Validiteit van de Vierdimensionale Klachtenlijst (4DKL) in de ambulante geestelijke gezondheidszorg

Berend Terluin, Marc Verbraak

SAMENVATTING

De Vierdimensionale Klachtenlijst (4DKL) is ontwikkeld in de eerstelijnsgezondheidszorg en meet vier dimensies van veelvoorkomende psychische klachten: distress, depressie, angst en somatisatie. De 4DKL zou ook voor de ambulante geestelijke gezondheidszorg (ggz) een aantrekkelijk instrument kunnen zijn voor het in kaart brengen en volgen van de klachten van patiënten. Voorwaarde is dat de 4DKL bij ambulante-ggz-patiënten hetzelfde meet als bij huisartsenpatiënten. Dat hebben wij onderzocht door 4DKL-gegevens van een groep ambulante-ggz-patiënten te vergelijken met 4DKL-gegevens van een groep huisartsenpatiënten en daarbij te kijken naar *differential item functioning* (DIF). Het bleek dat zes van de in totaal 50 items bij ambulante-ggz-patiënten iets anders functioneerden dan bij huisartsenpatiënten. Maar dat bleek op de schaalscores een verwaarloosbaar effect te hebben. De conclusie is dat de 4DKL bij ambulante-ggz-patiënten hetzelfde meet als bij huisartsenpatiënten.

Inleiding

De Vierdimensionale Klachtenlijst (4DKL) is een zelfinvulvragenlijst die ontwikkeld is in de eerstelijnsgezondheidszorg (Terluin, 1996). Geïntroduceerd in 1996, heeft de 4DKL zich inmiddels een vaste plaats in de Nederlandse huisartsenpraktijk verworven als hulpmiddel bij het opsporen, in kaart brengen en vooral bespreekbaar maken van psychische klachten (Sinnema et al., 2013). Behalve door huisartsen wordt de 4DKL ook gebruikt door bedrijfsartsen, (psychosomatische) fysiotherapeuten, psychologen en maatschappelijk werkers. Tevens is de 4DKL opgenomen in verschillende behandelrichtlijnen voor huisartsen, bedrijfsartsen en psychologen. Redenen om de 4DKL ook te gaan gebruiken in de praktijk van de ambulante geestelijke gezondheidszorg (ggz) kunnen zijn: de verbetering van de communicatie met verwijzers, de mogelijkheid om sommige vragenlijsten, zoals de Symptom Checklist (SCL-90), in de ambulante ggz (deels) te vervangen (Terluin, Neeleman-Van der Steen, Verbraak, Smitskamp, & Duijsens, 2009), en het feit dat de 4DKL gratis is voor niet-commercieel gebruik.

De 4DKL meet vier duidelijk te onderscheiden dimensies van psychopathologie: distress, depressie, angst en somatisatie, en is daarmee een van de weinige psychische klachtenlijsten die onderscheid trachten te maken tussen 'normale' (distress) en pathologische reactievormen (depressie, angst en somatisatie) (Terluin et al., 2006). De distressschaal meet de ervaring van spanning of stress en geeft aan hoe moeilijk de persoon het heeft met het hanteren van aanwezige stressoren. De depressieschaal meet specifieke symptomen van een (matige tot ernstige) depressieve stoornis, met name anhedonie (verlies van plezier) en negatieve cognities (waaronder suïcidale gedachten). De angstschaal meet specifieke symptomen van angststoornissen, met name van paniekstoornis, agorafobie, sociale fobie, obsessieve-compulsieve stoornis en posttraumatische stressstoornis. De somatisatieschaal meet de hoeveelheid lichamelijke spanningsklachten en geeft een indicatie van de (over)gevoeligheid van het lichaam voor dit type klachten. In de klinische praktijk kan de 4DKL gebruikt worden voor (1) het opsporen en bespreekbaar maken van psychosociale problemen (dat geldt uiteraard meer voor de huisartsenpraktijk dan voor de ambulante ggz, waar patiënten al voorgeselecteerd zijn op het hebben van een psychosociaal probleem), (2) het inschatten van de lijdensdruk (distress), (3) het inschatten van de kans dat iemand een depressieve stoornis of een angststoornis heeft (Terluin, Brouwers, Van Marwijk, Verhaak, & Van der Horst, 2009), en (4) het opvolgen van de klachten in de tijd (zie het klinisch vignet).

De beperking van de 4DKL is dat hij klachten meet, en dan alleen klachten die in voldoende mate voorkomen in een eerstelijnssetting en die zich laten vangen in een psychometrische schaal. De 4DKL meet dus geen niet-klachten zoals persoonlijkheidstrekken of -stoornissen en gedrag (denk aan antisociaal gedrag, verslavingen en eetstoornissen) of zeldzame klachten zoals psychotische verschijnselen. Wel is het zo dat wanneer iemand door zijn persoonlijkheid, gedrag of door zeldzame symptomen emotioneel in de problemen geraakt en daar ook daadwerkelijk onder lijdt, zich dat vertaalt in klachten – met name distress – die de 4DKL kan oppikken.

Een belangrijke voorwaarde voor het gebruik van de 4DKL in de ambulante ggz is dat de lijst in die setting valide is. De validiteit van de 4DKL is voornamelijk gebaseerd op onderzoek in de huisartsenpraktijk (Terluin et al., 2006), en aangezien ambulante-ggz-patiënten een selectie vormen van huisartsenpatiënten, is het allerminst vanzelfsprekend dat de 4DKL bij ggz-patiënten hetzelfde meet als bij huisartsenpatiënten en dat de 4DKL-scores op dezelfde manier kunnen worden geïnterpreteerd. Dat zal de communicatie met verwijzende huisartsen eerder bemoeilijken dan faciliteren. Om te bepalen of de 4DKL in de ambulante ggz hetzelfde meet als in de huisartsenpraktijk, hebben wij onderzoek gedaan naar *differential item functioning* (DIF; Dorans, Schmitt, & Bleistein, 1992; Zumbo, 1999; Michaelides, 2008) van de 4DKL.[1]

Klinisch vignet

Mevrouw Klok (37 jaar) is door haar huisarts naar een psycholoog verwezen. Sinds een maand of drie, sinds het einde van de zomer, heeft zij klachten van lusteloosheid, slecht slapen en verlies van eetlust. De huisarts had haar gevraagd of stress een rol zou kunnen spelen. Ze kon dat niet ontkennen met juist een hectische echtscheiding achter de rug. Daarop had hij haar gevraagd een klachtenlijst, de 4DKL, in te vullen. Tijdens het invullen van de lijst was mevrouw

Klok zich pas goed gaan realiseren hoeveel psychische spanningen zij eigenlijk nog had. Terug bij de huisarts kreeg ze de uitslag: Distress 23, Depressie 7, Angst 2 en Somatisatie 14. De huisarts vertelde mevrouw dat zij sterk verhoogd had gescoord op distress ("dat zijn spanningen") en depressie, en matig verhoogd op somatisatie ("dat zijn lichamelijke spanningsklachten"). Hij had gezegd dat hij aan haar distressscore kon zien dat zij het moeilijk had en aan haar somatisatiescore dat zij tamelijk gevoelig was voor lichamelijke stressreacties. Haar depressiescore zou kunnen passen bij een depressieve stoornis, een situatie, zo legde de huisarts het uit, waarin haar depressieve stemming als het ware was doorgeschoten en een eigen leven was gaan leiden. Hoewel mevrouw Klok inderdaad minder plezier ervoer in gewone activiteiten en zij regelmatig het gevoel had dat het leven niet de moeite waard was, had de huisarts het moeilijk gevonden om een duidelijke diagnose te stellen. Hij had haar aangeraden om eens met een psycholoog te gaan praten voor nader onderzoek en eventueel behandeling. Hij vermeldde de uitslag van de 4DKL in zijn verwijsbrief.

Bij de psycholoog vertelt mevrouw Klok kort haar verhaal. De psycholoog leest de verwijsbrief. De 4DKL-scores zeggen haar niets; zij heeft nog nooit van de 4DKL gehoord. Zij besluit haar gebruikelijke intake te doen. Later op de dag googelt ze op "4dkl" om te kijken wat de 4DKL eigenlijk voorstelt en klikt op een van de bovenste links, van het NHG (Nederlands Huisartsen Genootschap). Ze krijgt een pdf van de 4DKL met 50 vragen en een scoreformulier waarop de gebruikte cut-offscores per schaal staan vermeld. Op het formulier staat ook een link naar het EMGO-instituut van het VU Medisch Centrum te Amsterdam. Daar vindt ze een nascholingsartikel voor huisartsen waarin wordt uitgelegd hoe de scores van de 4DKL moeten worden geïnterpreteerd en hoe de scores met de patiënt besproken kunnen worden. Even later vindt ze op het internet nog een gratis omrekenprogramma om 4DKL-scores om te zetten in voor haar bekendere SCL-90-scores. Na het invullen van de 4DKL-scores van mevrouw Klok blijkt dat haar verwachte scores op enkele SCL-90-schalen als volgt zijn: depressie 45 (70% betrouwbaarheidsinterval: 39-52), angst 19 (16-21), agorafobie 8 (7-10), somatisatie 26 (24-28), psychoneuroticisme 186 (163-210).

Na een aantal maanden komt mevrouw Klok weer op het spreekuur van haar huisarts. De psycholoog heeft inderdaad een depressie bij haar vastgesteld, die was ontstaan in reactie op haar echtscheiding. Zij heeft haar geholpen de echtscheiding te verwerken. De huisarts vraagt hoe het nu gaat. Mevrouw Klok zegt dat het "wel goed" gaat. Het antwoord klinkt niet helemaal overtuigend. "Hoe goed dan?", vraagt de huisarts. Hij ziet dat graag geobjectiveerd en vraagt mevrouw Klok om de 4DKL nog eens in te vullen. De scores zijn nu: Distress 16, Depressie 4, Angst 0 en Somatisatie 5. De huisarts vertelt mevrouw Klok dat hij inderdaad kan zien dat het beter met haar gaat, maar eigenlijk nog niet helemaal goed. De scores voor distress en depressie zijn nog matig verhoogd. De huisarts vraagt door en mevrouw Klok vertelt dat ze zich alleen voelt zonder intieme relatie. Ze heeft haar leven wel weer op orde. Met haar kinderen gaat het goed, daar kan ze ook van genieten. Ook haar werk geeft haar plezier. Maar ze ziet zichzelf als moeder van een stel puberende kinderen en niet als een aantrekkelijke partner. Ze vreest de eenzaamheid. Dat is niet ter sprake gekomen bij de psycholoog. Ze hebben zich vooral gericht op haar echtscheiding, het werk en de kinderen. De huisarts raadt mevrouw Klok aan om opnieuw contact op te nemen met de psycholoog.

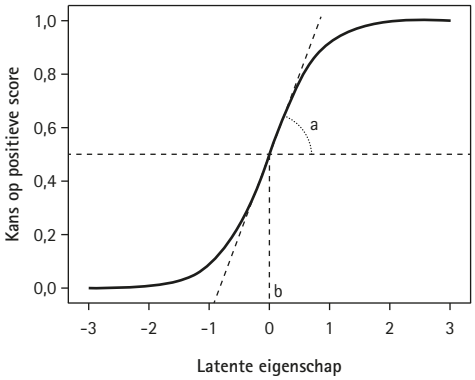
Methode

Differential item functioning (DIF)

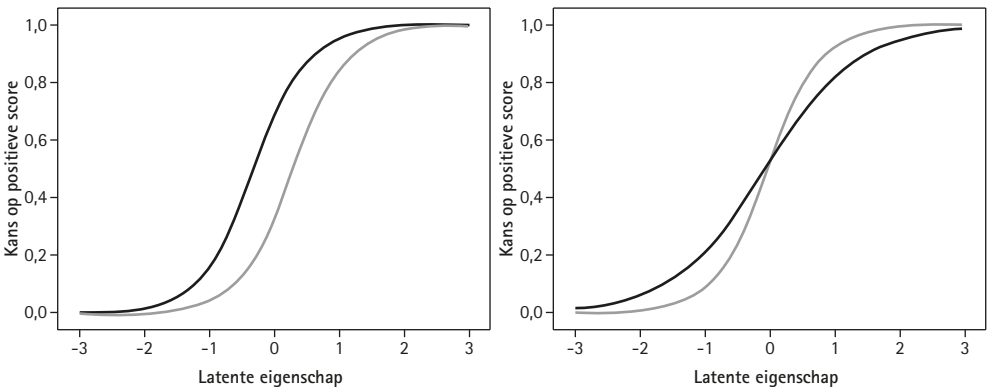
We geven eerst een korte theoretische uitleg over DIF en het onderzoek ernaar. Een multi-itemvragenlijst meet in twee groepen hetzelfde indien de betreffende items in die groepen dezelfde meeteigenschappen hebben of, anders gezegd, als de items hetzelfde 'functioneren'. Als voorbeeld nemen we een depressieschaal bestaande uit een aantal items en de depressiescore wordt berekend als de som van de itemscores. De gedachte achter zo'n schaal is dat er in de populatie een latente eigenschap depressie bestaat waarop iedere persoon een positie heeft. Sommige personen zijn weinig depressief en bevinden zich dus aan de lage kant van de latente eigenschap depressie, terwijl andere personen behoorlijk depressief zijn en aan de hoge kant zitten. De itemscores worden verondersteld een afspiegeling te zijn van die latente eigenschap en de depressiescore geeft een indruk over waar iemand zich ongeveer situeert op die latente eigenschap depressie. Die depressieschaal meet in twee groepen hetzelfde als de items in beide groepen dezelfde relaties hebben tot de onderliggende latente eigenschap. De relatie tussen een item en de latente eigenschap wordt gekarakteriseerd door de kenmerken *ernst* en *discriminerend vermogen* en wordt weergegeven met een responsfunctiecurve (RFC). Figuur 1 toont een voorbeeld van een item met twee responscategorieën (*ja*/*nee*). Getoond wordt de kans op een positieve respons (*ja*) als functie van iemands positie op de latente eigenschap. De *ernst* van een item wordt bepaald door de positie op de latente eigenschap waar de kans op een *ja* groter wordt (dus >0.5) dan de kans op een *nee*. Sommige items zijn ernstiger dan andere items. Een ernstig item, zoals bijvoorbeeld een vraag naar suïcidale gedachten, zit hoog op de latente eigenschap depressie en patiënten moeten dan ook een relatief ernstige depressie hebben voordat ze dat item scoren. Het discriminerend vermogen is het vermogen van een item om personen die hoog zitten op de latente eigenschap depressie, te onderscheiden van personen die daar laag op zitten. Het discriminerend vermogen van een item is tevens indicatief voor de item-totaal correlatie, dus voor de mate waarin het item past bij de schaal. Het discriminerend vermogen kan afgeleid worden uit de helling van de RFC. Hoe steiler de curve, hoe groter het discriminerend vermogen. Een depressieschaal wordt geacht in twee groepen hetzelfde te meten als de items van die schaal in beide groepen dezelfde eigenschappen qua ernst en discriminerend vermogen hebben in relatie tot de latente eigenschap depressie. Als die eigenschappen niet hetzelfde zijn en de items verschillend 'functioneren', wordt gesproken van *differential item functioning* (DIF; Zumbo, 1999). Een verschil in ernst wordt 'uniforme DIF' genoemd omdat de kans op het positief scoren van het item voor de ene groep over de hele latente eigenschap (uniform) groter is dan voor de andere groep. Een verschil in discriminerend vermogen wordt 'non-uniforme DIF' genoemd omdat het verschil in de kans om het item te scoren in een deel van de latente eigenschap verhoogd is terwijl het in het andere deel van de latente eigenschap verlaagd is ten opzichte van de andere groep (figuur 2).

Alle methoden voor DIF-analyse onderzoeken of de respons op een item verschilt tussen twee groepen wanneer gecorrigeerd wordt voor de latente eigenschap. Wanneer er geen sprake is van DIF, wordt de respons op een item uitsluitend bepaald door iemands positie op de latente eigenschap, ongeacht tot welke groep die persoon behoort. Om personen te kunnen matchen op hun positie op de latente eigenschap, moet DIF-analyse gebruik maken van een (meetbare) variabele die de (onmeetbare) latente eigenschap zo goed mogelijk benadert.

Die zogenaamde 'matchingvariabele' wordt doorgaans op de een of andere manier geconstrueerd uit de items van de schaal die onderzocht wordt. Hier doet zich echter het probleem voor dat als sommige items met DIF zijn behept, de daaruit geconstrueerde matchingvariabele geen onvertekende weergave is van de werkelijke latente eigenschap. Een goede matchingvariabele moet dus van DIF worden gezuiverd.



FIGUUR 1. Responsfunctiecurve (RFC) van een item met twee responscategorieën (*ja/nee*). De curve geeft de kans op een positieve respons als functie van de positie van de persoon op de latente eigenschap die door de schaal wordt gemeten. De RFC wordt gekenmerkt door het discriminerend vermogen (*a*) en de ernst (*b*) van het item. Het discriminerend vermogen kan worden afgeleid uit de helling van de RFC. De ernst wordt afgeleid uit de positie op de latente eigenschap waarbij een persoon 50% kans heeft om positief te antwoorden op het item. In dit geval heeft het item een gemiddelde ernst. De schaal van de latente eigenschap is willekeurig.



FIGUUR 2. Voorbeelden van differential item functioning (DIF). De figuren tonen de RFC's van een item bij twee groepen (groep A: zwarte curve; groep B: grijze curve). De linkerfiguur toont uniforme DIF, waarbij de kans op een positieve respons voor groep A over de hele latente eigenschap groter is dan (of gelijk is aan) die voor groep B. Het item heeft een verschillende ernst voor de groepen (voor groep B is het item ernstiger dan voor groep A). De rechterfiguur toont non-uniforme DIF, waarbij het verschil in de kans op een positieve respons tussen de groepen afhankelijk is van de positie op de latente eigenschap. Het item heeft voor beide groepen dezelfde ernst maar een verschillend discriminerend vermogen.

Design en onderzoeksgroepen

Dit onderzoek betrof een analyse van crosssectioneel verzamelde 4DKL-gegevens bij twee groepen patiënten. De eerste groep betrof alle ambulante patiënten die in februari 2006 in behandeling waren of voor behandeling waren aangemeld bij de HSK Groep, een landelijke tweedelijnsorganisatie voor ambulante hulpverlening bij psychische stoornissen. Alle patienten waren verwezen door hun huisarts of bedrijfsarts. Gedurende twee weken werd op alle deelnemende HSK-vestigingen aan alle dossiers van lopende of te starten behandelingen (intake of eerste sessie) een pakket met vragenlijsten toegevoegd. Behandelaars werden verzocht het pakket vragenlijsten aan de patiënten voor te leggen. Het pakket bevatte een brief met uitleg over het onderzoek en een toestemmingsformulier. De patiënten werden verzocht de vragenlijsten, zo mogelijk, direct na afloop van de zitting in te vullen of anders op de volgende afspraak ingevuld mee te brengen. Na invulling werden de vragenlijsten weer aan het dossier toegevoegd. Een onderzoeksassistent verzamelde alle ingevulde toestemmingsformulieren, en op een anonieme wijze de ingevulde vragenlijsten en aanvullende gegevens, waaronder leeftijd en geslacht, en de bij aanvang van de behandeling gestelde DSM-IV-diagnosen.
De tweede groep betrof alle patiënten van Gezondheidscentrum De Spil in Almere (NL) die tussen juli 2004 en december 2011 van hun huisarts een 4DKL ter invulling hadden gekregen in het kader van de gebruikelijke zorg. Voor de meeste patiënten was er sprake van (verdenking op) psychische problemen. De gegevens werden opgeslagen in een elektronisch databestand en naderhand anoniem uitgedraaid. Behalve de 4DKL-scores waren van deze groep alleen leeftijd en geslacht bekend. Door het gebruik van unieke identificatienummers konden herhaalde metingen worden onderscheiden van eerste metingen. Wij selecteerden voor dit onderzoek alleen eerste metingen om statistische onafhankelijkheid van de metingen te garanderen. We selecteerden in beide groepen patiënten van 18 tot en met 64 jaar.

Meetinstrument

De 4DKL is een vragenlijst met 50 items, die distress (16 items), depressie (6 items), angst (12 items) en somatisatie (16 items) meten. De 4DKL vraagt hoe vaak men bepaalde symptomen heeft ervaren in de afgelopen week. De items kennen 5 responsopties voor de patient, variërend van *nee* tot *heel vaak of voortdurend*, die voor het berekenen van schaalscores op een driepuntenschaal worden gescoord: *nee* = 0, *soms* = 1, *regelmatig/vaak/heel vaak of voortdurend* = 2.

Analyse

Patiënten bij wie voor een of meer 4DKL-schalen meer dan de helft van de itemscores ontbraken, werden uitgesloten. De overige ontbrekende itemscores werden geïmputeerd met behulp van de responsmethode, een methode die zowel rekening houdt met verschillen tussen personen als verschillen tussen items (Van Ginkel & Van der Ark, 2005).
Omdat DIF-analyse ervan uitgaat dat de schalen (in voldoende mate) unidimensionaal zijn (dat wil zeggen dat de scores door slechts één factor worden bepaald), hebben we de

4DKL-schalen onderzocht met 'multigroep' confirmatieve factoranalyse. We gebruikten hiervoor het softwarepakket 'lavaan' zoals geïmplementeerd in het gratis programma R (http://www.r-project.org/) (Rosseel, 2012). We evalueerden de vergelijkbaarheid in beide groepen van eenfactormodellen waarbij de itemscores als geordende categorieën werden behandeld. De mate van modelpassendheid beoordeelden we aan de hand van de volgende indexen: comparative fit index (CFI) > 0.95, Tucker-Lewis index (TLI) > 0.95 en root mean square error of approximation (RMSEA) < 0.06 (Hu & Bentler, 1999). Om adequate modelpassendheid te verkrijgen, werden residuale correlaties toegestaan tussen itemparen met een vergelijkbare inhoud.

Voor het onderzoek naar DIF hebben wij gekozen voor de Mantel-Haenszel (M-H)-methode (Michaelides, 2008). De M-H-methode gebruikt in eerste instantie de gewone somscore van de items van een schaal als matchingvariabele. Vervolgens worden de gemiddelde itemscores van de groepen vergeleken voor elk niveau van de matchingvariabele. Na standaardisatie van de somscores over de groepen wordt een gemiddeld verschil in gemiddelde itemscores (een 'standardized mean difference'; SMD) berekend. In tabel 1 wordt een vereenvoudigd voorbeeld getoond van de berekening van de SMD voor een item met twee responsopties van een hypothetische schaal met slechts twee items (waardoor de somscore varieert tussen 0 en 4). Als we kijken naar de patiënten die een schaalscore van 2 hebben, dan zien we dat groep A gemiddeld 0.80 scoort op het item, en groep B 0.50. Het verschil in gemiddelde itemscore tussen de groepen voor patiënten die een schaalscore van 2 hebben, is 0.80-0.50 = 0.30. Het verschil in gemiddelde itemscore is 0.00 voor patiënten die een schaalscore van 1 hebben, en 0.12 voor patiënten die een schaalscore van 3 hebben. Per definitie bestaat er geen verschil in gemiddelde itemscores voor patiënten die minimaal (score = 0) of maximaal (score = 4) op de schaal scoren. Immers, dan moeten de patiënten ofwel allemaal 0 ofwel allemaal 1 voor het item hebben gescoord. Om de SMD te berekenen wordt een gestandaardiseerde populatiesamenstelling berekend, in dit geval door per schaalscore het gemiddelde aantal patiënten van beide groepen te nemen.

Daarna wordt voor deze gestandaardiseerde populatie het gemiddelde verschil in itemscore tussen groep A en groep B berekend door het totale verschil in itemscore te delen door het aantal patiënten (exclusief patiënten die minimaal of maximaal op de schaal hebben gescoord). De SMD bedraagt in dit voorbeeld 0.17 en dat betekent dat patiënten van groep A gemiddeld 0.17 punten hoger scoren op het item dan patiënten van groep B. Volgens conventie wordt bij een statistisch significante SMD van 5% van de itemscorerange (in dit voorbeeld is dat 0.05 punten op een itemrange van 0 tot 1) gesproken van klinisch belangrijke DIF (Dorans, Schmitt, & Bleistein, 1992). Worden een of meer items met DIF gevonden, dan wordt het item met de meeste DIF uit de matchingvariabele verwijderd. Hierna wordt de analyse herhaald. Dat 'zuiveringsproces' wordt net zolang herhaald tot geen nieuwe items met DIF meer worden ontdekt.

Omdat de groepen niet alleen verschilden qua setting maar ook qua verhouding tussen vrouwen en mannen, hebben we de DIF-analyse herhaald voor vrouwen en mannen om te controleren of de gevonden DIF niet aan geslachtsverschillen moest worden toegeschreven.

[Tabel 1] Voorbeeld van berekening van de standardized mean difference (SMD) voor een item met twee responsopties (0 en 1). De hypothetische schaal bestaat uit slechts twee items zodat de schaalscore een range heeft van 0 tot 4. Zie de tekst voor een toelichting.

	Groep A				Groep B				Berekeningen		
	Aantallen patiënten				Aantallen patiënten						
	Itemrespons				Itemrespons						
Schaal-score	0	1	Totaal	Gemiddelde itemscore	0	1	Totaal	Gemiddelde itemscore	Verschil tussen gemiddelde itemscores	Gestandaar-diseerde N*	Totaal verschil in gemiddelde itemscores
0	6	0	6	0.00	4	0	4	0.00	0.00	5	
1	6	2	8	0.25	3	1	4	0.25	0.00	6	0.00
2	2	8	10	0.80	5	5	10	0.50	0.30	10	3.00
3	1	3	4	0.75	3	5	8	0.63	0.12	6	0.72
4	0	2	2	1.00	0	4	4	1.00	0.00	3	
Totaal	15	15	30	0.50	15	15	30	0.50	0.00	30	3.72
										22	SMD = 3.72/22 = 0.17

* Gestandaardiseerde N = gemiddelde van aantal patiënten van groep A en B.

Voor het beoordelen van de grootte van het effect van de gevonden DIF op de 4DKL-schaalscores zijn we als volgt te werk gegaan. DIF-bevattende items werden gesplitst in twee items, voor elk van de groepen een item, zodat het item voor de ambulante-ggz-groep ontbrekende waarden had voor de huisartsengroep en het item voor de huisartsengroep ontbrekende waarden had voor de ambulante-ggz-groep. Die gesplitste items werden vervolgens per schaal samen met de DIF-vrije items (die waarden bevatten voor beide groepen) onderworpen aan een Rasch-analyse. Die methode hanteert een item-respons-theoriemodel dat ervan uitgaat dat de items alleen verschillen in ernst en niet in discriminerend vermogen. De items worden echter wel getest op hun passendheid (qua discriminerend vermogen) bij het Rasch-model met behulp van zogenaamde *infit* en *outfit statistics* (Bond & Fox, 2007). Door de DIF-items te splitsen wordt de ernst van de items per groep bepaald en zodoende worden voor DIF gecorrigeerde Rasch-scores voor alle patiënten berekend. Ten slotte hebben we de gewone (dus DIF-houdende) schaalscores per groep grafisch vergeleken met de DIF-vrije Rasch-scores. Het effect van DIF (in de items) op de schaalscore was zichtbaar aan het uit elkaar gaan van de curves voor de schaalscores van beide groepen.

Voor de Rasch- en M-H-analyse gebruikten we het programma jMetrik 2.1, dat gratis op het internet verkrijgbaar is (http://www.itemanalysis.com). Voor de overige analyses gebruikten we SPSS 20 voor Windows.

Resultaten

Onderzoeksgroepen

Op zeventien HSK-vestigingen werden 2.500 pakketten met vragenlijsten uitgezet. In totaal vulden 1.074 patiënten de vragenlijsten in (respons 43%). Drie patiënten waren ouder dan 64 jaar en van één patiënt was de leeftijd onbekend. Van de resterende 1.070 werden 11 patienten uitgesloten omdat ze in een of meer 4DKL-schalen 50% of meer ontbrekende itemscores hadden, en 22 patiënten werden uitgesloten omdat de DSM-IV-diagnose ontbrak. Van de resterende 1.037 patiënten had 3,4% een of meer ontbrekende 4DKL-itemscores. In totaal ontbrak niet meer dan 0,11% van alle itemscores; die werden succesvol geïmputeerd. De ambulante-ggz-patiënten bestonden uit 507 vrouwen (49%) en 530 mannen (51%). De gemiddelde leeftijd bedroeg 42.0 jaar (*SD* = 9.5); vrouwen 39.0 jaar (*SD* = 9.5) en mannen 44.8 jaar (*SD* = 8.5). De bij de intake gestelde DSM-IV-diagnosen waren: depressieve stoornis (*major depression* en/of dysthymie) bij 23%, angststoornis(sen) bij 22% en burn-out (ongedifferentieerde stoornis met werkgerelateerde moeheid als hoofdklacht) bij 27% van de patiënten.

Het databestand van Gezondheidscentrum De Spil bevatte 1.745 4DKL-lijsten van 1.141 huisartsenpatiënten. Daarvan hadden 1.021 patiënten een leeftijd van 18 tot en met 64 jaar. Wegens het ontbreken van 50% of meer itemscores in een of meer 4DKL-schalen werden 23 patiënten uitgesloten. Van de resterende 998 huisartsenpatiënten had 21,8% een of meer ontbrekende 4DKL-itemscores. In totaal ontbrak niet meer dan 0,83% van alle itemscores; die werden succesvol geïmputeerd. De huisartsenpatiënten bestonden uit 633 vrouwen (63%) en 365 mannen (37%). De gemiddelde leeftijd bedroeg 39.2 jaar (*SD* = 12.2); vrouwen

38.5 jaar (SD = 12.3) en mannen 40.5 jaar (SD = 12.2). De huisartsenpatiënten hadden aanzienlijk hogere gemiddelde 4DKL-scores dan de ambulante-ggz-patiënten (tabel 2). Voor zover we dat mogen interpreteren als een werkelijk verschil in psychische klachten (dat moet in het vervolg van dit artikel nog blijken), was de waarschijnlijke oorzaak gelegen in het feit dat de meeste huisartsenpatiënten een nieuw psychisch probleem hadden, om welke reden de huisarts een 4DKL geïndiceerd achtte, terwijl de meeste ambulante-ggz-patiënten al enige tijd onder behandeling waren, waardoor hun klachten al waren verminderd.

[Tabel 2] Demografische kenmerken en 4DKL-scores van de onderzoeksgroepen (gemiddelden en standaarddeviaties).

		Ambulante-ggz-patiënten			Huisartsenpatiënten		
4DKL-schalen	Range	Vrouwen N = 507 M (SD)	Mannen N = 530 M (SD)	Totaal N = 1.037 M (SD)	Vrouwen N = 633 M (SD)	Mannen N = 365 M (SD)	Totaal N = 998 M (SD)
Distress	0-32	14.7 (8.7)	13.5 (8.8)	14.1 (8.8)	20.7 (8.7)	19.1 (8.9)	20.1 (8.8)
Depressie	0-12	1.9 (2.9)	2.1 (3.0)	2.0 (3.0)	4.2 (4.0)	4.2 (4.1)	4.2 (4.0)
Angst	0-24	3.8 (4.8)	3.6 (4.7)	3.7 (4.7)	7.4 (6.6)	6.4 (6.3)	7.0 (6.5)
Somatisatie	0-32	9.1 (6.4)	8.2 (6.7)	8.6 (6.6)	16.3 (7.5)	13.9 (7.3)	15.4 (7.5)

Confirmatieve factoranalyse

Multigroep confirmatieve factoranalyse liet zien dat een eenfactormodel goed paste voor de 4DKL-schalen mits correlatie van de residuale variantie van een aantal items werd geaccepteerd (tabel 3). De factorstructuur was in beide groepen hetzelfde. De items waarvan de residuale variantie correleerde, waren de distressitems 20 en 39 (slaapproblemen) en 47 en 48 (symptomen na een aangrijpende gebeurtenis), de depressie-items 33 en 46 (suïcidale gedachten) en de somatisatie-items 2, 4 en 5 (klachten van het bewegingsapparaat), 9, 12 en 13 (gastro-intestinale klachten) en 15 en 16 (thoracale klachten).

DIF-analyse

We vonden DIF in vier distress- en twee somatisatie-items (tabel 3). De items van de depressie- en angstschalen vertoonden geen DIF. Alle distressitems met DIF waren ernstiger voor ambulante-ggz-patiënten, terwijl de beide somatisatie-items met DIF juist minder ernstig waren. De ambulante-ggz-patiënten hadden dus een hogere drempel voor het scoren van de betreffende distressitems maar een lagere drempel voor de somatisatie-items in vergelijking met de huisartsenpatiënten. Het item met de meeste DIF was item 26 ('snel geïrriteerd'). De SMD van -0.23 betekende dat ambulante-ggz-patiënten gemiddeld 0,23 punten lager scoorden op het betreffende item dan huisartsenpatiënten, gecorrigeerd voor het niveau van distress. DIF op basis van geslacht troffen we alleen aan in item 41 ('snel emotioneel', SMD = -0.18), dat ernstiger bleek voor mannen. Omdat de groep ambulante-ggz-patiënten uit relatief meer mannen bestond, kan het geslachtsverschil de gevonden DIF in item 41 (deels) verklaren. Het geslachtsverschil kon de in de andere items gevonden DIF evenwel niet verklaren.

[Tabel 3] Resultaten van de multigroep confirmatieve factoranalyse.

	CFI	TLI	RMSEA	95% BI RMSEA
Distress[a]	0.995	0.994	0.059	0.055-0.063
Depressie[b]	0.999	0.999	0.052	0.038-0.067
Angst	0.995	0.994	0.047	0.042-0.053
Somatisatie[c]	0.988	0.985	0.047	0.043-0.051

CFI = comparative fit index
TLI = Tucker-Lewis index
RMSEA = root mean square error of approximation
95% BI = 95% betrouwbaarheidsinterval
[a] correlatie van residuale variantie toegestaan tussen itemparen 20-39 en 47-48.
[b] correlatie van residuale variantie toegestaan tussen itemparen 33-46 en 34-35.
[c] correlatie van residuale variantie toegestaan tussen itemparen 2-4, 2-5, 4-5, 9-12, 9-13, 12-13 en 15-16.

Het effect van DIF op de schaalscore

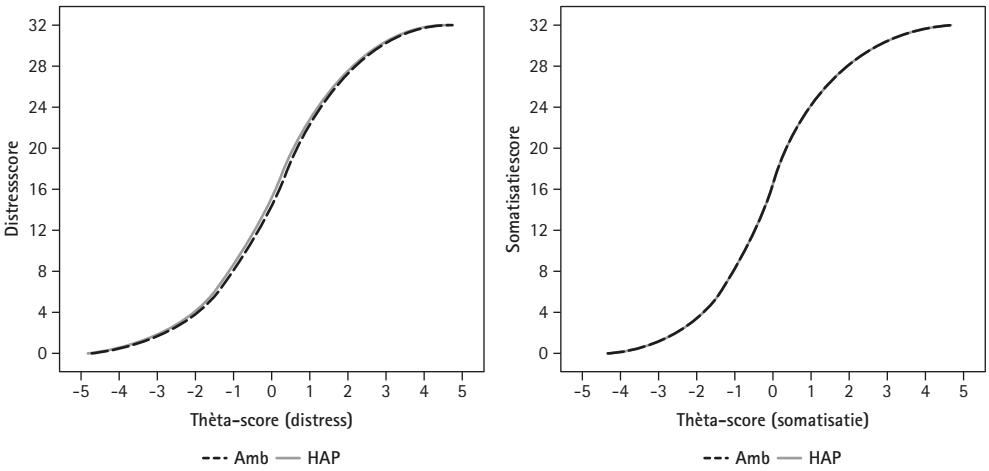
De Rasch-analyse liet zien dat alle DIF-items even goed, zo niet beter pasten in hun betreffende schalen bij de ambulante-ggz-patiënten dan bij de huisartsenpatiënten. Opmerkelijk was dat de slaapitems (20 en 39) bij de ambulante-ggz-patiënten een iets betere passing aan het Rasch-model lieten zien dan bij de huisartsenpatiënten (gegevens opvraagbaar bij de eerste auteur). De 4DKL-schalen vormen bij ambulante-ggz-patiënten dus een even goede of zelfs iets betere basis voor het inschatten van de ernst van de latente eigenschap dan bij huisartsenpatiënten.

[Tabel 4] 4DKL-items met differential item functioning (DIF).

4DKL-schaal	Item nr.	Korte omschrijving	SMD*	Opmerking**
Distress	20	onrustig slapen	–0.12	ernstiger
	26	snel geïrriteerd	–0.23	ernstiger
	39	moeite om in slaap te komen	–0.15	ernstiger
	41	snel emotioneel	–0.11	ernstiger
Somatisatie	4	pijn in de nek	+0.12	minder ernstig
	9	opgeblazen gevoel in de buik	+0.16	minder ernstig

* SMD = standardized mean difference
** ernstiger: het item is ernstiger voor ambulante-ggz-patiënten dan voor huisartsenpatiënten

Gebruik makend van de voor DIF gecorrigeerde Rasch-scores toont figuur 3 het effect van DIF op de gewone somscores van de distress- en somatisatieschalen. Bij een matige hoeveelheid distress scoorden de ambulante-ggz-patiënten net iets lager dan de huisartsenpatiënten. Het verschil op de schaalscore was echter verwaarloosbaar klein. Voor de somatisatiescore was het nauwelijks waarneembaar dat de ambulante-ggz-patiënten iets hoger scoorden dan de huisartsenpatiënten.



Amb = ambulante-ggz-patiënten; HAP = huisartsenpatiënten.

FIGUUR 3. Impact van DIF op de schaalscore. De gewone somscore voor distress (linkerfiguur) en somatisatie (rechterfiguur) als functie van de voor DIF gecorrigeerde Rasch-score (thèta).

Beschouwing

We hebben in dit onderzoek naar de meeteigenschappen van de 4DKL bij patiënten uit de ambulante ggz in vergelijking tot huisartsenpatiënten voor zes van de vijftig items kleine verschillen gevonden in het functioneren van die items binnen de schalen waartoe ze behoren. Het betrof alleen items van de distress- en somatisatieschalen. Op de schaalscores voor distress en somatisatie had dat echter praktisch geen invloed. Voor de items van de depressie- en angstschalen hebben we in het geheel geen verschillen kunnen constateren. Kortom, de 4DKL bleek bij ambulante-ggz-patiënten over dezelfde schaaleigenschappen te beschikken als bij huisartsenpatiënten.

Een mogelijke beperking van dit onderzoek betreft de selectie van de onderzoeksgroepen. Het is onwaarschijnlijk dat patiënten van een instelling als de HSK Groep in alle opzichten representatief zijn voor alle ambulante-ggz-patiënten en dat de patiënten van een gezondheidscentrum in Almere representatief zijn voor alle huisartsenpatiënten. De vraag is echter hoe waarschijnlijk het is dat HSK-patiënten hun psychische klachten op een wezenlijk andere manier beleven en verwoorden dan patiënten van andere ambulante-ggz-instellingen en hoe waarschijnlijk het is dat Almeerse huisartsenpatiënten hun psychische klachten anders beleven en verwoorden dan elders in Nederland wonende huisartsenpatienten. Wij achten de kans daarop heel klein. Voorts is werkgerelateerde problematiek in de HSK-populatie oververtegenwoordigd (wat zich uit in een oververtegenwoordiging van mannen) maar de verdeling van de DSM-IV-diagnosen in de onderzoeksgroep illustreert dat het hier wel degelijk om ambulante-ggz-patiënten ging. Almere is een grote stad maar de bewoners hebben dezelfde psychische problemen en ondervinden dezelfde klachten als overal elders. Een zwak punt van dit onderzoek, althans een risico, was dat de man-vrouw-

verhouding verschilde in de twee groepen. Het is mogelijk dat een kleine hoeveelheid DIF, die we in de distressschaal hebben gevonden, moet worden toegeschreven aan het verschil in die verhouding tussen vrouwen en mannen. Maar dat zou onze eindconclusie niet kunnen veranderen. Een sterk punt van dit onderzoek was de omvang van de onderzoeksgroepen waardoor de kans dat toevallige factoren de resultaten hebben beïnvloed, gering moet worden geacht.

Besluit

De 4DKL meet bij ambulante-ggz-patiënten hetzelfde als bij huisartsenpatiënten. 4DKL-scores van ambulante-ggz-patiënten kunnen een-op-een worden vergeleken met 4DKL-scores bij huisartsenpatiënten.

Noot

1. Praktische informatie met betrekking tot het gebruik van de 4DKL
   - de vragenlijst (vragenlijst en scoreformulier) is voor niet-commerciële doeleinden gratis te verkrijgen op: https://www.nhg.org/sites/default/files/content/nhg_org/uploads/standaard/download/4dkl.pdf
   - nascholingsartikel voor huisartsen over de interpretatie van scores en bespreking met de patienten [CB1] op: http://www.emgo.nl/quality-of-our-research/research-tools/4dsq
   - omrekenprogramma om 4DKL-scores om te zetten in SCL-scores: http://www.datec.nl/4dkl/SCL.htm.

Literatuur

Bond, T.G., & Fox, C.M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences. Second edition.* New York: Routledge.

Dorans, N.J., Schmitt, A.P., & Bleistein, C.A. (1992). The standardization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement, 29*, 309-319.

Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Michaelides, M.P. (2008). An illustration of a Mantel-Haenszel procedure to flag misbehaving common items in test equating. *Practical Assessment, Research & Evaluation, 13*, 7.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*, 2.

Sinnema, H., Franx, G., Spijker, J., Ruiter, M., Van Haastrecht, H., Verhaak, P., & Nuyen, J. (2013). Delivering stepped care for depression in general practice: Results of a survey amongst general practitioners in the Netherlands. *European Journal of General Practice, 19*, 221-229.

Terluin, B. (1996). De Vierdimensionale Klachtenlijst (4DKL). Een vragenlijst voor het meten van distress, depressie, angst en somatisatie. *Huisarts en Wetenschap, 39*, 538-547.

Terluin, B., Brouwers, E.P.M., Van Marwijk, H.W.J., Verhaak, P.F.M., & Van der Horst, H.E. (2009). Detecting depressive and anxiety disorders in distressed patients in primary care; comparative diagnostic accuracy of the Four-Dimensional Symptom Questionnaire (4DSQ) and the Hospital Anxiety and Depression Scale (HADS). *BMC Family Practice, 10*, 58.

Terluin, B., Neeleman-Van der Steen, C.W.M., Verbraak, M.J.P.M., Smitskamp, J.F., & Duijsens, I.J. (2009). Kunnen SCL-90-scores worden voorspeld op basis van 4DKL-scores? Vergelijking van de Vierdimensionale Klachtenlijst (4DKL) en de Symptom Checklist (SCL-90). *De Psycholoog, 44*, 498-507.

Terluin, B., Van Marwijk, H.W.J., Adèr, H.J., De Vet, H.C.W., Penninx, B.W.J.H., Hermens, M.L.M., ... Stalman, W.A.B. (2006). The Four-Dimensional Symptom Questionnaire (4DSQ): A validation study of a multidimensional self-report questionnaire to assess distress, depression, anxiety and somatization. *BMC Psychiatry, 6*, 34.

Van Ginkel, J.R., & Van der Ark, L.A. (2005). SPSS syntax for missing value imputation in test and questionnaire data. *Applied Psychological Measurement, 29*, 152-153.

Zumbo, B.D. (1999). *A Handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores.* Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Summary

The Four-Dimensional Symptom Questionnaire (4DSQ), which measures distress, depression, anxiety and somatization, has been developed in primary care. In a mental health care outclinic setting the 4DSQ might also be a valuable instrument for the assessment and monitoring of patients' symptoms. Prerequisite is that the 4DSQ measures the same constructs in mental health care outpatients as in primary care patients. 4DSQ-data were collected in mental health care outpatients (n = 1037) and compared with data of primary care patients (n = 998) using differential item functioning (DIF) analysis. DIF was detected in six items. However, the impact of DIF on the scale scores was negligible. In conclusion, the 4DSQ was found to measure mental health care outpatients and in primary care patients similarly.

Personalia

Dr. Berend Terluin is huisarts en senioronderzoeker aan de Afdeling Huisartsgeneeskunde en Ouderengeneeskunde van het EMGO+ Institute for Health and Care Research, VU Medisch Centrum, Amsterdam.
E-mail: b.terluin@vumc.nl.

Prof. dr. Marc Verbraak is klinisch psycholoog en bijzonder hoogleraar gezondheidszorgpsychologie aan het Behavioural Science Institute, Radboud Universiteit Nijmegen. Hij is tevens hoofdopleider gezondheidszorgpsychologen SPON bij de Radboud Universiteit en inhoudelijk directeur bij de HSK Groep te Arnhem.
E-mail: m.verbraak@hsk.nl.