschil is waar te nemen in hoe vaak "niet aan de orde gekomen" wordt gescoord door de VA-G tijdens de drie verschillende sessies in de tijd. Analyse van de gegevens laat zien dat in de 1e meting, in sessie 1 10 keer "niet aan de orde gekomen" wordt gescoord, 2 keer in sessie 2, en 14 keer in sessie 3. Bij de 2e meting was dit, 10 keer in sessie 1, 1 keer in sessie 2, en 14 keer in sessie 3. Hieruit blijkt geen leereffect.

4.1.3 Verdeling van klassen per subrubriek

In tabel 4.4 wordt de verdeling van het voorkomen van de verschillende gescoorde klassen per subrubriek weergegeven. Dit is weergegeven per subrubriek voor de 24 films die elk door 5 VA-G zijn gescoord.

Tabel 4.4: Verdeling geschatte klasse over de subrubrieken, 1e meting

Subrubriek		*Normaal*	*Geringe beperking*	*Milde beperking*	*Matige beperking*	*Ernstige beperking*	*Extreme beperking*	*NG*
		0	*1*	*2*	*3*	*4*	*5*	
1	Hygiëne zelfzorg	117 (97.5)	1 (0.8)					2 (1.7)
2	Mobiliteit	92 (76.7)	25 (20.8)	2 (1.7)				1 (0.8)
3	Slapen	17 (14.2)	30 (25)	55 (45.8)	12 (10)	5 (4.2)		
4	Seksuele functie	82 (68.3)	4 (3.3)	5 (4.2)	2 (1.7)	1 (0.8)		26 (21.7)
5	Basale communicatie	117 (97.5)	2 (1.7)					1 (0.8)
6	Sociale communicatie	43 (35.8)	50 (41.6)	16 (13.3)	5 (4.2)	6 (5)		
7	Communicatieve vaardigheden	43 (35.8)	42 (35)	28 (23.3)	5 (4.2)	2 (1.7)		
8	Structuur aanbrengen	80 (66.7)	29 (24.2)	8 (6.7)				3 (2.5)
9	Huishoudelijke activiteiten	93 (77.5)	17 (14.2)	8 (6.7)				2 (1.7)
10	Omgaan stressvolle gebeurtenissen	28 (23.3)	63 (52.5)	21 (17.5)	3 (2.5)	1 (0.8)		4 (3.3)

N=5 artsen x 24 films=120 per subrubriek
NG="niet aan de orde gekomen"

Wat valt op:

- Er is sprake van een scheve verdeling van de klassen: er wordt vaak klasse 0 gescoord, nooit klasse 5, en weinig klasse 3 of klasse 4.
- Bij 5 van de 10 subrubrieken wordt meer dan in tweederde van de situaties (> n=80 keer van het theoretische maximum n=120) een klasse 0 gescoord. Voor de subrubrieken: 1 'Hygiëne zelfzorg', 2 'Mobiliteit', 4 'Seksuele functie', 5 'Basale communicatie' en 9 'Huishoudelijke activiteiten', geldt dat voornamelijk klasse 0 gescoord wordt.
- Voor de subrubriek 3 'Slapen' wordt relatief het vaakst hogere klassen gescoord, gevolgd door de subrubriek 10 'Omgaan met stressvolle gebeurtenissen', 7 'Communicatieve vaardigheden', en 6 'Sociale communicatie'.

4.2 Intradoktervariatie

In deze paragraaf wordt de *intra*doktervariatie beschreven. Van iedere VA-G wordt voor elke film de 1e en de 2e meting vergeleken, en bepaald hoe groot de variatie is.
De onderstaande tabel 4.5 geeft de verdeling weer hoe vaak –binnen een VA-G- een overeenkomstige score is gegeven (theoretisch maximum n=120), hoe vaak het verschil +/-1 klasse was, +/- 2 klassen was, etc. Wanneer een VA-G zowel bij de 1e als bij de 2e meting "niet aan de orde gekomen" gescoord heeft, wordt dit gerekend als 0, een overeenkomstige score. Ook geeft de tabel weer, in de laatste kolom, in hoeverre er bij een gevonden verschil er sprake is van een hogere score tijdens de 1e meting in vergelijking met de 2e meting.

Tabel 4.5: Overeenkomsten en verschillen in gescoorde klassen binnen een VA-G tussen de 1e meting en de 2e meting, per subrubriek

Subrubriek N(%)		*0**	*+/- 1*	*+/- 2*	*+/- 3*	*+/- 4*	*NG$*	*1e >2e #*
1	Hygiëne zelfzorg	116 (96.7)					4 (3.3)	-
2	Mobiliteit	103 (85.8)	14 (11.7)	1 (0.8)	1 (0.8)		1 (0.8)	9 (56)
3	Slapen	85 (70.8)	32 (26.7)	1 (0.8)			2 (1.7)	21 (65)
4	Seksuele functie	105 (87.5)	5 (4.2)	4 (3.3)		1 (0.8)	5 (4.2)	8 (80)
5	Basale communicatie	113 (94.2)	4 (3.3)				3 (2.5)	1 (25)
6	Sociale communicatie	74 (61.7)	42 (35)	4 (3.3)				25 (54)
7	Communicatieve vaardigheden	69 (57.5)	44 (36.7)	7 (5.8)				30 (59)
8	Structuur aanbrengen	95 (79.2)	20 (16.7)	3 (2.5)			2 (1.7)	13 (56)
9	Huishoudelijke activiteiten	100 (83.3)	18(15)	2 (1.7)				7 (35)
10	Omgaan stressvolle gebeurtenissen	77 (64.2)	39 (32.5)				4 (3.3)	25 (64)

N=5 artsen x 24 films=120 per subrubriek

* 0=identieke score 1e meting en 2e meting. +/- 1 is een verschil van 1 klasse.

$ NG is "niet aan de orde gekomen" gescoord bij de 1e óf de 2e meting.

Wanneer er een verschil gevonden is tussen de 1e en de 2e meting, wordt hier het aantal keren weergegeven dat de score bij de 1e meting hoger is dan bij de 2e meting.

Wat valt op:

- Subrubriek 1 'Hygiëne zelfzorg' heeft de hoogste *intra*dokterovereenstemming: in 96.7% wordt bij de 2e meting een identieke klasse gescoord ten opzichte van de 1e meting. In 3.3% wordt 'niet aan de orde gekomen' gescoord bij de 1e of de 2e meting.
- Subrubriek, 6 'Sociale communicatie', 7 'Communicatieve vaardigheden' en 10 'Omgaan met stressvolle gebeurtenissen' laten de meeste variatie zien binnen een VA-G tussen de 1e en de 2e meting (voor alle drie subrubrieken geldt een *intra*dokterovereenstemming van <65%).
- Een verschil van 2 klassen of meer per subrubriek komt weinig voor (<6%).
- Over het algemeen is de *intra*doktervariatie goed: er is een hoge mate van overeenstemming binnen een VA-G in het scoren van de subrubrieken tijdens de 1e en de 2e meting.

- Wanneer er een verschil wordt gevonden tussen de 1e en de 2e meting dan kan over het algemeen gesteld worden dat de score bij de 1e meting hoger is dan de 2e meting. Dat geldt voor 7 van de 9 subrubrieken (bij subrubriek 1 is er geen sprake van een klasse verschil tussen de 1e en de 2e meting, doordat alleen verschillen gevonden zijn in "niet aan de orde gekomen").

In tabel 4.6 worden vervolgens dezelfde gegevens gebruikt als in tabel 4.5, alleen wordt nu een verschil van 1 klasse beschouwd als 'niet praktijkrelevant verschil'[21]. Een verschil van 1 klasse binnen een VA-G tussen de 1e en de 2e meting is nu samengevoegd met een identieke score (zie kolom 0 of +/- 1 in tabel 4.6).

Tabel 4.6: Overeenkomsten en verschillen in gescoorde klassen binnen een VA-G tussen de 1e en de 2e meting met toestaan van 1 klasse verschil, per subrubriek N(%)

		0 of +/- 1	*+/- 2*	*+/- 3*	*+/-4*	*NG$*
1	Hygiëne zelfzorg	116 (96.7)				4 (3.3)
2	Mobiliteit	117 (97.5)	1 (0.8)	1 (0.8)		1 (0.8)
3	Slapen	117 (97.5)	1 (0.8)			2 (1.7)
4	Seksuele functie	110 (91.7)	4 (3.3)		1 (0.8)	5 (4.2)
5	Basale communicatie	117 (97.5)				3 (2.5)
6	Sociale communicatie	116 (96.7)	4 (3.3)			
7	Communicatieve vaardigheden	113 (94.2)	7 (5.8)			
8	Structuur aanbrengen	115 (95.8)	3 (2.5)			2 (1.7)
9	Huishoudelijke activiteiten	118 (98.3)	2 (1.7)			
10	Omgaan stressvolle gebeurtenissen	116 (96.7)				4 (3.3)

N=5 artsen x 24 films=120 per subrubriek

* 0=of +/- 1 identieke score tussen 1e meting en 2e meting of een verschil van 1 klasse.

$ NG is niet gescoord in de 1e óf de 2e meting.

Wat valt op:

- Wanneer 1 klasse verschil niet als praktijkrelevant verschil wordt gezien, hebben alle subrubrieken meer dan 90% overeenstemming tussen de 1e en de 2e meting.
- De meeste *intra*doktervariatie, overigens nog zeer gering, wordt gevonden voor subrubriek 4 'Seksuele functie' en 7 'Communicatieve vaardigheden'. Bij subrubriek 4 komt dit vooral door de 5% waarbij "niet aan de orde gekomen" gescoord is bij de 1e of de 2e meting.
- Conclusie: de *intra*doktervariatie is zeer laag, de *intra*dokterbetrouwbaarheid is zeer hoog.

[21] Beslissing van de klankbordgroep dd 18 mei 2011: analyses uitvoeren met de aanname dat een variatie van inschatting van 1 klasse in de praktijk niet als een praktijkrelevant verschil hoeft te worden aangemerkt.

De gevonden identiek gescoorde klassen in tabel 4.5 worden verder uitgewerkt in tabel 4.7, om inzicht te krijgen in een eventueel leereffect. Het voorkomen van identiek gescoorde klassen, een verschil van 0, tussen de 1e en de 2e meting wordt uitgesplitst voor de drie verschillende sessies van de dataverzameling (1e meting van sessie 1 in september 2010, sessie 2 in januari 2011 en sessie 3 in april 2011). De resultaten daarvan staan vermeld in onderstaande tabel 4.7.

Tabel 4.7: Voorkomen van identiek gescoorde klassen bij de 1e en de 2e meting, uitgesplitst per sessie

		Sessie			*Totaal*
Subrubriek N(%)		*1*	*2*	*3*	
1	Hygiëne zelfzorg	39 (33.6)	40 (34.5)	37 (31.9)	116 (100)
2	Mobiliteit	37 (35.9)	31 (30.1)	35 (34.0)	103 (100)
3	Slapen	28 (32.9)	27 (31.8)	30 (35.3)	85 (100)
4	Seksuele functie	34 (32.4)	35 (33.3)	36 (34.3)	105 (100)
5	Basale communicatie	37 (32.7)	38 (33.6)	38 (33.6)	113 (100)
6	Sociale communicatie	25 (33.9)	19 (25.7)	30 (40.5)	74 (100)
7	Communicatieve vaardigheden	25 (36.2)	21 (30.4)	23 (33.3)	69 (100)
8	Structuur aanbrengen	28 (29.5)	30 (31.6)	37 (39.0)	95 (100)
9	Huishoudelijke activiteiten	32 (32)	32 (32)	36 (36)	100 (100)
10	Omgaan stressvolle gebeurtenissen	25 (32.5)	26 (33.8)	26 (33.8)	77 (100)

Wat valt op:

- Het voorkomen van identiek gescoorde klassen bij de 1e en de 2e meting is voor de meeste subrubrieken per sessie vergelijkbaar. Er lijkt bij deze subrubrieken geen sprake te zijn van een leereffect.
- Bij subrubriek 6 'Sociale communicatie' en subrubriek 8 'Structuur aanbrengen' is een klein verschil te zien tussen de drie sessies. Bij subrubriek 8 'Structuur aanbrengen' wordt in de 3e sessie vaker bij de 2e meting een identieke klasse gescoord dan in de 1e en 2e sessie. Dit kan mogelijk duiden op een leereffect, maar is op het totaal zeer gering.

4.2.1 Intradoktervariatie en percentage overeenstemming

Door de scheve verdeling van klassen op de verschillende subrubrieken kan er geen betrouwbare kappa berekend worden. Daarom is in dit onderzoek het percentage overeenstemming berekend met Ag09 (zie paragraaf 3.6.1 en bijlage 8). De ranges, de VA-G met het laagste percentage overeenstemming en de VA-G met het hoogste percentage overeenstemming, (kwadratisch gewogen, zie bijlage 8) zijn weergegeven per subrubriek in tabel 4.8. Ook wordt weergegeven bij welke VA-G het laagste percentage overeenstemming hoort. Als voorbeeld: bij subrubriek 2 'Mobiliteit' hoort het laagste percentage overeenstemming, in dit geval 92%, bij VA-G 'M'.
In deze berekening wordt "niet aan de orde gekomen" meegenomen als missende waarden.

Tabel 4.8: Range percentage intradokterovereenstemming per VA-G, per subrubriek*

Subrubriek		*Percentage overeenstemming*	*VA-G met laagste percentage overeenstemming*
1	Hygiëne zelfzorg	100	-
2	Mobiliteit	92-100	M
3	Slapen	95-99	K
4	Seksuele functie	78-100	L
5	Basale communicatie	98-100	K
6	Sociale communicatie	93-99	K
7	Communicatieve vaardigheden	86-98	E
8	Structuur aanbrengen	95-100	E
9	Huishoudelijke activiteiten	94-99	L
10	Omgaan stressvolle gebeurtenissen	96-100	H en I

Alle films samen, per VA-G, het aantal beoordeelde films per VA-G varieert van n=4-13.
*kwadratisch gewogen percentage overeenstemming, met gebruikte weging: vaststaand aantal klassen (in deze studie n=5)

Wat valt op:

- Subrubriek 1 'Hygiëne zelfzorg': voor alle VA-G geldt 100% overeenstemming tussen de 1^{e} en de 2^{e} meting.
- Voor de subrubriek 4 'Seksuele functie' geldt de meeste variatie in gescoorde klassen tussen de 1^{e} en de 2^{e} meting: met een range van 78% tot 100%.
- Er is geen sprake van een specifieke VA-G die consequent de minste overeenstemming scoort, VA-G 'K' blijkt bij 3 subrubrieken de minste overeenstemming te scoren. Hierbij moet rekening worden gehouden dat VA-G 'K' met 12 films veel films heeft beoordeeld en de kans dus groter is dat hij/zij een keer lager uitvalt. Daarnaast hebben de laagst scorende VA-G nog altijd een hoog percentage *intra*dokterovereenstemming.

In de volgende tabel 4.9, worden dezelfde gegevens weergegeven als in tabel 4.8 alleen wordt nu "niet aan de orde gekomen" in de berekening meegenomen als de klasse 0.
De ranges, de VA-G met het laagste percentage overeenstemming en de VA-G met het hoogste percentage overeenstemming (kwadratisch gewogen, zie bijlage 8) zijn weergegeven per subrubriek. Dit resulteert niet in grote verschillen met tabel 4.8. Het belangrijkste verschil is dat het percentage overeenstemming van subrubriek 4 'Seksuele functie' stijgt.

Tabel 4.9: Range percentage intradokterovereenstemming per VA-G, per subrubriek, met "niet aan de orde gekomen" als klasse 0

Subrubriek			
		Percentage overeenstemming	*VA-G met laagste percentage overeenstemming*
1	Hygiëne zelfzorg	100	-
2	Mobiliteit	92-100	M
3	Slapen	95-99	K
4	Seksuele functie	84-100	L
5	Basale communicatie	98-100	K
6	Sociale communicatie	93-99	K
7	Communicatieve vaardigheden	86-98	E
8	Structuur aanbrengen	95-100	L
9	Huishoudelijke activiteiten	94-100	E
10	Omgaan stressvolle gebeurtenissen	96-100	H

Alle films samen, voor alle VA-G (totaal VA-G is n=14)

*kwadratisch gewogen percentage overeenstemming, met gebruikte weging: vaststaand aantal klassen (in deze studie n=5)

Conclusie intradokterbetrouwbaarheid

Op basis van de bovenstaande resultaten geldt als algemene conclusie dat de *intra*dokterbetrouwbaarheid goed tot zeer goed is.

4.3 Interdoktervariatie

In deze paragraaf richten we ons op de *inter*doktervariatie: in hoeverre verschillen of komen VA-G overeen in hun gescoorde klassen na het zien van dezelfde film? In de weergave van de volgende resultaten worden alleen de resultaten van de 1e meting weergegeven, de resultaten van de 2e meting zijn toegevoegd in de bijlage 9.

Tabel 4.10 geeft het aantal films weer met het aantal VA-G dat overeenkomstig scoort per subrubriek (bij de 1e meting). Hierbij is "niet aan de orde gekomen" meegenomen als een aparte klasse. Dus wanneer 5 VA-G "niet aan de orde gekomen" hebben gescoord, wordt dit gerekend als 100% overeenstemming, wanneer 1 VA-G "niet aan de orde gekomen" heeft gescoord en 4 VA-G een klasse 0, dan wordt dit gerekend als 4 VA-G overeenstemmend.

Tabel 4.10: Aantal films met aantal VA-G, N(%) met overeenkomstige gescoorde klasse per subrubriek (1e meting)

Subrubriek		*5* *100% Overeen-stemming*	*4* *4 VA-G Overeen-stemming*	*3* *3 VA-G Overeen-stemming*	*2* *2 VA-G Overeen-stemming*
1	Hygiëne zelfzorg	21 (87.5)	3 (12.5)	0	0
2	Mobiliteit	14 (58.3)	6 (25.0)	3 (12.5)	1 (4.2)
3	Slapen	2 (8.3)	12 (50.0)	7 (29.2)	3 (12.5)
4	Seksuele functie	10 (41.7)	9 (37.5)	4 (16.7)	1 (4.2)
5	Basale communicatie	21 (87.5)	3 (12.5)	0	0
6	Sociale communicatie	4 (16.7)	6 (25.0)	9 (37.5)	5 (20.8)
7	Communicatieve vaardigheden	2 (8.3)	4 (16.7)	10 (41.7)	8 (33.3)
8	Structuur aanbrengen	10 (41.7)	6 (25.0)	7 (29.2)	1 (4.2)
9	Huishoudelijke activiteiten	13 (54.2)	6 (25.0)	5 (20.8)	0
10	Omgaan stressvolle gebeurtenissen	7 (29.2)	6 (25.0)	10 (41.7)	1 (4.2)

N=24 films, 5 VA-G zelfde score is 100% overeenstemming

Wat valt op:

- Twee subrubrieken vallen op door de hoge overeenstemming, namelijk 1 'Hygiëne zelfzorg' en 5 'Basale communicatie', beide 87.5%. Voor beide subrubrieken geldt dat van de 24 films, er bij 21 films door alle 5 VA-G hetzelfde is gescoord. Bij de overige 3 films geldt bij deze subrubriek dat 4 van de 5 VA-G identiek scoorden.
- De grootste variatie in gescoorde klassen tussen de VA-G per film is waar te nemen voor de subrubrieken 3 'Slapen', 6 'Sociale communicatie' en 7 'Communicatieve vaardigheden'. Daar zien we respectievelijk slechts 2, 4 en 2 films -van de 24 films- waarbij 5 VA-G identieke klassen scoren.
- Bij de subrubrieken 6 'Sociale communicatie' en 7 'Communicatieve vaardigheden' zijn er bij respectievelijk 5 en 8 films maar 2 VA-G die identieke klassen scoren.

In de volgende tabel 4.11 worden de gegevens op dezelfde wijze beschreven, alleen wordt nu een verschil van 1 klasse van de 5e VA-G niet als een praktijkrelevant verschil beschouwd. De 5de VA-G met een afwijkende score ten opzichte van 4 VA-G met een identieke score wordt beschouwd als overeenstemmend, mits de afwijkende score 1 klasse is. In tabel 4.11 is deze situatie weergegeven in de kolom *'5 of 4* VA-G overeenstemming'*. Wanneer de 5e VA-G een afwijkende score heeft van meer dan 1 klasse komt deze in de laatste kolom van tabel 4.11, *'4**, 3 of 2 VA-G overeenstemming'*. Hierbij wordt 'niet aan de orde gekomen' gezien als een aparte klasse, die afwijkend is.

Tabel 4.11: Aantal VA-G, N(%) met overeenkomstige gescoorde klasse per subrubriek (1e meting), 5 of 4 VA-G overeenstemming

Subrubriek		*5 of 4* VA-G overeenstemming*	*4**, 3 of 2 VA-G overeenstemming*
1	Hygiëne zelfzorg	22 (91.7)	2 (8.3)
2	Mobiliteit	19 (79.2)	5 (20.8)
3	Slapen	13 (54.2)	11 (45.8)
4	Seksuele functie	12 (50.0)	12 (50.0)
5	Basale communicatie	23 (95.8)	1 (4.2)
6	Sociale communicatie	10 (41.7)	14 (58.3)
7	Communicatieve vaardigheden	6 (25.0)	18 (75.0)
8	Structuur aanbrengen	15 (62.5)	9 (37.5)
9	Huishoudelijke activiteiten	18 (75.0)	6 (25.0)
10	Omgaan stressvolle gebeurtenissen	13 (54.2)	11 (45.8)

N=24 films

* 4 VA-G overeenstemming en de 5e VA-G niet meer dan 1 klasse verschil van de andere 4 VA-G

** 4 VA-G overeenstemming en de 5e VA-G meer dan 1 klasse verschil van de andere 4 VA-G

Wat valt op:

- Voor 5 van de 10 subrubrieken geldt dat in ≥ 62.5% van de films 5 of 4 VA-G overeenstemmen in hun gescoorde klasse op die specifieke subrubriek.
- Bij subrubriek 1 ‘Hygiëne zelfzorg’ en 5 ‘Basale communicatie’ scoren meer dan 91.7% overeenstemming (5 of 4 VA-G). Dit zijn tevens de twee subrubrieken waar het vaakst klasse 0 wordt gescoord (zie ook tabel 4.4).
- De subrubrieken 6 ‘Sociale communicatie’ en 7 ‘Communicatieve vaardigheden’ zijn de 2 subrubrieken met het laagste percentage overeenkomstige gescoorde klassen tussen de VA-G.

In tabel 4.12 wordt nader ingegaan op de resultaten zoals die zijn gepresenteerd in tabel 4.10. Wanneer in een groep van 5 VA-G de 5e VA-G een afwijkende klasse scoort dan de andere 4 VA-G, hoe groot is dan de afwijking van deze anders scorende VA-G? Dit wordt per subrubriek weergegeven in tabel 4.12.

Tabel 4.12: Voorkomen van de afwijkende scores van de 5e VA-G wanneer 4 VA-G overeenstemmend gescoord hebben, per subrubriek

Subrubriek		*-2*	*-1*	*0**	*+1*	*+2*	*+3*	*+4*	*NG*
1	Hygiëne zelfzorg				1				2
2	Mobiliteit		1		4				1
3	Slapen		6		5				1
4	Seksuele functie			1	1	2	1	1	3
5	Basale communicatie				2				1
6	Sociale communicatie		1		4	1			
7	Communicatieve vaardigheden				4				
8	Structuur aanbrengen				5	1			
9	Huishoudelijke activiteiten				5	1			
10	Omgaan stressvolle gebeurtenissen		2		3				1

N=61, dit zijn de films met 4 VA-G overeenstemmend gescoord uit tabel 4.10

0* 4 VA-G hebben "niet aan de orde gekomen" gescoord, 1 VA-G klasse 0.

Wat valt op:

- De afwijking van de 5e VA-G is meestal (72%) maar 1 klasse (kolom -1 of kolom +1).
- Een keer wordt door 4 VA-G de klasse 0 gescoord en door 1 VA-G klasse 4, bij de subrubriek 4 'Seksuele functie'. Dit is een afwijking die verder niet verklaard kan worden.
- De afwijking van de 5e VA-G is in n=9 films "niet aan de orde gekomen".

4.3.1 *Interdoktervariatie en percentage overeenstemming*

In tabel 4.13 wordt het percentage overeenstemming weergegeven (gepaard kwadratisch gewogen, zie ook bijlage 8). Er worden twee verschillende methoden van wegingen gepresenteerd: (1) vaststaand aantal klassen (in deze studie is dat n=5) en (2) daadwerkelijk aantal klassen (afhankelijk van het aantal klassen waarop bij de specifieke subrubriek is gescoord). Daarnaast wordt in de laatste kolom de overeenstemming weergegeven per klasse.

Tabel 4.13: Percentage interdokterovereenstemming tussen VA-G, per subrubriek*

Sub-rubriek	*Wegings-methode (1): Vaststaand aantal klassen*	*Wegings-methode (2): Daadwerkelijk aantal klassen*	*Overeenstemming per klasse*				
			0	1	2	3	4
1	100	98	98	98	-	-	-
2	99	98	83	81	97	100	-
3	97	97	91	69	66	86	95
4	78	78	77	92	90	96	98
5	100	97	97	97	-	-	-
6	93	93	75	59	78	92	93
7	93	93	68	55	68	93	97
8	97	97	76	75	91	100	100
9	98	92	81	82	94	-	-
10	97	97	86	68	82	97	98

N=24 films en 5 VA-G per film

*Gepaard kwadratisch gewogen

Wat valt op:

- Het percentage overeenstemming tussen de VA-G voor de verschillende subrubrieken is hoog. Voor de beide wegingsmethoden is de range van interdokterovereenstemming respectievelijk 78% - 100% en 78% - 98%.
- Klasse “0” en de hogere klasse “4” zijn gemakkelijker te scoren, hier wordt vaak een hoger percentage overeenstemming gevonden dan voor de klasse “1” en “2”.

In de onderstaande tabel 4.14 worden de resultaten van dezelfde analyses gepresenteerd, echter ditmaal is “niet aan de orde gekomen” in de berekeningen meegenomen als klasse 0.

Tabel 4.14: Percentage interdokterovereenstemming tussen VA-G, “niet aan de orde gekomen” meegenomen als klasse 0*

Sub-rubriek	*Wegings-methode 1: Vaststaand aantal klassen*	*Wegings-methode 2: Daadwerkelijk aantal klassen*	*Overeenstemming per klasse*				
			0	1	2	3	4
1	100	98	98	98	-	-	-
2	99	98	83	82	97	100	-
3	97	97	91	69	67	87	95
4	94	94	82	93	93	97	98
5	100	97	97	97	-	-	-
6	93	93	75	59	78	92	93
7	93	93	68	55	68	93	97
8	98	98	75	73	92	100	100
9	98	92	82	82	94	-	-
10	97	97	84	67	83	97	98

N=24 films en 5 VA-G per film.

*Gepaard kwadratisch gewogen.

Wat valt op:

- Voor de 9 subrubrieken verandert er nauwelijks iets.
- Alleen subrubriek 4 ‘Seksuele functie’ heeft een hoger percentage overeenstemming wanneer “niet aan de orde gekomen” meegenomen wordt als klasse 0, 94% t.o.v. 78%.
- Het percentage overeenstemming tussen de VA-G voor de verschillende subrubrieken is hoog. Voor de beide wegingsmethoden is de range van interdokterovereenstemming respectievelijk 93% - 100% en 92% - 98%.

4.4 Invaliditeitspercentage, IP

Deze paragraaf gaat in op de uitkomstmaat van de bestudeerde schattingsmethodiek, het invaliditeitspercentage (IP). Naast de verdeling van uitkomstmaten tussen de VA-G per film wordt ook ingegaan op de bijdragen van de verschillende rubrieken aan het invaliditeitspercentage. Tot slot wordt beschreven hoe de hoogte van het IP eruit komt te zien wanneer de verschillende subrubrieken even zwaar wegen bij de totstandkoming van het IP.
In tabel 4.15 is de verdeling van de afgeronde percentages IP per film weergegeven. Per film wordt de verdeling weergegeven van het afgeronde IP van de 5 VA-G en de spreiding tussen de laagste en hoogste IP per groepje van 5 VA-G.

Tabel 4.15: Verdeling van het afgerond % IP van 5 VA-G per film en de spreiding in % afgerond IP tussen de 5 VA-G per film

	Afgerond % IP										
Film	*0*	*5*	*10*	*15*	*20*	*25*	*30*	*35*	*40*	*45*	*Spreiding*
1		2	3								5
2		1		1	1	2					20
3	1	2	2								10
4	3		1	1							15
5	2	1	2								10
6	4	1									5
7				1			2	1		1	30
8				2		1	1	1			20
9					1	1	3				10
10			2	2	1						10
11	5										0
12			2	2	1						10
13	5										0
14			4				1				20
15	1	4									5
16	2	1	2								10
17				3	2						5
18		2	2	1							10
19		3	2								5
20	4		1								10
21		3	1		1						15
22	1		4								10
23		2	2		1						15
24		4		1							10

5 VA-G per film, “niet aan de orde gekomen” is in de berekeningen meegenomen als klasse 0.

In de laatste kolom van tabel 4.15 is in groen de spreiding van 0, 5 of 10% weergegeven. In geel is weergegeven de spreiding van 15% en in oranje is weergegeven de spreiding van 20, 25, of 30%.

Wat valt op:

- Bij 4 films is het maximale IP 5 %.
- Hoger dan een IP van 45% wordt niet gescoord.
- 17 van de 24 film laten een spreiding tussen de 5 VA-G zien van maximaal 10%.
- Daarvan hebben 7 films een zeer kleine spreiding tussen de VA-G: 0 of 5%.
- 7 van de 24 films laten een spreiding zien van meer dan 10% tussen de 5 VA-G.
- Film 7 laat de grootste spreiding zien tussen de VA-G: van 30% (van minimaal 15% tot maximaal 45%).

De 7 films met een spreiding in de uitkomstmaat, IP, van meer dan 10% worden in tabel 4.16 nader bekeken. Per subrubriek wordt de mate van overeenstemming van de 5 VA-G weergegeven.

Tabel 4.16: Zeven films nader bekeken: Aantal VA-G, N(%) met overeenkomstige gescoorde klasse per subrubriek

Subrubriek		*5*	*4*	*3*	*2*
		100% overeen-stemming	*4 VA-G overeen-stemming*	*3 VA-G overeen-stemming*	*2 VA-G overeen-stemming*
1	Hygiëne zelfzorg	6 (85.7)	1 (14.3)		
2	Mobiliteit	2 (28.6)	2 (28.6)	2 (28.6)	1 (14.3)
3	Slapen	4 (57.1)	1 (14.3)	2 (28.6)	
4	Seksuele functie	3 (42.9)	1 (14.3)	3 (42.9)	
5	Basale communicatie	5 (71.4)	2 (28.6)		
6	Sociale communicatie	1 (14.3)	1 (14.3)	2 (28.6)	3 (42.9)
7	Communicatieve vaardigheden		2 (28.6)	2 (28.6)	3 (42.9)
8	Structuur aanbrengen	1 (14.3)	1 (14.3)	5 (71.4)	
9	Huishoudelijke activiteiten	2 (28.6)	2 (28.6)	3 (42.9)	
10	Omgaan stressvolle gebeurtenissen	2 (28.6)	1 (14.3)	3 (42.9)	1 (14.3)

N=7 films, 5 VA-G; zelfde score is 100% overeenstemming

Wat valt op:

- De drie subrubrieken 6 ‘Sociale communicatie’, 7 ‘Communicatieve vaardigheden’ en 8 ‘Structuur aanbrengen’ laten de minste overeenstemming zien tussen de VA-G.
- Voor de subrubriek 7 ‘Communicatieve vaardigheden’ geldt dat bij geen van de 7 films een groep van 5 VA-G allemaal dezelfde klasse scoren (100% overeenstemming).

In tabel 4.17 zoomen we nogmaals in op de 7 films waar de meeste spreiding is tussen de VA-G in de uitkomstmaat IP. Per film wordt de spreiding weergegeven van de laagste en de hoogste rubriekscore per groep van 5 VA-G. De spreiding (verschil laagste scorende VA-G en hoogste scorende VA-G) wordt per rubriek weergegeven.

De rubriekscores komen als volgt tot stand: rubriek 1 'Activiteiten dagelijks leven' bestaat uit vier subrubrieken, rubriek 2 'Sociaal functioneren' uit drie subrubrieken, rubriek 3 'Concentratie, doorzettingsvermogen en tempo' uit twee subrubrieken en de laatste, rubriek 4 'Adaptatie stressvolle omstandigheden' bestaat uit één subrubriek.
De berekeningen gaan als volgt:
Rubriek 1= (S1+S2+S3+S4)/4, waarbij S=subrubriek
Rubriek 2= (S5+S6+S7)/3
Rubriek 3= (S8+S9)/2
Rubriek 4= S10
Vervolgens wordt er met de scores van de vier rubrieken een 'totaalrubriekscore' berekend =(R1+R2+R3+R4)/4. Deze 'totaalrubriek score' is een getal tussen 0 en 5 en vertegenwoordigt een percentage van 0 tot 100%, het invaliditeitspercentage (IP). Dit wordt afgerond en weergegeven in stappen van 5%, het afgeronde IP[22].

Tabel 4.17: Zeven films nader bekeken: Spreiding per rubriek en afgerond % IP voor de films met een IP-spreiding groter dan 10%

	Spreiding tussen de laagste en hoogste score binnen 5 VA-G				
Film	Rubriek 1	Rubriek 2	Rubriek 3	Rubriek 4	Afgerond % IP
2	0.25	1.33	1.50	1.00	20
4	1.00	0.67	1.00	1.00	15
7	0.75	2.00	2.00	2.00	30
8	0.50	1.67	0.50	2.00	20
14	1.00	1.00	1.50	1.00	20
21	0.75	1.00	0.50	1.00	15
23	0.75	1.67	0.50	0.00	15

Wat valt op:

- Bij rubriek 1 wordt relatief een lagere spreiding gevonden vergeleken met de andere rubrieken.
- Er bestaat voor deze 7 films geen eenduidig beeld over een specifieke rubriek met een grote spreiding en daarmee de oorzaak van de spreiding in het IP.
- De gevonden geringste overeenstemming in de subrubrieken 6, 7 en 8 (zie tabel 4.16) vallen respectievelijk onder rubriek 2, 2 en 3.

Film 7 heeft de grootste spreiding in IP, namelijk 30%. In tabel 4.18 wordt voor deze film nader gekeken naar de gescoorde klassen per subrubriek van de 5 VA-G.

[22] PTSS protocol – PTSS WIA IP Protocol, Project WIA Zorg wet, Defensie 2007

Tabel 4.18: Film 7 nader bekeken: de verdeling van de gescoorde klassen

Subrubriek		*0*	*1*	*2*	*3*	*4*	*5*	*NG*
1	Hygiëne zelfzorg	4	1					
2	Mobiliteit	2	3					
3	Slapen				2	3		
4	Seksuele functie	5						
5	Basale communicatie	5						
6	Sociale communicatie		1[a]		1	3		
7	Communicatieve vaardigheden		2	1	1	1[a]		
8	Structuur aanbrengen	1[a]	1	3				
9	Huishoudelijke activiteiten	1[a]	1	3				
10	Omgaan stressvolle gebeurtenissen			4		1[b]		

N=1 film x 5 VA-G

a = VA-G a scoort het laagst.

b = VA-G b scoort het hoogst.

Wat hierbij opvalt, is het volgende:

- De VA-G met de hoogste IP (45%), (VA-G b) heeft bij deze film op subrubriek 10 'Omgaan stressvolle gebeurtenissen' een relatief hoge klasse gescoord.
- De VA-G met de laagste IP (15%), (VA-G a) heeft op subrubriek 6 'Sociale communicatie' een relatief lage klasse gescoord.
- Wanneer deze twee waarden aangepast worden naar de meest voorkomende waarden van de andere VA-G dan wordt de spreiding tussen de VA-G een stuk lager; 15% (VA-G a 20% en VA-G b 35%).

Eerder is in tabel 4.15 de verdeling van het afgeronde IP weergegeven. In de volgende tabel 4.19 wordt op dezelfde wijze gescoord als bij tabel 4.15. Alleen wordt nu voor de 7 films met een spreiding van meer dan 10%, de meest afwijkende 5^{e} VA-G weggelaten. Dit is in de tabel weergegeven met een "-". De spreiding wordt voor deze situatie weergegeven in de laatste kolom.

Tabel 4.19: Verdeling van het afgerond % IP van 5 VA-G per film en de spreiding in % afgerond IP tussen de 5 VA-G per film*

	Afgerond % IP										
Film	*0*	*5*	*10*	*15*	*20*	*25*	*30*	*35*	*40*	*45*	*Spreiding*
1		2	3								5
2		-		1	1	2					10
3	1	2	2								10
4	3		1	-							10
5	2	1	2								10
6	4	1									5
7				-			2	1		1	15
8				2		1	1	-			15
9					1	1	3				10
10			2	2	1						10
11	5										0
12			2	2	1						10
13	5										0
14			4				-				0
15	1	4									5
16	2	1	2								10
17				3	2						5
18		2	2	1							10
19		3	2								5
20	4		1								10
21		3	1		-						5
22	0		4								10
23		2	2		-						5
24		4		1							10

N=113, 5 of 4 VA-G x 24 films, “niet aan de orde gekomen” is in de berekeningen meegenomen als 0.

* De 5e meest afwijkende VA-G is weggelaten in deze tabel bij de films met een spreiding groter dan 10%.

Wat valt op:

- 22 van de 24 films laten op deze manier een spreiding tussen de VA-G zien van maximaal 10%.
- Slechts 2 films laten dan een spreiding zien van 15%.
- Geen enkele film laat een spreiding tussen de VA-G zien groter dan 15%.

4.4.1 Totstandkoming IP

De uitkomstmaat IP is berekend volgens de ontwikkelde schattingsmethodiek. Elke rubriek kent een verschillend aantal subrubrieken (zie paragraaf 4.4).

In tabel 4.20 wordt ingegaan op de gemiddelde scores van de rubrieken en van het IP. De gemiddelde scores van de waarden op basis waarvan het afgeronde IP tot stand komt wordt weergeven: de vier rubriekscores, de totaalscore en het invaliditeitspercentage. De variërende n heeft te maken met het feit dat bij deze berekening de “niet aan de orde gekomen” scores als missing zijn meegenomen in de analyses.

Tabel 4.20: Gemiddelde scores, rubrieken en invaliditeitspercentages

Rubriek		*N*	*min*	*max*	*gemiddelde*	*SD*
1	Activiteiten dagelijks leven	91	0.0	1.50	0.5*	0.4
2	Sociaal functioneren	119	0.0	2.67	0.7	0.6
3	Concentratie, doorzettingsvermogen en tempo	115	0.0	2.0	0.3	0.5
4	Adaptatie stressvolle omstandigheden	116	0.0	4.0	1.0	0.8
	Totaalrubriekscore	83	0	2.4	0.7	0.5
	Invaliditeitspercentage	83	0	47.9	13.5	10.0
	Afgerond Invaliditeitspercentage	83	0	45.0	11.3	9.7

N_{max}=5 artsen x 24 films=120, * de getallen zijn afgerond

Wat valt op:

- De gemiddelde score van rubriek 4 ‘Adaptatie stressvolle omstandigheden’ is relatief hoog.
- Rubriek 3 ‘Concentratie, doorzettingsvermogen en tempo’ heeft de laagste gemiddelde score van de 4 rubrieken.

De volgende tabel 4.21 heeft dezelfde basis als tabel 4.20. Het verschil is, dat in tabel 4.21 een klasse 0 toegekend is wanneer “niet aan de orde gekomen” is gescoord bij een subrubriek. De N per rubriek is hiermee steeds 120.

Tabel 4.21: Gemiddelde scores per rubriek en invaliditeitspercentages, berekend met “niet aan de orde gekomen” als klasse 0

Rubriek		*N*	*min*	*max*	*gemiddelde*	*SD*
1	Activiteiten dagelijks leven	120	0.0	1.5	0.5*	0.4
2	Sociaal functioneren	120	0.0	2.7	0.7	0.6
3	Concentratie, doorzettingsvermogen en tempo	120	0.0	2.0	0.3	0.5
4	Adaptatie stressvolle omstandigheden	120	0.0	4.0	1.0	0.8
	Totaalrubriekscore	120	0.0	2.4	0.6	0.5
	Invaliditeitspercentage	120	0.0	47.9	12.5	9.7
	Afgerond invaliditeitspercentage	120	0	45.0	10.2	9.5

N=5 artsen x 24 films=120, *de getallen zijn afgerond

Wat valt op:

- Het gemiddelde afgeronde IP daalt licht van 13.5% (zie tabel 4.20) naar 12.5% wanneer “niet aan de orde gekomen” als klasse 0 wordt berekend. De gemiddelde scores van de rubrieken blijven nagenoeg hetzelfde.

In tabel 4.22 staan tenslotte de resultaten weergegeven wanneer het IP op twee verschillende manieren wordt berekend: (A) de formule van de schattingsmethodiek, waarbij gerekend wordt met vier rubrieken; (B) de berekening waarbij alle tien subrubrieken een even grote bijdrage hebben. Zowel de totaalrubriekscore, het IP als het afgeronde IP wordt daarbij weergegeven[23].

Tabel 4.22: Totaalrubriekscore, IP en afgerond IP berekend op twee manieren (A en B)

Rubriek	*A: 4 rubrieken*					*B: Subrubrieken even zwaar*				
	n	*min*	*max*	*gem*	*SD*	*n*	*min*	*max*	*gem*	*SD*
Totaalrubriekscore	83	0	2.4	0.7	0.5	83	0.0	2.0	0.6	0.4
IP	83	0	47.9	13.5	10.0	83	0.0	40.0	12.1	8.9
Afgerond IP	83	0	45.0	11.3	9.7	83	0.0	40.0	10.2	9.2

N_{max}=5 artsen x 24 films=120

Wat valt op:

- Over het geheel genomen komen de waarden iets hoger uit wanneer de formule van de huidige schattingsmethodiek wordt toegepast. Gemiddeld voor het IP 11.3% versus 10.2% bij een berekening waarin de 10 subrubrieken allemaal evenveel meewegen in het totaal.

[23] In bijlage 10 is een extra tabel weergegeven waarin vier hypothetische situaties worden doorgerekend naar het IP.

Hoofdstuk 5 Conclusies en discussie

In het huidige onderzoek staat een nieuw instrument centraal dat sinds 1 juli 2008 wordt gebruikt in het Zorgloket MOD. Het instrument is een schattingsmethodiek voor het bepalen van het percentage invaliditeit bij militairen met psychische aandoeningen. In de praktijk gebeurt dat bij post-actieve militairen door een verzekeringsarts van het MOD, die naar aanleiding van een beoordelingsgesprek met een post-actieve militair de ernst van de beperkingen scoort op 4 rubrieken, die onderverdeeld zijn in 10 subrubrieken[24].
Het huidige onderzoek richt zich op de variatie in uitkomsten van de schattingsmethodiek wanneer vijf artsen deze toepassen op een gefilmd beoordelingsgesprek. De *intra*doktervariatie is de variatie binnen eenzelfde arts bij het beoordelen van dezelfde film met een tussenperiode van 2 maanden. De *inter*doktervariatie is onderzocht door de variatie in uitkomsten tussen vijf artsen te bestuderen. Een geringe variatie binnen een arts en tussen artsen zijn goede indicaties voor een betrouwbaar instrument.
In dit laatste hoofdstuk wordt allereerst puntsgewijs ingegaan op de conclusies van het onderzoek. Daarna volgen enkele discussiepunten die naar aanleiding van het onderzoek kunnen worden geformuleerd.

5.1 Conclusies

Subrubrieken die "niet aan de orde" zijn gekomen

- Over het algemeen genomen is er in het gehele onderzoek door de VA-G heel weinig gebruik gemaakt van de scoringsmogelijkheid "niet aan de orde gekomen" [25]. Dat betekent dat in de gefilmde gesprekken voldoende informatie aanwezig is om de 10 subrubrieken te kunnen scoren.
- Dit komt overeen met de conclusie van het expertpanel, die alle films heeft gezien tijdens het inclusie-, en exclusietraject. Alle films die zij hebben bekeken bevatten voldoende informatie en zijn geïncludeerd in het onderzoek.
- Bij de subrubriek 4 'Seksuele functie' wordt het vaakst "niet aan de orde gekomen" gescoord. Dat geldt zowel tijdens de 1e als de 2e meting.

Verdeling van de gescoorde klassen per subrubriek

- Er bestaat een scheve verdeling van de gescoorde klassen die de ernst van de beperkingen aangeven. Klasse 0, wat staat voor "normaal", is relatief veel gescoord. Klasse 5, wat staat voor "extreme beperkingen", is nooit gescoord. De tussenliggende klassen 1 tot en met 4 zijn wel gescoord, met een duidelijk accent op de klassen 1 en 2.
- Bij de subrubrieken 3 'Slapen', 10 'Omgaan met stressvolle gebeurtenissen', 6 'Sociale communicatie' en 7 'Communicatieve vaardigheden' zijn relatief vaker hogere klassen gescoord.

[24] Het gaat om de subrubrieken: 1 'Persoonlijke hygiëne en zelfzorg', 2 'Mobiliteit', 3 'Slapen', 4 'Seksuele functie', 5 'Basale communicatie', 6 'Communicatief sociaal functioneren', 7 'Communicatieve vaardigheden', 8 'Structuur aanbrengen', 9 'Huishoudelijke activiteiten', 10 'Omgaan met stressvolle gebeurtenissen'.
[25] In totaal zijn er per subrubriek n=120 scores gegeven (n=5 artsen x n=24 films). Voor alle subrubrieken samen is dat n=1200. Tijdens de 1e meting is daarvan n=40 keer "niet aan de orde" gescoord (3.3%). Voor de 2e meting is dat n=35 (2.9%).

- De scheve verdeling van de gescoorde klassen maakt een betrouwbare berekening van kappa's niet mogelijk.

Intradoktervariatie

- Over het geheel genomen is de *intra*doktervariatie klein. Met andere woorden, de overeenstemming binnen een VA-G is hoog tussen de gescoorde klasse op de 1e meting in vergelijking met de gescoorde klasse –voor dezelfde film- op de 2e meting.
- Wanneer een arts verschillend scoort op de 2e meting in vergelijking met de 1e meting dan is dit verschil, per subrubriek, in minder dan 6% 2 klassen of meer.
- Het percentage overeenstemming (kwadratisch gewogen) binnen een VA-G voor alle films en alle VA-G is hoog, voor alle subrubrieken variërend van 84% - 100%[26].
- Wanneer er een verschil wordt gevonden binnen een VA-G tussen de gescoorde klasse van dezelfde film voor de 1e en de 2e meting, dan is dat verschil voor het overgrote deel niet meer dan 1 klasse[27].
- Wanneer de analyses zijn uitgevoerd met de aanname dat een variatie van inschatting van 1 klasse in de praktijk niet als een praktijkrelevant verschil hoeft te worden aangemerkt[28], dan varieert de overeenstemming binnen een VA-G per subrubriek van 91% - 98%.

Interdoktervariatie

- Over het geheel genomen laten de resultaten meer variatie zien in gescoorde klassen *tussen* vijf VA-G dan in gescoorde klassen *binnen* een VA-G.
- Er zijn twee subrubrieken waar alle vijf VA-G vaak overeenstemmende klassen scoren, namelijk bij 88% van de gescoorde films bij zowel 1 'Hygiëne zelfzorg' en 5 'Basale communicatie'. Voor beide subrubrieken geldt dat nagenoeg alle VA-G de klasse 0 scoren.
- Voor de meeste subrubrieken, 6 van de 10, blijkt dat voor een groot deel van de films (67% - 100%) vier of vijf VA-G overeenstemmend scoren.
- Wanneer er een verschil wordt gevonden tussen de vijf artsen, is de afwijking van de vijfde arts meestal, 72% van de gevallen maar 1 klasse.
- De grootste variatie in gescoorde klassen tussen vier of vijf VA-G is gevonden voor de subrubrieken 6 'Sociale communicatie' en 7 'Communicatieve vaardigheden'. Bij deze subrubrieken wordt respectievelijk 25% en 41.7% overeenstemming gevonden tussen vier of vijf artsen.
- Het percentage overeenstemming (gepaard kwadratisch gewogen) in gescoorde klassen voor alle films en alle VA-G is hoog, voor alle subrubrieken variërend van 92% - 100%[29]. Overall is de *inter*doktervariatie klein.
- Voor de klasse "0" en "4" wordt een hoger percentage overeenstemming gevonden dan voor de klasse "1" en "2".

[26] Dit geldt voor de analyses waarbij "niet aan de orde gekomen" is berekend als de klasse 0.
[27] Ook hier geldt dat er voor alle subrubrieken samen n=1200 scores zijn gegeven. Slechts in 2% daarvan, n=24, is er meer dan 1 klasse verschil gescoord binnen een VA-G.
[28] Beslissing van de klankbordgroep dd 18 mei 2011.
[29] Dit geldt voor de analyses waarbij "niet aan de orde gekomen" is berekend als de klasse 0.

Invaliditeitspercentage, de uitkomstmaat IP

- De variatie tussen vijf VA-G in de uitkomstmaat afgerond IP per film is relatief klein: bij 17 van de 24[30] films is deze spreiding maximaal 10%.
- De spreiding in afgerond IP voor de 7 films waar de variatie groter is dan 10% wordt voor het grootste deel verklaard door spreiding op subrubriek 6, 7 en 8.Wanneer in de groep van vijf VA-G de vijfde VA-G met de meest afwijkende gescoorde klasse uit de analyse wordt weggelaten –voor 7 films wordt gerekend met een groepje van vier VA-G en voor 17 films met een groepje van vijf VA-G- dan geldt voor 22 van de 24 films een maximale spreiding van 10% afgerond IP tussen vier of vijf VA-G.

Overall conclusie

- De *intra*doktervariatie is laag, de *intra*dokterbetrouwbaarheid van het nieuwe schattingsinstrument is goed tot zeer goed.
- De *inter*doktervariatie is relatief laag, de *inter*dokterbetrouwbaarheid van het nieuwe schattingsinstrument is goed.
- De non-respons analyse geeft geen aanleiding om te veronderstellen dat er vertekenende selectie heeft plaatsgevonden in het onderzoek. De resultaten en de conclusies geven een representatief beeld voor de groep waarbij het schattingsinstrument in de praktijk wordt toegepast.

5.2 Discussie

Subrubrieken die "niet aan de orde" zijn gekomen

- In de totale studie is voor een zeer klein aantal subrubrieken de score "niet aan de orde gekomen" gekozen door de VA-G. Daarbij kan de vraag worden gesteld wat de betekenis is van deze score: (1) Is het onderwerp van de betreffende subrubriek niet aan bod gekomen in het gesprek, of (2) Is de informatie over de betreffende subrubriek niet opgemerkt door de VA-G? Een duidelijk antwoord kan echter niet worden gegeven, ook door het kleine aantal. De resultaten voor subrubriek 4 'Seksuele functie' laten echter zien dat VA-G die tijdens de 1e meting de mogelijkheid "niet aan de orde gekomen" scoren, dit ook doen – voor dezelfde film- tijdens de 2e meting (zie ook tabel 4.3). Deze hoge *intra*dokterovereenstemming doet veronderstellen dat het onderwerp niet in het gesprek aan bod is gekomen.
- Een tweede opmerking betreft de betekenis van "niet aan de orde gekomen" in termen van de analyse en de resultaten van de studie. In het huidige onderzoek is daarmee op twee manieren omgegaan: (1) als missing data, en (2) als een score klasse 0. Beide manieren zijn echter inhoudelijk niet correct. Er is immers geen sprake van een missende waarde; het gegeven dat er niet is gescoord heeft een andere betekenis. Daarnaast kan er niet zonder meer vanuit worden gegaan dat een klasse 0 "normaal" gescoord had kunnen worden in de situatie van "niet aan de orde gekomen". Het is aan de andere kant ook niet aannemelijk dat in geval van ernstige beperkingen een subrubriek niet aan de orde komt in een gesprek. Dat blijkt ook uit de data: de subrubrieken waarvoor de hoogste klassen worden gescoord zijn ook de subrubrieken waar de score "niet aan de orde gekomen" ontbreekt. Het

[30] Dit geldt voor de analyses waarbij "niet aan de orde gekomen" is berekend als de klasse 0.

nauwelijks voorkomen van de score “niet aan de orde gekomen”, plus het geringe verschil in resultaat van beide manieren van analyseren geven aanleiding te veronderstellen dat dit geen effect heeft op de conclusies van de studie.

Verdeling van de gescoorde klassen per subrubriek

- De resultaten laten een duidelijke scheve verdeling zien van de gescoorde klassen. Klasse 5 is nooit gescoord en hoe lager de score, hoe vaker deze voorkomt. De scheve verdeling maakt een betrouwbare berekening van kappa’s niet mogelijk. De gebruikte beschrijvingen van de data en toegepaste statistieken zijn helder om betekenisvolle conclusies te kunnen trekken.
- De scheve verdeling van de klassen en daarbij vooral het relatief vaak voorkomen van de score klasse 0 houdt in dat er een relatief kleine variatie is tussen de verschillende klassen. Dit heeft als consequentie dat er ook een relatief kleine variatie bestaat tussen de VA-G. Tegelijkertijd laten de resultaten zien dat er wel degelijk gescoord wordt voor de klassen 1 tot en met 4. Dat geldt voornamelijk voor de subrubrieken ‘Slapen’, ‘Sociale communicatie’, ‘Communicatieve vaardigheden’ en ‘Omgaan met stressvolle gebeurtenissen’. Bovendien valt op te merken dat ook overeenstemming binnen een VA-G of tussen een groep VA-G met de score klasse 0 voor een specifieke subrubriek geen variatie betekent en daarmee duidt op betrouwbaarheid van het instrument.

Intradoktervariatie

- In het huidige onderzoek is gekozen voor een interval van twee maanden tussen de 1^{e} en de 2^{e} meting voor de berekening van *intra*doktervariatie. Een scan van de literatuur gaf geen duidelijke aanwijzing voor een ideaal interval bij onze onderzoekvragen. De vraag die gesteld kan worden is of de tussentijd mogelijk te kort was, zodat de VA-G de gescoorde klassen van de 1^{e} meting zich nog konden herinneren bij de 2^{e} meting. In dat geval zou de *intra*dokterbetrouwbaarheid overschat zijn. Er lijkt echter geen sprake te zijn van overschatting, omdat de deelnemende VA-G desgevraagd aangeven dat ze de gescoorde klassen van de 1^{e} meting niet meer weten tijdens de 2^{e} meting.
- Een vraag in het verlengde hiervan is, of er gedurende de looptijd van het onderzoek sprake is van een leereffect? In totaal zijn er drie sessies van bijeenkomsten met VA-G geweest waarin films werden bekeken en gescoord. Een leereffect zou betekenen dat in de derde sessie de *intra*doktervariatie kleiner zou zijn dan in de eerste sessie. De resultaten van afzonderlijke analyses voor de *intra*doktervariatie (zie tabel 4.7) laten zien dat er geen sprake is van een leereffect. Ook analyses op het voorkomen van de score “niet aan de orde gekomen” voor de drie sessies wijzen niet op een leereffect.
- Een discussiepunt over de lage intradoktervariatie in deze studie is de scheve verdeling van de gescoorde klassen in de richting van lage klassen. Daarbij kan worden opgemerkt dat deze wel representatief zijn voor de groep post-actieve militairen waarbij het instrument in de praktijk wordt toegepast (zie non-respons analyse).

Interdoktervariatie

- Een relevante vraag die in het huidige onderzoek meermaals is gesteld luidt: “Wat is een verschil?”. Kan slechts dan gesproken worden van overeenstemming wanneer er sprake is van 100% dezelfde score binnen een groep van 5 VA-G, of is een afwijkende score van een

van de 5 VA-G ook als overeenstemming te beschouwen? En is er ook sprake van overeenstemming als er 1 klasse verschil in score voorkomt? In verschillende discussies met zowel het expertpanel als de klankbordgroep kwam naar voren dat in de praktijk een verschil van 1 klasse ook als overeenstemming kan worden beschouwd. Een verschil van 2 klassen is een echt verschil. In de analyse en de presentatie van de resultaten hebben we ervoor gekozen om beide manieren te presenteren: (1) 100% overeenstemming, en (2) 1 klasse verschil te beschouwen als overeenstemming. Uiteraard geldt dat de overeenstemming in het tweede geval hoger is, maar ook in het eerste geval geldt dat het instrument weinig *intra*doktervariatie en *inter*doktervariatie laat zien. Daarnaast is ook het kwadratisch gewogen percentage overeenstemming weergegeven, in deze berekening weegt een verschil van 1 klasse minder zwaar dan een verschil van 2 klassen.
- In de literatuur en in vergelijkbaar onderzoek wordt een percentage overeenstemming van 70% of hoger als goed beschouwd en van 90% of meer als excellent[31].

Invaliditeitspercentage, de uitkomstmaat IP
- Bij de uitkomstmaat % IP (afgerond) kan de vraag gesteld worden hoe deze tot stand komt. Het PIM-protocol hanteert een formule waarin de tien subrubrieken een verschillend gewicht hebben voor de uitkomstmaat, omdat elke rubriek is samengesteld uit een verschillend aantal subrubrieken (variërend van 1 tot 4). De resultaten van dit onderzoek laten zien dat de gemiddelde uitkomst van de IP echter nauwelijks verschilt wanneer deze op verschillende manieren wordt berekend: te weten de berekening volgens het PIM-protocol of de berekening waarin alle subrubrieken evenveel gewicht in de schaal leggen. De methode van het PIM-protocol laat gemiddeld een iets hogere IP zien.
- Verder valt hierbij nog op te merken dat de subrubriek 10 “Omgaan met stressvolle gebeurtenissen” relatief veel gewicht (1/4 x daadwerkelijke score) in de schaal legt en dat deze subrubriek een lage *intra*doktervariatie en *inter*doktervariatie kent. De subrubriek 3 ‘Slapen’ kent een relatief hogere variatie, maar legt ook relatief minder gewicht in de schaal (1/16 x daadwerkelijke score) bij de totstandkoming van de IP.
- In dit verband kan de vraag worden geformuleerd wat er gebeurt met het % IP wanneer er op 1 subrubriek 1 klasse verschil wordt doorgerekend. Met de gegevens zijn de hypothetische situaties geanalyseerd en doorgerekend voor de subrubriek 3 ‘Slapen’ en de subrubriek 10 ‘Omgaan met stressvolle gebeurtenissen’: wat is het effect op het % IP van min 1 klasse of van plus 1 klasse? De effecten voor de subrubriek 3 ‘Slapen’ zijn kleiner dan voor de subrubriek 10 ‘Omgaan met stressvolle gebeurtenissen’ (zie bijlage 10)
- Tot slot past bij dit onderdeel een opmerking over de subrubrieken 6 en 7 over ‘Sociale communicatie’ en ‘Communicatieve vaardigheden’. Beide subrubrieken vallen in negatieve zin op in de *intra*doktervariatie en *inter*doktervariatie. Ook tijdens een discussie bij de afronding van de gegevensverzameling met de VA-G kwam naar voren dat juist deze beide subrubrieken de meeste interpretatieruimte bevatten bij het toepassen van het instrument op de gefilmde beoordelingsgesprekken. De formulering van de klassen bij deze specifieke subrubrieken verdient aandacht bij een eventuele revisie van de schattingsmethodiek.

[31]Spanjer J, Krol B, Popping R, Groothoff JW, Brouwer S. Disability assessment interview: the role of detailed information on functioning in addition to medical history-taking. Journal of Rehabilitation Medicine 2009;41:267-72
Landis JR, Kocht GG. The measurement of Observer Agreement for Categorical Data. Biometics, Vol. 33, No1. 1977, pp 159-174

Algemene discussiepunten:

- Het huidige onderzoek is een betrouwbaarheidstudie van een schattingsinstrument. Dat is gedaan op basis van het bestuderen van de *intra*doktervariatie en de *inter*doktervariatie. Een belangrijke aanvulling hierbij is, dat een goede betrouwbaarheid van een instrument een beperkte betekenis heeft voor een uitspraak over de validiteit van dat instrument. Bij de validiteit gaat het over de vraag of het instrument ook daadwerkelijk meet wat het beoogt te meten. Een instrument dat niet betrouwbaar is en bij herhaald invullen onder exact dezelfde omstandigheden tot andere resultaten leidt kan nooit valide zijn. Een goede betrouwbaarheid is in die zin een van de onderdelen van het bepalen van de validiteit van het instrument.
- Wel kan op basis van het huidige onderzoek worden geconcludeerd dat het instrument betrouwbaar is. De uitkomst van het schattingsinstrument bij de beoordeling van de ernst van beperkingen bij psychische aandoeningen van een post-actieve militair is onafhankelijk van de - getrainde - verzekeringsarts die het instrument toepast.

Bijlagen

Bijlage 1 Achtergrond van de opdracht van het Ministerie van Defensie

De nieuwe schattingsmethodiek staat beschreven in het rapport "*Schade in Schalen – de psyche in beeld"*[32] (2005) en in het *"PTSS – PTSS WIA IP Protocol"* (2007)[33]. Dit protocol is aangeboden aan de Tweede Kamer en aan de Bonden. Het protocol is per ministeriële regeling ingevoerd. Bij invoering is toegezegd de nieuwe schattingssystematiek te valideren. De vragen, die hierbij naderhand op de voorgrond traden, luidden: Is valideren überhaupt mogelijk? Wat heeft men met die opmerking bedoeld? Om deze vragen te beantwoorden werd een forum van deskundigen geraadpleegd. Dat proces, met uitkomsten en voorstellen van het forum en de keuzes die het Ministerie van Defensie op basis daarvan heeft gemaakt, staan hieronder kort beschreven.

Een nadere specificatie van "Validatie": Brainstormsessie Defensie met forum van deskundigen
Om het begrip validatie nader te specificeren en uit te werken is een brainstormsessie georganiseerd met een forum van hoogleraren. De concrete vraag aan dit forum luidde: "Geef aan hoe en welk onderzoek het best kan worden uitgevoerd om, op zo verantwoorde wetenschappelijk acceptabele wijze, zo snel en goedkoop mogelijk, antwoord te krijgen op de vraag of de richtlijnen betrouwbaar/valide zijn."
Het forum van hoogleraren bestond uit:

- Prof. Dr. R.van der Bosch, hoogleraar psychiatrie en voorzitter NVVP
- Prof. Dr. F. Koerselman, hoogleraar psychiatrie
- Prof. Dr. H. Willems, bijzonder hoogleraar sociale verzekeringsgeneeskunde
- Prof. Dr. D. Wiersma, hoogleraar epidemiologie

Uitkomsten brainstormsessie met forum van deskundigen
De belangrijkste uitkomst van de brainstormsessie was de conclusie dat een rechtstreeks validatie onderzoek niet mogelijk was, omdat er een "gouden standaard" ontbrak.
Ook was het niet (goed) mogelijk om de nieuwe schattingsmethodiek te vergelijken met de "oude methodiek", want: (1) er zijn geen "oude" richtlijnen; (2) hierdoor ook een enorme discretionaire ruimte bestaat; (3) als een arts onderricht heeft gehad in een nieuwe wijze van werken en indirect ook is gewezen op zijn tekortkomingen kan hij bij een beoordeling feitelijk niet meer scoren op de oude manier. De vergaarde kennis en kunde spelen dan toch mee op de achtergrond.
Rechtstreekse validatie was niet mogelijk. Daarom werd gekozen om indirect te valideren.

Indirecte validatie is mogelijk door:

1. aan te geven dat de richtlijnen door ter zake deskundigen zijn gemaakt,
2. adviezen van leden van de Gezondheidsraad
3. beoordeling door de RZO, Raad voor Zorg en Onderzoek Veteranen (Raad Tiesinga)
4. een betrouwbaarheidsonderzoek met inter- en intra- doktervariatie
5. beoordeling van deze vier items door de klankbordgroep

Voorstellen voor opzet betrouwbaarheidsonderzoek van forum van deskundigen
Het forum kwam tot de conclusie dat een betrouwbaarheidsonderzoek naar de nieuwe schattingsmethodiek, binnen de genoemde kaders, het best tegemoet kwam aan de vraag van de Tweede Kamer en de Bonden. Daarvoor zagen zij de volgende drie mogelijkheden:

1. Dubbele separate spreekkamerbeoordeling door twee artsen op één dag met filmopnames. Deze filmopnames dan voorleggen aan een groep verzekeringsartsen.

[32] *Schade in Schalen – de psyche in beeld. Het vaststellen van de mate van psychische invaliditeit in het kader van de militaire pensioenvoorschriften.* Advies Commissie WPC-PIM, 's Gravenhage, juni 2005
[33] PTSS protocol – PTSS WIA IP Protocol, Project WIA ZORG wet, Defensie 2007

2. Dubbele beoordeling door twee artsen op hetzelfde moment met filmopname en separate rapportages. Eén dokter voert het gesprek, de andere luistert.
3. Filmopname van de beoordeling door één arts. Deze filmopnames dan voorleggen aan een groep verzekeringsartsen.

Keuze van het Ministerie van Defensie.
Op basis van de hierboven genoemde uitkomsten en voorstellen van het forum van deskundige hoogleraren heeft het Ministerie van Defensie de keuze laten vallen op de derde mogelijkheid voor de uitvoering van het betrouwbaarheidsonderzoek van de nieuwe schattingsmethodiek: filmopname van de beoordeling door één arts (één beoordeling) en deze filmopnames voorleggen aan een groep verzekeringsartsen. De motivatie voor deze keuze, is gelegen in de volgende punten: (1) gevoeligheid materie; (2) kwetsbaarheid en belasting onderzoeksgroep; (3) terughoudendheid artsen van het Zorgloket, en (4) praktische overwegingen.

Vraagstelling van het betrouwbaarheidsonderzoek
Daarmee was de opdracht geformuleerd tot de uitvoering van een onderzoek met als centrale vraagstelling: “Is de nieuwe schattingsmethodiek ter beoordeling van de mate van invaliditeit bij psychische aandoeningen betrouwbaar en zo ja in welke mate?”

Context van het onderzoek: haalbaarheid en beperkingen
Tijdens de aanloop van het onderzoek is er een aantal punten naar voren gekomen die van invloed zouden kunnen zijn op het onderzoek. Deze punten zijn hieronder weer gegeven.

- Het causaliteitsvraagstuk is vaak niet eenduidig bij psychisch letsel. De opdracht voor het huidige onderzoek betrof niet de betrouwbaarheid van de beoordeling van het causaliteitsvraagstuk, maar sec de betrouwbaarheid van de beoordelingssystematiek van beperkingen. Als twee onderwerpen tegelijk in het onderzoek met elkaar worden vergeleken, is de kans op ruis erg groot. Daarnaast waren de artsen van het Zorgloket dit aangaande terughoudend. Tot slot bestond de kans dat belangenbehartigers dan mogelijk gegevens opvragen om te gebruiken in juridische procedures tegen het Ministerie van Defensie.
- De belasting van de betrokkene: de groep waarover het onderzoek gaat, post-actieve militairen die een rekest indienen vanwege psychische klachten (PTSS) is een kwetsbare groep. De inschatting bestond dat bij deze groep weerstand en reserve aangaande dit onderzoek zou zijn. Datzelfde gold voor de Bonden.
- Pre – existentie / endogene predispositie: ook hierop wordt in het huidige onderzoek geen onderscheid gemaakt, in verband met de grote discretionaire ruimte dit aangaande en daardoor de beïnvloeding op de scores van de beperkingen (en uiteindelijk het percentage invaliditeit).
- Financiële kanten: de begroting van het Ministerie van Defensie stond bij de toezegging van een onderzoek bij de invoering van de nieuwe methodiek al onder grote druk door vele uitzendingen. Het huidig onderzoek was het maximaal haalbare.

Bijlage 2: Organisatie en betrokken partijen

Onderzoeksteam TGO UMCG
Dr. J Tuinstra
Dr. AS Fokkens
Drs. NP Verheij
Dr. RH Bakker
Prof. dr. JW Groothoff
Prof. dr. JJL van der Klink

Wetenschappelijke adviesgroep
Dr. S Brouwer
Drs. AM van de Ven

Statistiek
Dr. R Popping
Dr. RE Stewart

Ministerie van Defensie / hoofddirectie Personeel (opdrachtgever)
Lgen. JGA Leijh
JH Paulusma-de Waal, arts M&G

Expertpanel
Het expertpanel is ingesteld door de opdrachtgever en heeft twee taken: (1) verzorgt de training in het gebruik van de schattingsmethodiek door de VA-G; (2) beoordeelt of de films met beoordelingsgesprekken geïncludeerd kunnen worden.
JH Wijers, verzekeringsarts
S Knepper, verzekeringarts

Klankbordgroep
De klankbordgroep is ingesteld door de opdrachtgever en heeft als taak te toetsen of het onderzoek volgens de regelen der kunst van de wetenschap wordt uitgevoerd en gerapporteerd. De klankbordgroep is vier keer bijeen geweest, een bijeenkomst over de opzet van het onderzoek, twee bijeenkomsten over tussenrapportages van resultaten, waarin beslissingen zijn genomen over analyses en een bijeenkomst waarin het eindconcept van het rapport is besproken.
Prof.dr. JHBM Willems
Prof.dr. Koerselman
Dr. WAHJ van Stiphout, arts M&G, epidemioloog B
Dr. NHTh Croon, (voorzitter), verzekeringsarts

Zorgloket MOD
Vier verzekeringsartsen, VA-U
Secretariaat/beheer
Dhr. J Rouschop Management
***Dhr. Reinoudt (vanaf december 2010)
Dhr. C Kolenberg Manager zorgcoördinatie
Dertien zorgcoördinatoren / casemanagers

Onafhankelijke verzekeringsartsen
Veertien verzekeringsartsen, VA-G

Bijlage 3 Folder met toelichting voor de cliënt.

NB: de opmaak die hier is weergegeven is platte tekst, de cliënt ontvangt een folder in kleur.

Geachte meneer/mevrouw,
Wij vragen u vriendelijk om mee te werken aan een onderzoek. Het betreft een onderzoek naar de nieuwe schattingsmethodiek voor het bepalen van het percentage van invaliditeit bij psychische aandoeningen. Deze schattingsmethodiek wordt toegepast door de verzekeringsartsen van het Zorgloket MOD (Militaire Oorlogs- en Dienstslachtoffers). Het onderzoek wordt verricht op verzoek van de Tweede Kamer en de Bonden. Het Universitair Medisch Centrum Groningen (UMCG) voert het onderzoek uit. U beslist zelf of u mee wilt doen. Voordat u deze beslissing neemt is het belangrijk om meer te weten over het onderzoek. Deze informatie kunt u lezen in deze folder.

Waarom dit onderzoek?

Voor de invaliditeitsbeoordeling gebruikt het Zorgloket MOD sinds 1 juli 2008 richtlijnen, waaronder het PTSS (Post Traumatische Stressstoornis) protocol. Dit protocol maakt gebruik van een unieke methode om de mate van invaliditeit te schatten bij psychische aandoeningen. De methode is gemaakt door deskundigen en positief beoordeeld door leden van de Gezondheidsraad. Het is nu van belang te onderzoeken, hoe betrouwbaar de methode is in de praktijk. De betrouwbaarheid geeft aan in hoeverre verschillende artsen bij de beoordeling van één en dezelfde persoon tot vergelijkbare invaliditeitspercentages komen. De betrouwbaarheid heeft geen relatie met de hoogte van het invaliditeitspercentage. Een eventuele verbetering van de schattingsmethodiek aan de hand van de onderzoeksresultaten zal dan ook niet leiden tot wijzigingen van de invaliditeitspercentages.

De uitvoering van het onderzoek.

Het onderzoek wordt verricht bij het Zorgloket MOD. Het UMCG gaat gesprekken van veteranen met de verzekeringsarts op film opnemen. Aan de hand van deze films kunnen meerdere onafhankelijke artsen tegelijkertijd bij dezelfde persoon het percentage invaliditeit schatten, gebruik makend van de nieuwe methode. Dat doen ze op een ander tijdstip en een andere plaats dan het Zorgloket. De onderzoekers van het UMCG garanderen bij de filmopname de veiligheid en de geheimhouding.
Het onderzoek staat verder geheel los van uw beoordeling bij het Zorgloket MOD en heeft dan ook **geen enkele invloed** op de uitslag van de beoordeling.

Wat wordt er van u gevraagd?

Als u mee wilt doen aan dit onderzoek wordt u schriftelijk om toestemming gevraagd om eenmalig het gesprek met u en uw verzekeringsarts te mogen filmen. Het beoordelingsgesprek vindt verder volgens de normale gang van zaken plaats. Het enige verschil is dat het gesprek gefilmd wordt. Er is alleen een kleine camera in de gespreksruimte aanwezig, geen extra persoon. Meedoen is op vrijwillige basis. U kunt zich te allen tijde terug trekken uit het onderzoek. Uw besluit om wel of niet mee te doen aan dit onderzoek heeft geen enkele invloed op uw beoordeling bij het zorgloket MOD.

Wat gebeurt er met de film?

De filmbeelden zijn volledig vertrouwelijk en worden anoniem verwerkt bij het UMCG. Alle betrokkenen bij het onderzoek hebben een geheimhoudingsplicht. De filmbeelden worden uitsluitend gebruikt voor dit onderzoeksdoel en zijn eigendom van het UMCG. De beelden mogen op geen enkele wijze gebruikt worden voor andere doeleinden en zijn niet opeisbaar. Na afloop van het onderzoek worden alle filmbeelden vernietigd.

Wilt u meedoen?

Wij hopen dat u het gebruik van een betrouwbare schattingsmethodiek bij de medische beoordeling van uw invaliditeit net zo belangrijk vindt als wij, en dat u

mee wilt werken aan ons onderzoek.
Als u besluit om mee te werken aan dit onderzoek wilt u dan, binnen twee weken, het volledig ingevulde toestemmingsformulier in bijgevoegde antwoordenvelop terugsturen? Een postzegel is niet nodig. Als u besluit om niet mee te werken aan dit onderzoek, vragen wij u vriendelijk of u het niet-deelname formulier wilt invullen en terugsturen. U wordt dan niet meer door ons benaderd voor deelname aan het onderzoek. Hartelijk dank hiervoor.

Heeft u vragen?

Wanneer u vragen heeft kunt u contact opnemen met de onderzoekers van het UMCG, drs. Andrea Fokkens en dr. Jolanda Tuinstra.
Zij staan u graag te woord en zijn bereikbaar op telefoon (050) 363 29 88 / (050) 363 90 80 of email a.s.fokkens@med.umcg.nl

Bijlage 4: Toestemmingsformulier cliënt
NB: de opmaak die hier is weergegeven is platte tekst, de cliënt ontvangt een formulier in kleur.

Toestemmingsverklaring

Ondergetekende,
Meneer/mevrouw …… geeft aan deel te nemen aan het onderzoek naar de *Betrouwbaarheid van de nieuwe schattingsmethodiek bij het vaststellen van de mate van invaliditeit op psychische gronden*.

Daarbij geldt:
• Ik ben duidelijk ingelicht over de aard van het onderzoek.
• Ik stem vrijwillig in met deelname aan dit onderzoek.
• Ik geef toestemming voor het eenmalig filmen van het beoordelingsgesprek met de verzekeringsarts van het Zorgloket MOD.
• Ik ben goed geïnformeerd dat deelname aan het onderzoek geen enkele invloed heeft op de beoordeling door het Zorgloket MOD.
• Ik ben goed geïnformeerd dat ik me te allen tijde kan terugtrekken uit het onderzoek.
• Ik weet dat de filmbeelden eigendom zijn van het UMCG.
• Ik weet dat de filmbeelden niet opvraagbaar zijn door betrokkene of gemachtigde.
• Ik ben goed geïnformeerd dat de filmbeelden uitsluitend gebruikt worden voor dit onderzoek door het UMCG en op geen enkele andere wijze gebruikt worden.
• Ik weet dat de filmbeelden na afloop van het onderzoek worden vernietigd.
• Als ik iemand meeneem naar het beoordelingsgesprek, zoals mijn partner, is dit toestemmingsformulier ook geldig voor hem of haar.

Handtekening: Datum
…… ……/……/……

Graag het volledig ingevulde formulier in bijgevoegde antwoordenvelop retourneren. Een postzegel is niet nodig. Hartelijk dank hiervoor!

Bijlage 5: Niet deelname formulier cliënt

NB: de opmaak die hier is weergegeven is platte tekst, de cliënt ontvangt een formulier in kleur.

Verklaring van niet-deelname

Ondergetekende,
Meneer/mevrouw …… geeft aan niet deel te nemen aan het onderzoek naar de *Betrouwbaarheid van de nieuwe schattingsmethodiek bij het vaststellen van de mate van invaliditeit op psychische gronden*.
Daarbij geldt:
• Ik geef geen toestemming voor het eenmalig filmen van het beoordelingsgesprek met de verzekeringsarts van het Zorgloket MOD.
Handtekening: Datum
…… ……/……/……
We zouden het erg op prijs stellen wanneer u hieronder de belangrijkste reden aangeeft voor het afzien van deelname aan dit onderzoek:
……
……
……
……
……
Graag het formulier in bijgevoegde antwoordenvelop retourneren. Een postzegel is niet nodig. Hartelijk dank hiervoor!

Bijlage 6: Formulier ter invulling VA-U voor de achtergrond informatie

Achtergrond informatie cliënt

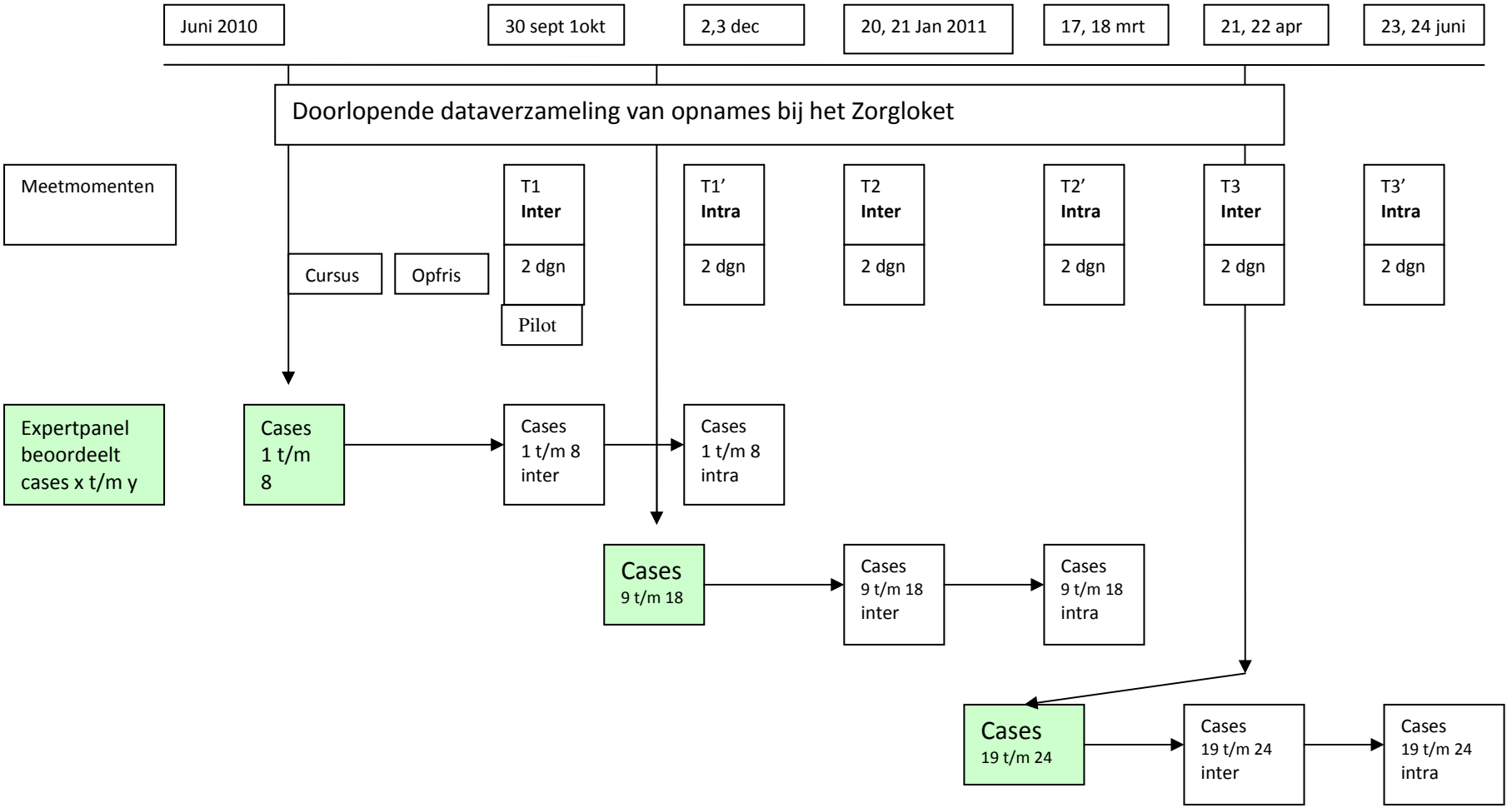
Datum: ____/_____/____ Code film: _____

Persoonsprofiel
Leeftijd: Geslacht: M / V

Dienstverloop
Data uitzendingen: Datum ontslag: Claim:

Medisch dossier
(eventuele doorgemaakte aandoeningen / behandelingen)

Bijlage 7 Opzet onderzoek



Bijlage 8
Informatie over analyse techniek

Voor de analyses in dit onderzoek is gebruik gemaakt van een aangepaste versie van het computer programma Ag09 geschreven door Popping[34]. Dit programma berekent een uitgebreide versie van de interrater agreement index Scott's π. De algemene formule voor de overeenstemmingsindex is:

$I = (P_0 - P_e) / (1 - P_e)$
P_0 = *waargenomen percentage overeenstemming*
P_e = *verwachte percentage overeenstemming*

Bij het berekenen van waargenomen en verwachte overeenstemming ziin twee uitgangspunten van belang. Ten eerste zijn de berekeningen gebaseerd op alle paren van beoordelingen van een object, waarbij elk object minimaal twee maal beoordeeld moet zijn. Ten tweede worden alle beoordelaars als gelijkwaardig beschouwd (zie ook Popping, 2009[35]). Het is mogelijk de overeenstemming per klasse te berekenen (het object valt wel of niet binnen de klasse en als er meerdere klassen zijn en deze zijn minimaal geordend dan is weging mogelijk, teneinde rekening te houden met de mate van verschillen). Als een object twee maal aan dezelfde klasse wordt toegewezen dan is er hier perfecte overeenstemming. Als de ene toegewezen klasse één hoger of lager is dan de andere toegewezen klasse, dan is er een beetje verschil tussen de toewijzingen. Maar als de ene toewijzing aan de laagst mogelijke klasse is en de andere toewijzing aan de hoogst mogelijke klasse dan is er een enorm verschil. Via weging kan rekening gehouden worden met de mate van deze verschillen. In het eerste geval wordt dan uitgegaan van een beetje overeenstemming, in het tweede geval van volstrekt geen overeenstemming. Dit moet uitgedrukt worden in de vorm van getallen en deze getallen worden dan als de gewichten aangeduid. Vormen de klassen een oplopende schaal (ordinale data) dan wordt doorgaans gebruik gemaakt van een lineaire relatie tussen de gewichten. Zijn de klassen niet alleen geordend, maar is er ook een gelijke afstand tussen alle opeenvolgende klassen (interval niveau) dan wordt een kwadratische lineaire relatie gebruikt. Overeenstemming wordt altijd berekend per kenmerk (in deze studie de subrubrieken).
Verder is een aanname dat alle klassen ongeveer in dezelfde mate gebruikt worden. Vaak is dit niet het geval[36], de verhouding in gebruik tussen klassen is dan heel scheef. Een gevolg van deze scheve verdeling is dat de verwachte overeenstemming als het ware kunstmatig heel hoog wordt. Dit heeft weer tot gevolg dat de index een hele lage uitkomst op zal leveren. Als deze situatie zich voordoet, en dat is het geval in de huidige studie, beperkt men zich vaak tot de rapportage van het waargenomen percentage (gewogen) overeenstemming. Dit mede omdat voor deze situatie nog geen andere goede maten beschikbaar zijn.

De volgende berekeningen zijn voor de huidige studie uitgevoerd in het analyse programma;
1) de overeenstemming tussen de 1e en de 2e meting binnen een VA-G, 2) de gepaarde overeenstemming tussen alle VA-G over alle films (apart voor de 1e en de 2e meting),en 3) de overeenstemming per klasse

[34] Ag09 : A Computer Program for Interrater Agreement for Judgments. R. Popping. *Social Science Computer Review* 2010 28: 391
[35] Some views on agreement to be used in content analysis studies. R. Popping Quality & Quantity, 2009, DOI 10.1007/s11135-009-9258-3.
[36] High agreement but low kappa: II. Resolving the paradoxes. Cicchetti DV, Feinstein AR. *J Clin Epidemiol*. 1990;43(6):551-8

wordt bepaald. De uitkomsten worden weergegeven in pi en in kwadratisch gewogen percentage overeenstemming, we hebben immers te maken met interval data. In het statistische programma worden 2 verschillende uitgangspunten bij de weging gekozen. Ten eerste, de weging met een vaststaand theoretisch aantal klassen, in de huidige studie 5. Of ten tweede, een weging met het daadwerkelijke aantal klassen, welke verschilt per subrubriek. Bijvoorbeeld wanneer alleen klasse 0 en 1 gescoord is, is het daadwerkelijk aantal klassen 2. Het theoretisch aantal klassen loopt van 0 tot en met 4 (is 5 klassen), en niet van 0 tot en met 5, omdat in dit onderzoek de hoogste klasse 5 nooit gescoord is door de VA-G. De keuze voor theoretisch of feitelijk beschikbare aantal klassen is een theoretische. Beide zijn uitgevoerd en gerapporteerd in afzonderlijke kolommen in de tabel.

Ter illustratie worden hieronder de matrixen weergegeven, voor de verschillende situaties:

Geen gewichten:

	0	1	2	3	4
0	1,00	0,00	0,00	0,00	0,00
1	0,00	1,00	0,00	0,00	0,00
2	0,00	0,00	1,00	0,00	0,00
3	0,00	0,00	0,00	1,00	0,00
4	0,00	0,00	0,00	0,00	1,00

Kwadratisch gewogen (5 klassen):

	0	1	2	3	4
0	1,00	0,94	0,75	0,44	0,00
1	0,94	1,00	0,94	0,75	0,44
2	0,75	0,94	1,00	0,94	0,75
3	0,44	0,75	0,94	1,00	0,94
4	0,00	0,44	0,75	0,94	1,00

Kwadratisch gewogen met daadwerkelijke klassen (bij 4 klassen)

	0	1	2	3
0	1,00	0,89	0,56	0,00
1	0,89	1,00	0,89	0,56
2	0,56	0,89	1,00	0,89
3	0,00	0,56	0,89	1,00

Kwadratisch gewogen met theoretisch aantal klassen (bij 4 klassen)

	0	1	2	3
0	1,00	0,94	0,75	0,44
1	0,94	1,00	0,94	0,75
2	0,75	0,94	1,00	0,94
3	0,44	0,75	0,94	1,00

Kwadratisch gewogen met daadwerkelijke klassen (bij 2 klassen)

	0	1
0	1,00	0,00
1	0,00	1,00

Kwadratisch gewogen met theoretisch aantal klassen (bij 2 klassen)

	0	1
0	1,00	0,94
1	0,94	1,00

Bijlage 9: Tabellen 2e meting.

De tabellen zijn oplopend genummerd, tussen haakjes en cursief staat de verwijzing naar de tabel in hoofdstuk 4.

Tabel 1 *(tabel 4.4):* Verdeling geschatte klasse over de subrubrieken, 1e meting

Subrubriek		*Normaal*	*Geringe beperking*	*Milde beperking*	*Matige beperking*	*Ernstige beperking*	*Extreme beperking*	*NG*
		0	*1*	*2*	*3*	*4*	*5*	
1	Hygiëne zelfzorg	115 (95.8)	1 (0.8)					4 (3.3)
2	Mobiliteit	96 (80.0)	21 (17.5)	2 (1.7)	1 (0.8)			
3	Slapen	21 (17.6)	30 (25.2)	55 (46.2)	8 (6.7)	5 (4.2)		
4	Seksuele functie	87 (72.5)	4 (3.3)	2 (1.7)	2 (1.7)			25 (20.8)
5	Basale communicatie	117 (98.3)	2 (1.7)					1 (0.8)
6	Sociale communicatie	44 (36.7)	42 (35.0)	28 (23.3)	5 (4.2)	2 (1.7)		
7	Communicatieve vaardigheden	45 (37.5)	47 (39.2)	21 (17.5)	4 (3.3)	3 (2.5)		
8	Structuur aanbrengen	87 (72.5)	21 (17.5)	10 (8.3)	1 (0.8)			1 (0.8)
9	Huishoudelijke activiteiten	88 (73.3)	19 (15.8)	11 (9.2)				2 (1.7)
10	Omgaan stressvolle gebeurtenissen	39 (32.5)	58 (48.3)	20 (16.7)	2 (1.7)	1 (0.8)		

N=5 artsen x 24 films=120, NG=Niet gescoord

Tabel 2 *(tabel 4.10)***:** Aantal films met aantal VA-G, N(%) met overeenkomstige gescoorde klasse per subrubriek (2e meting)

Subrubriek		*5* 100% overeenstemming	*4* 4 VA-G overeen stemmend	*3* 3 VA-G overeen Stemmend	*2* 2 VA-G overeen stemmend
1	Hygiëne zelfzorg	19 (79.2)	5 (20.8)	0	0
2	Mobiliteit	15 (62.5)	4 (16.7)	4 (16.7)	1 (4.2)
3	Slapen	5 (20.8)	9 (37.5)	9 (37.5)	1 (4.2)
4	Seksuele functie	15 (62.5)	6 (25.0)	2 (8.3)	1 (4.2)
5	Basale communicatie	18 (75.0)	6 (25.0)	0	0
6	Sociale communicatie	4 (16.7)	7 (29.2)	9 (37.5)	4 (16.7)
7	Communicatieve vaardigheden	3 (12.5)	7 (29.2)	10 (41.7)	4 (16.7)
8	Structuur aanbrengen	12 (50.0)	3 (12.5)	7 (29.2)	29 (8.3)
9	Huishoudelijke activiteiten	14 (58.3)	3 (12.5)	7 (29.2)	0
10	Omgaan stressvolle gebeurtenissen	5 (20.8)	5 (20.8)	13 (54.2)	1 (4.2)

N=24 films, 5 artsen zelfde score is 100% overeenstemming

Tabel 3 *(tabel 4.11)*: Aantal films met aantal VA-G, N(%) met overeenkomstige gescoorde per subrubriek (1e meting), 5 of 4 VA-G overeenstemming.

Subrubriek		*5 of 4* VA-G overeenstemmend*	*4**, 3 of 2 VA-G overeenstemmend*
1	Hygiëne zelfzorg	24 (100)	-
2	Mobiliteit	19 (79.2)	5 (20.8)
3	Slapen	14 (58.3)	10 (41.7)
4	Seksuele functie	20 (83.3)	4 (16.7)
5	Basale communicatie	24 (100)	-
6	Sociale communicatie	11 (45.8)	13 (54.2)
7	Communicatieve vaardigheden	10 (41.7)	14 (58.3)
8	Structuur aanbrengen	15 (62.5)	9 (37.5)
9	Huishoudelijke activiteiten	17 (70.8)	7 (29.2)
10	Omgaan stressvolle gebeurtenissen	14 (58.3)	10 (41.7)

N=24 films

* 4 artsen overeenstemmend en de 5de arts niet meer dan 1 klasse verschilt van de andere 4

** 4 artsen overeenstemmend en de 5de arts meer dan 1 klasse verschilt van de andere 4

Tabel 4 *(tabel 4.12)*: Voorkomen van de afwijkende scores van de 5e VA-G wanneer 4 VA-G overeenstemmend gescoord hebben, per subrubriek

Subrubriek		*-2*	*-1*	*0**	*+1*	*+2*	*+3*	*+4*	*NG*
1	Hygiëne zelfzorg				1				4
2	Mobiliteit		2		2				
3	Slapen		6		3				
4	Seksuele functie			1	2		1		2
5	Basale communicatie				4				2
6	Sociale communicatie		3		4				
7	Communicatieve vaardigheden		1		6				
8	Structuur aanbrengen				3				
9	Huishoudelijke activiteiten				3				
10	Omgaan stressvolle gebeurtenissen		2		3				

* vier artsen hebben “niet aan de orde gekomen” aangegeven, 1 arts 0.

Tabel 5 *(tabel 4.13)*:Percentage interdokterovereenstemming* tussen VA-G, per subrubriek

Subrubriek	*Weging: Vaststaand aantal klassen*	*Weging: Daadwerkelijk aantal klassen*	*Overeenstemming per klasse*				
			0	1	2	3	4
1	100	98	98	98	-	-	-
2	98	96	82	82	97	98	-
3	97	97	99	66	100	100	78
4	80	80	87	92	98	96	100
5	100	93	93	93	-	-	-
6	94	94	73	60	82	90	96
7	94	94	70	64	73	95	96
8	97	97	80	75	89	100	98
9	99	95	85	79	93	-	-
10	96	96	75	59	78	97	98

N=24 films en 5 VA-G per film
*Gepaard kwadratisch gewogen

Tabel 6 *(tabel 4.14)*: Percentage interdokterovereenstemming* tussen VA-G, "niet aan de orde gekomen" meegenomen als klasse 0.

Subrubriek	*Weging: Vaststaand aantal klassen*	*Weging: Daadwerkelijk aantal klassen*	*Overeenstemming per klasse*				
			0	1	2	3	4
1	100	98	98	98	-	-	-
2	98	98	82	82	97	98	-
3	97	97	91	70	70	93	95
4	97	97	88	94	97	97	100
5	100	93	93	93	-	-	-
6	94	94	73	60	82	90	-
7	94	94	70	64	73	95	96
8	97	97	80	75	89	100	98
9	99	95	86	80	93	-	-
10	96	96	75	59	78	97	98

N=24 films
*Gepaard kwadratisch gewogen

Tabel 7 *(tabel 4.15):* Verdeling van het geschatte afgerond % IP van 5 VA-G per film en de spreiding in % afgerond IP tussen de 5 VA-G per film

	Afgerond IP											
Film	*0*	*5*	*10*	*15*	*20*	*25*	*30*	*35*	*40*	*45*	*50*	*Spreiding*
1	1	2	1		1							20
2	1			1	1	2						25
3		1	3	1								10
4	2	3										5
5	2	2	1									10
6	4	1										5
7				1			1	1	1		1	35
8				1		1		1	2			25
9				1	1							5
10		1	2	1	1							15
11	5											0
12		1	3	1								10
13	5											0
14			2	1	1	1						15
15	3	2										5
16	2	2	1									10
17			2	2		1						20
18	1	1	3									10
19	1	2	2									10
20	2	3										5
21	1	2	2									10
22	2	2	1									10
23		1	4									5
24		4	1									5

5 VA-G per film, "niet aan de orde gekomen" is in de berekeningen meegenomen als klasse 0.

Tabel 8 *(tabel 4.20):* Gemiddelde scores rubrieken en invaliditeitspercentage

Rubriek		*N*	*min*	*max*	*gemiddelde*	*SD*
1	Activiteiten dagelijks leven	92	0.0	2.0	0.5	0.4
2	Sociaal functioneren	118	0.0	2.7	0.6	0.6
3	Concentratie, doorzettingsvermogen en tempo	117	0.0	3.0	0.4	0.7
4	Adaptatie stressvolle omstandigheden	120	0.0	4.0	0.9	0.8
	Totaalrubriekscore	90	0.0	2.6	0.6	0.5
	Invaliditeitspercentage	90	0.0	51.7	11.9	10.2
	Afgerond invaliditeitspercentage	90	0.0	50	9.7	10.1

N_{max}=5 artsen x 24 films=120, *de getallen zijn afgerond

Tabel 9 *(tabel 4.21)***:** Gemiddelde scores per rubriek en invaliditeitspercentage, berekend met "niet gescoord" als klasse 0

Rubriek		*N*	*min*	*max*	*gemiddelde*	*SD*
1	Activiteiten dagelijks leven	120	0.0	2.0	0.5	0.4
2	Sociaal functioneren	120	0.0	2.7	0.6	0.6
3	Concentratie, doorzettingsvermogen en tempo	120	0.0	3.0	0.4	0.6
4	Adaptatie stressvolle omstandigheden	120	0.0	4.0	0.9	0.8
	Totaalrubriekscore	120	0.0	2.6	0.6	0.5
	Invaliditeitspercentage	120	0.0	51.7	11.9	10.4
	Afgerond invaliditeitspercentage	120	0.0	50.0	9.7	10.4

N_{max}=5 artsen x 24 films=120, *de getallen zijn afgerond

Tabel 10 *(tabel 4.22)***:** Totaalrubriekscore, IP en afgerond IP berekend op twee manieren (A en B)

Rubriek	*A: 4 rubrieken*					*B: Subrubrieken even zwaar*				
	n	*min*	*max*	*gem*	*SD*	*n*	*min*	*max*	*gem*	*SD*
Totaalrubriekscore	90	0.0	2.6	0.6	0.5	90	0.0	2.1	0.5	0.4
Invaliditeitspercentage	90	0.0	51.7	11.9	10.2	90	0.0	42	10.9	8.9
Afgerond invaliditeitspercentage	90	0.0	50	9.7	10.1	90	0.0	40.0	9.0	9.0

N_{max}=5 artsen x 24 films=120

Bijlage 10 IP, hypothetische berekeningen

De uitkomstmaat IP is berekend volgens de ontwikkelde schattingsmethodiek. Elke rubriek kent een verschillend aantal subrubrieken, waardoor de bijdrage per subrubriek verschillend is. Zo bestaat rubriek 1 'Activiteiten dagelijks leven' uit vier subrubrieken, rubriek 2 'Sociaal functioneren' uit drie subrubrieken, rubriek 3 'Concentratie, doorzettingsvermogen en tempo' uit twee subrubrieken en de laatste, rubriek 4 'Adaptatie stressvolle omstandigheden' uit een subrubriek.

De rubriekscores komen als volgt tot stand:

Rubriek 1= (S1+S2+S3+S4)/4, waarbij S=subrubriek

Rubriek 2= (S5+S6+S7)/3

Rubriek 3= (S8+S9)/2

Rubriek 4= S10

Totaalscore =(R1+R2+R3+R4)/4

In tabel B10.1 wordt een totaaloverzicht gegeven van het voorkomen van het IP in stappen van 5% voor alle films samen. In dezelfde tabel wordt dit gegeven wanneer "niet aan de orde gekomen" is berekend als klasse 0 (kolom IP met NG=0).

In deze tabel wordt een overzicht weergegeven van vier hypothetische IP berekeningen. Per kolom wordt telkens 1 bepaalde subrubriek aangepast, de overige subrubrieken blijven onveranderd. Zo staan in de 4^{e} kolom de resultaten voor de volgende conditie: de subrubriek 3 'Slapen' wordt per film per VA-G verlaagd met -1 (bij klasse 0 blijft dit 0). Datzelfde gebeurt met deze subrubriek voor +1 (5^{e} kolom). Dergelijke analyses zijn ook uitgevoerd voor de subrubriek 10 'Omgaan met stressvolle gebeurtenissen' (De voorlaatste en laatste kolom van de tabel).

Tabel B10.1: Overzicht frequenties IP bij verschillende condities

5% klas-sen IP	*IP (n=120)*	*IP met NG=0*	*Slapen -1*	*Slapen + 1*	*Omgaan stressv. -1*	*Omgaan stressv. +1*
Geen berekening mogelijk door NG	37	-	37	37	37	37
0	15	28	16	14	33	
5	20	26	28	14	20	15
10	18	30	10	22	11	20
15	11	14	12	11	8	18
20	8	8	7	11	4	11
25	4	4	4	1	5	8
30	5	7	4	7	1	4
35	1	2	1	2		5
40					1	1
45	1	1	1	1		1

IP, NG=0: "niet aan de orde gekomen" berekend als 0, Slapen -1: subrubriek slapen met 1 klasse verlaagd, slapen +1: subrubriek slapen met 1 klasse verhoogd, omgang stress -1: subrubriek omgaan stressvolle gebeurtenissen met 1 klasse verlaagd, omgang stress + 1: subrubriek omgaan stressvolle gebeurtenissen met 1 klasse verhoogd.

Wat valt op:

- Wanneer “niet aan de orde gekomen” als klasse 0 wordt meegerekend, stijgt de frequentie voornamelijk bij 0% IP.

- Het effect van min 1 klasse of plus 1 klasse is voor de subrubriek 3 ‘Slapen’ kleiner dan voor de subrubriek 10 ‘Omgaan met stressvolle gebeurtenissen”. Dit komt doordat de subrubriek “Slapen” relatief minder bijdraagt aan het IP en subrubriek 10 ’Omgaan met stressvolle gebeurtenissen’ relatief meer. Dit is wel afhankelijk van de daadwerkelijke score: voor subrubriek 3 ‘slapen’ geldt $^{1}/_{16}$ x daadwerkelijke score, en voor subrubriek 10 ‘omgaan met stressvolle gebeurtenissen geldt ¼ x daadwerkelijke score.