Test-hertest betrouwbaarheid van de OECD vragenlijst voor lichamelijke beperkingen

H.C. Boshuizen, A.M.J. Chorus, D.J.H. Deeg*

In 1981 werd door de OECD een vragenlijst gepubliceerd voor het vaststellen van prevalenties van langdurige lichamelijke beperkingen via enquêtes in de open bevolking. In het hier beschreven onderzoek werd de test-hertest betrouwbaarheid van deze vragenlijst onderzocht bij ouderen (55-85 jaar, n=356). Daartoe ontvingen zij per post driemaal een vragenlijst, de tweede lijst zes weken en de derde lijst achttien weken na de eerste vragenlijst. Van hen vulden 216 personen (61%) de lijst zowel de eerste als de tweede maal in en 231 personen (65%) deden dat de eerste en de derde maal. Het antwoord op de vraag of men met één ander persoon een gesprek kan voeren was slecht reproduceerbaar (ongewogen kappa 0,09, gewogen kappa 0,22). De ongewogen kappa's van de overige vragen varieerden van 0,33 tot 0,72, de gewogen kappa's van 0,38 tot 0,90. De correlatiecoëfficiënten voor de totale somscore op de hele vragenlijst tussen de verschillende responsmomenten waren zeer hoog (0,92 en 0,94). Voor de somscores 'aantal handelingen waarmee men minimaal grote moeite had' en 'aantal handelingen dat niet uitgevoerd kan worden' waren deze correlatiecoëfficiënten iets lager, maar wel boven de 0,85. In de subgroep respondenten die in de maand voor het tweede responsmoment geen contact met een arts hadden gehad en voor wie er daarom geen aanwijzingen bestonden dat de gezondheid veranderd was, waren de kappa's en correlatiecoëfficiënten niet merkbaar hoger dan in de totale groep respondenten. De test-hertest betrouwbaarheid (afgemeten aan waarden van de gewogen en ongewogen kappa) was tussen de eerste en derde afname niet duidelijk anders dan tussen de eerste en tweede afname. De kappa's in dit onderzoek waren in de meeste gevallen duidelijk hoger dan in eerder onderzoek van het CBS. De mate van reproduceerbaarheid van de antwoorden op de afzonderlijke vragen blijkt voor de meeste vragen bevredigend; de reproduceerbaarheid van de somscores is uitstekend.

Trefwoorden: Test-hertest betrouwbaarheid, langdurige beperkingen, OECD-indicator, Cohen's kappaInleiding

In 1981 werd door een werkgroep van de OESO (Organisatie voor Economische Samenwerking en Ontwikkeling, ook bekend onder de Engelse afkorting OECD) een vragenlijst voor langdurige lichamelijke beperkingen gepubliceerd[1], bedoeld voor afname in enquêtes in de algemene populatie. Deze vragenlijst richt zich op aspecten van Algemene Dagelijkse Levensverrichtingen (ADL, zoals zelfredzaamheid in zich kunnen aan- en uitkleden, in en uit bed komen etc.), mobiliteit en communicatie. Door deze vragenlijst in verschillende landen te gebruiken, zouden internationale vergelijkingen van prevalenties van langdurige lichamelijke beperkingen mogelijk worden.

Als een van de weinige landen wordt deze vragenlijst in Nederland continue afgenomen door het CBS in haar gezondheidsenquête. Daarnaast wordt de vragenlijst, in enkele varianten, ook in ander Nederlands onderzoek regelmatig gebruikt, onder andere in de patiëntenquête van de Nationale Studie van het NIVEL[2] en de longitudinale studie naar sociaal economische gezondheidsverschillen.[3] In het ontwikkelingsproces van deze vragenlijst zijn enkele onderzoeken gedaan naar de betrouwbaarheid en validiteit. Door het CBS zijn enkele onderzoeken naar aspecten van de meetkwaliteit van deze vragenlijst gepubliceerd.[4-7] Validiteitsgegevens zijn echter nog schaars.[8] Over de test-hertest betrouwbaarheid van deze vragenlijst is slechts tweemaal gepubliceerd. Eenmaal is gepubliceerd over onderzoek bij een groep personen die eerder had deelgenomen aan een pretest voor de Amerikaanse volkstelling van 1980.[9] Dit onderzoek vertoont echter enkele tekortkomingen: het betrof slechts 11 van de 16 vragen uit de OECD-vragenlijst. Daarnaast verschilde de antwoordformulering bij beide gelegenheden: de eerste maal was sprake van een driedeling ('zonder moeite', 'met moeite' en 'nee, kan niet'); de tweede maal van een tweedeling ('heeft geen moeite', 'heeft wel

* *H.C. Boshuizen[1], A.M.J. Chorus[1], D.J.H. Deeg[2]*
[1] TNO Preventie en Gezondheid, Divisie Volksgezondheid, Leiden
[2] Vakgroep Psychiatrie, vakgroep Sociologie en Sociale Gerontologie, EMGO instituut, Vrije Universiteit, Amsterdam

moeite'). Voor het berekenen van het percentage overeenstemming werden de antwoordcategorieën 'met moeite' en 'nee' samengevoegd. Het percentage overeenstemming werd op geheel eigen wijze gedefinieerd als het percentage respondenten dat tweemaal een beperking aangaf binnen de groep respondenten die òf de eerste maal, òf de tweede maal, òf beide keren een beperking aangaf.

Door het CBS[4,5] werden 189 respondenten tweemaal geïnterviewd met gemiddeld 4,5 maand tussenpauze. Hier waren de antwoordcategorieën: 'ja, zonder moeite', 'ja, met enige moeite', 'ja, met grote moeite', 'neen, dat kan ik niet'. Naast het percentage overeenstemming (hier zoals gebruikelijk gedefinieerd als het totaal aantal overeenstemmende antwoorden op het totaal aantal personen), werd Cohen's kappa voor overeenstemming berekend. De kappa's varieerden van 0,01 tot 0,57. Volgens de auteur bleek bij nadere bestudering dat het aannemelijk was dat deze lage kappa's veroorzaakt werden doordat de beperkingen in de tijd fluctueerden. Van de respondenten met wisselende antwoorden bleek 80% te lijden aan een chronische aandoening met een wisselend karakter.

Andere gegevens over de test-hertest betrouwbaarheid zijn ons niet bekend. Inmiddels is de formulering van de vragen door het CBS enigszins aangepast. Ook wordt de vragenlijst nu schriftelijk in plaats van mondeling afgenomen. De manier van afnemen heeft invloed op de resultaten, in de zin dat bij schriftelijke afname het percentage personen dat 'ja, zonder moeite' antwoordt lager is.[5-6] Daarom is besloten een nieuw test-hertest onderzoek te doen, nu met de nieuwe formuleringen, bij schriftelijke afname en met ongeveer zes weken tussentijd. Omdat deze vragenlijst in een doorsneepopulatie lage prevalenties oplevert, is deze beter geschikt voor gebruik in oudere populaties.[8] Daarom vond deze studie plaats in een oudere populatie. Daarbij werd zowel naar de test-hertest betrouwbaarheid gekeken in de gehele onderzoeksgroep, als in een subgroep waarvoor verondersteld wordt dat de gezondheidstoestand stabiel was. Hiertoe werden de personen die aangaven dat zij in de maand voorafgaande aan de tweede meting een arts hadden bezocht uit de analyse gelaten. Tevens werden test-hertest gegevens met 4,5 maand tussentijd vergeleken met de eerdere publicatie van het CBS.

Methode

Een steekproef van ouderen in Sassenheim in de leeftijd van 55-89 jaar die in het kader van de *Longitudinal Aging Study Amsterdam* (LASA) hebben meegewerkt aan een pilot-onderzoek[10] (n=359), is met een tussenpauze van eerst zes weken en daarna nog eens twaalf weken (achttien weken na de eerste vragenlijst) benaderd met een vragenlijst per post. Naast de OECD vragenlijst bevatte deze onder andere ook de vraag of men in de afgelopen maand een arts heeft geraadpleegd. De oorspronkelijke steekproef voor de LASA-pilotstudie is gebaseerd op het bevolkingsregister van de bevolking van 55-89 jaar (dus inclusief bewoners van tehuizen).

De eerste vragenlijst werd geretourneerd door 248 ouderen, 220 retourneerden de tweede lijst en 236 de derde lijst. Hierbij beantwoordden 216 respondenten minimaal één vraag uit de OECD vragenlijst in zowel de eerste als de tweede vragenlijst en 231 respondenten deden dat zowel in de eerste als de derde vragenlijst. Daarbij beantwoordden 179 respondenten alle 16 vragen uit de OECD vragenlijst in zowel de eerste als de tweede vragenlijst en 165 respondenten deden dat in zowel de eerste als de derde vragenlijst. In *tabel* 1 worden de 216 respondenten die de eerste en de tweede vragenlijst invulden vergeleken met de deelnemers van het pilot-onderzoek die eenmaal of beide malen niet respondeerden. De respondenten zijn gemiddeld jonger, hoger opgeleid en hebben minder beperkingen dan de oorspronkelijke onderzoeksgroep.

Per vraag is Cohen's kappa berekend. Cohen's kappa is een maat voor overeenstemming waarin is gecorrigeerd voor de overeenstemming die ook op basis van toeval zou kunnen zijn ontstaan. Omdat het hier gaat om metingen op een ordinale schaal, is ook een gewogen kappa berekend.[11] In een gewogen kappa tellen afwijking tussen twee naastgelegen antwoordcategorieën minder zwaar dan een afwijking met meer verschil. Als gewicht werd hier het kwadraat van het verschil in antwoordcategorieën gebruikt (dus gewicht 1 bij twee antwoorden in naastliggende categorieën, gewicht 4 bij twee antwoorden met één categorie ertussen, en gewicht 9 bij de combinatie 'ja, zonder moeite' en 'nee, dat kan ik niet'). Dit is conform algemene aanbevelingen.[12] Voor zowel gewogen als ongewogen kappa is een 95% betrouwbaarheidsinterval berekend gebaseerd op 1,96 maal de standaardfout voor het geval dat kappa ongelijk aan nul is.

Tabel 1 *Kenmerken van degenen die in zowel de eerste als tweede vragenlijst minimaal één vraag uit de OECD-vragenlijst beantwoordden (respondenten) in vergelijking met de deelnemers uit de LASA pilotstudie die dit niet deden (non-respondenten)*

	Respondenten (n=216)	Non-respondenten (n=143)
Zelfstandig wonend**	94%	83%
Leeftijd (gemiddelde (SD)**	69,4 (9,1)	73,9 (10,4)
% vrouwen	49%	55%
Laag opgeleid (alleen LO)*	52%	65%
Kan zonder moeite buitenshuis 5 minuten aan een stuk lopen zonder stil te staan**	89%	72%
Kan met bril en/of contactlenzen zonder moeite de gewone, kleine letters in de krant lezen (n=106)[a]	82%	76%
Kan evt. met hoortoestel zonder moeite een gesprek volgen in een groep van drie of vier personen (n=107)[a]	70%	62%

Verschil tussen respondenten – non-respondenten: * = p <0,05; ** = p<0,005

[a] De vragen over horen en zien zijn slechts aan een derde van de pilotpopulatie gesteld.

Tabel 2 *Verdeling antwoordcombinaties (in procenten; N=3254) bij herhaalde afname van de OECD vragenlijst voor lichamelijke beperkingen, over alle paren gezamenlijk*[a]

Eerste meting / Herhaalde meting	ja, zonder moeite	ja, met enige moeite	ja, met veel moeite	nee, dat kan ik niet	
ja, zonder moeite	74,0	3,4	0,3	0,4	78,2
ja, met enige moeite	4,5	6,3	1,2	0,4	12,4
ja, met veel moeite	0,3	1,0	1,0	0,7	3,0
nee, dat kan ik niet	0,2	0,5	0,6	5,1	6,4
	79,1	11,2	3,1	6,6	100,0

[a] Gebaseerd op 216 respondenten die samen 3254 vragen (maximaal 16 per respondent) tweemaal beantwoordden.

Voor een aantal vragen antwoordt de overgrote meerderheid van de respondenten tweemaal 'ja, zonder moeite'. Wanneer bijvoorbeeld de antwoorden van 97% van de personen in één antwoordcategorie vallen, betekent dit dat voor deze antwoordcategorie al een percentage overeenstemming van $0{,}97^2 = 94\%$ wordt verwacht op grond van toeval. Deze 94% 'toevalsovereenstemming' telt niet mee bij de bepaling van kappa. Dit betekent dat de nauwkeurigheid waarmee kappa gemeten wordt als het ware gebaseerd is op minder dan 6% van de populatie, en daarom laag is. Bovendien is in deze gevallen de celvulling gering voor de overige combinaties van antwoorden (anders dan tweemaal het meest voorkomende antwoord). De hier berekende betrouwbaarheidsintervallen zijn gebaseerd op asymptotische benaderingen, dat wil zeggen dat zij alleen gelden wanneer het aantal beschikbare gegevens groot genoeg is. Wat 'groot genoeg' is, is echter uit de ons bekende literatuur niet te halen. Er zijn wel enkele simulatiestudies verricht met kappa's voor twee-bij-twee tabellen[13-16]. Deze laten zien dat vooral bij zeer ongelijke verdeling over de twee categorieën (prevalentie vlak bij 0% of 100%) de betrouwbaarheidsintervallen te nauw worden geschat. Om al te onbetrouwbare berekeningen uit te sluiten is er hier, arbitrair, voor gekozen om geen kappa's te berekenen als er minder dan zes personen zijn die iets anders antwoorden dan tweemaal 'ja, zonder moeite'.

Fleiss[11] karakteriseert kappa's groter of gelijk aan 0,75 als wijzend op uitstekende overeenkomst; kappa's beneden 0,40 als slechte overeenkomst en tussenliggende waarden als redelijk tot goede overeenkomst. Voor gewogen kappa's wordt dezelfde interpretatie gehanteerd.[11]

Ter vergelijking met de resultaten van Wilson en McNeil[9] is ook een percentage overeenstemmende antwoorden berekend volgens hun definitie (= het percentage overeenstemmende antwoorden binnen de groep personen die minimaal op één van de meettijdstippen een beperking aangeeft). Daarbij werden de drie hoogste antwoordcategorieën ('met enige moeite' tot 'kan ik niet') samengevoegd.

Voor de OECD vragenlijst werden verschillende somscores berekend, conform methoden die in verschillende onderzoeken gangbaar zijn. In de eerste somscore (de 'totaalscore') is aan ieder antwoord 1 (voor 'ja, zonder moeite') tot 4 (voor 'neen, dat kan ik niet') punten toegekend en zijn deze opgeteld over alle vragen. Deze somscore heeft dus een range van 16 (geen enkele moeite met alle handelingen) tot 64 punten (kan geen enkele handeling verrichten). Een tweede somscore ontstaat door het tellen van het aantal maal dat 'ja, met grote moeite' of 'nee, dat kan ik niet' werd geantwoord (deze score noemen we verder 'aantal handelingen waar men minimaal grote moeite mee heeft'). De derde somscore ('aantal handelingen dat niet uitgevoerd kan worden') ontstaat door het tellen van het aantal maal dat 'nee, dat kan ik niet' werd geantwoord. Deze laatste twee somscores hebben een range van 0 tot 16. De somscores zijn alleen berekend voor de respondenten die alle 16 vragen hebben beantwoord. Tussen de somscores bij eerste afname van de vragenlijst enerzijds en bij tweede, respectievelijk derde afname anderzijds zijn correlatiecoëfficiënten berekend. Een 95% betrouwbaarheidsinterval van de correlatiecoëfficiënten werd berekend met behulp van Fisher's z-transformatie.[17]

Resultaten

Tabel 2 geeft de verdeling van de combinatie van antwoordcategorieën tussen de eerste afname van de vragenlijst en afname zes weken later, waarbij alle 16 vragen samengenomen zijn. Het percentage overeenstemmende antwoorden (van het totaal aantal antwoorden) bedraagt 86,4%, de ongewogen kappa 0,63, de gewogen kappa 0,83. Het percentage overeenstemmende antwoorden zoals gedefinieerd door Wilson en McNeil[9] bedraagt 64,6%.

Tabel 3 geeft de kappa's en gewogen kappa's per afzonderlijke vraag voor de zes weken test-hertest betrouwbaarheid, zowel voor de totale steekproef als voor alleen de personen die in de maand voor de tweede meting geen arts bezocht hebben. In deze tabel wordt ook het aantal respondenten gegeven dat de eerste maal antwoordt 'ja, zonder enige moeite', benevens het percentage missende antwoorden.

Bij bijna alle vragen (uitzondering is de vraag of men 100 meter hard kan lopen) blijken de meeste personen, vaak meer dan 90%, in te vullen dat zij de handeling zonder enige moeite kunnen uitvoeren. Daardoor neemt de onnauwkeu-

Tabel 3 *Percentage 'ja, zonder moeite', percentage item-nonrespons, kappa en gewogen kappa [met 95% betrouwbaarheidsinterval] bij herhaalde meting na 6 weken voor de totale populatie, en voor respondenten zonder artsbezoek*[a]

	alle respondenten (n=216)				Respondenten zonder artsbezoek (n=129)		
	prevalentie 'ja, zonder moeite'	Percentage missende antwoorden	Kappa	gewogen kappa	prevalentie 'ja zonder moeite'	kappa	gewogen kappa
Kunt u een gesprek voeren met één andere persoon? (zo nodig met hoorapparaat)	91,0	4,2	0,09 [-0,05 - 0,23]	0,22 [0,01 - 0,43]	92,4	0,27 [0,05 - 0,50]	0,39 [0,11 - 0,67]
Kunt u een gesprek volgen in een groep van drie of meer personen? (zo nodig met hoorapparaat)	71,5	4,6	0,55 [0,44 - 0,66]	0,68 [0,55 - 0,81]	77,1[	0,51 [0,35 - 0,67]	0,70 [0,51 - 0,89]
Kunt u normaal, verstaanbaar praten?	96,5	5,1	0,45 [0,15 - 0,75]	0,59 [0,25 - 0,93]	99,1	-	-
Kunt u op een afstand van 4 meter het gezicht van iemand herkennen? (zonodig met bril of contactlenzen)	92,0	4,2	0,48 [0,26 - 0,69]	0,38 [0,13 - 0,63]	95,7	-	-
Zijn uw ogen goed genoeg om de gewone kleine letters in de krant te kunnen lezen (zo nodig met bril of contactlenzen)	79,1	4,2	0,42 [0,28-0,56]	0,57 [0,39 - 0,75]	83,9	0,42 [0,21 - 0,64]	0,52 [0,23 - 0,82]
Kunt u hard voedsel bijten en kauwen, zoals bijvoorbeeld een harde appel?	68,4	1,9	0,53 [0,42 - 0,64]	0,73 [0,62 - 0,85]	74,6	0,48 [0,33 - 0,62]	0,69 [0,52 - 0,86]
Kunt u zelf uw eten snijden, zoals bijvoorbeeld vlees?	96,1	2,3	0,64 [0,42 - 0,86]	0,86 [0,73 - 0,99]	97,5	-	-
Kunt u de nagels van uw tenen knippen?	63,2	2,3	0,72 [0,64 - 0,81]	0,90 [0,84 - 0,95]	69,2	0,78 [0,67 - 0,89]	0,92 [0,86 - 0,98]
Kunt u, als u staat, bukken en iets van de grond oppakken?	74,0	2,3	0,69 [0,60 - 0,79]	0,82 [0,72 - 0,91]	76,9	0,71 [0,59 - 0,83]	0,85 [0,76 - 0,93]
Kunt u een voorwerp van 5 kilo, bijvoorbeeld een volle boodschappentas, 10 meter dragen?	69,8	2,3	0,54 [0,43 - 0,64]	0,75 [0,65 - 0,85]	72,1	0,51 [0,38 - 0,65]	0,65 [0,48 - 0,82]
Kunt u zich zelf aan- en uitkleden?	95,1	1,9	0,36 [0,14 - 0,58]	0,63 [0,40 - 0,85]	97,5	-	-
Kunt u zelf in en uit bed stappen?	97,1	1,9	0,56 [0,29 - 0,83]	0,83 [0,64 - 1,0]	98,4	-	-
Kunt u zich op dezelfde verdieping van de ene naar de andere kamer verplaatsen?	97,1	1,9	0,63 [0,36 - 0,91]	0,80 [0,56 - 1,0]	98,3	-	-
Kunt u een trap van 15 treden op- en aflopen zonder stil te moeten staan?	76,2	1,9	0,63 [0,43 - 0,67]	0,76 [0,64 - 0,88]	82,8	0,57 [0,39 - 0,75]	0,70 [0,48 - 0,92]
Kunt u 400 meter aan een stuk lopen, zonder stil te staan? zo nodig met stok)	76,6	1,9	0,63 [0,52 - 0,74]	0,85 [0,79 - 0,92]	79,5	0,53 [0,38 - 0,69]	0,75 [0,61 - 0,89]
Zou u 100 meter hard kunnen lopen?	22,4	1,9	0,61 [0,53 - 0,69]	0,88 [0,84 - 0,92]	27,0	0,60 [0,49 - 0,71]	0,87 [0,82 - 0,92]

- = Kappa's niet berekend omdat het aantal personen dat niet tweemaal 'ja, zonder moeite' antwoordt te klein is.
[a] respondenten die in de 4 weken voor de tweede afname van de vragenlijst geen contact met een arts hebben gehad.

righeid van de gevonden kappa toe, hetgeen te zien is aan de soms erg brede betrouwbaarheidsintervallen.

Kijken we naar de ongewogen kappa's, dan blijken de antwoorden op de meeste vragen bij tweede afname een redelijk tot goede overeenstemming te vertonen met de oorspronkelijke antwoorden. Voor de vragen over 'een gesprek kunnen voeren met één ander persoon' en 'zichzelf kunnen aan- en uitkleden' is de overeenstemming echter slecht; voor de vraag naar 'een gesprek kunnen voeren met één andere persoon' is de overeenstemming zelfs zo slecht dat deze niet duidelijk beter is dan op grond van toeval zou worden verwacht. Wanneer naar de gewogen kappa's wordt gekeken, is het beeld aanzienlijk gunstiger: nu is de overeenstemming in ruim de helft van de gevallen zelfs uitstekend, slechts voor de vraag over 'een gesprek kunnen voeren met één andere persoon' en 'een gezicht op vier meter afstand herkennen' blijft of wordt de overeenkomst slecht, al is deze voor de eerst genoemde vraag nu wel groter dan verwacht op grond van toeval.

Wanneer alleen naar personen wordt gekeken die in de laatste vier weken voor het moment van de herhaalde meting geen contact met een arts hebben gehad (een groep waarin een plotselinge verslechtering van de gezondheid in de meetperiode minder waarschijnlijk is), verandert het

Tabel 4 *Percentage 'ja, zonder moeite', kappa en gewogen kappa [met 95% betrouwbaarheidsinterval] bij herhaalde meting na 18 weken voor dit onderzoek en eerder onderzoek van het CBS*[4]

	deze studie (n=231)			CBS 1983 (n=189)	
	prevalentie geen moeite	kappa	gewogen kappa	prevalentie geen moeite	kappa
Kunt u een gesprek voeren met één andere persoon? (zo nodig met hoorapparaat)	92,3	0,27 [0,08 - 0,47]	0,41 [0,16 - 0,66]	98,0	0,33
Kunt u een gesprek volgen[1] in een groep van drie of meer personen? (zo nodig met hoorapparaat)	71,6	0,55 [0,43 - 0,66	0,66 [0,52 - 0,81]	89,3	0,31
Kunt u normaal, verstaanbaar praten?	95,9	0,61 [0,33 - 0,90]	0,63 [0,28 - 0,98]	99,3	---
Kunt u op een afstand van 4 meter het gezicht van iemand herkennen? (zonodig met bril of contactlenzen)[2]	92,2	0,33 [0,15 - 0,52]	0,49 [0,21 - 0,77]	96,9	0,33
Zijn uw ogen goed genoeg om de gewone kleine[3] letters in de krant te kunnen lezen (zo nodig met bril of contactlenzen)[2]	80,0	0,41 [0,29 - 0,53]	0,60 [0,44 - 0,76]	92,4	0,39
Kunt u hard voedsel bijten en kauwen, zoals bijvoorbeeld een harde appel?	67,4	0,52 [0,42 - 0,63]	0,73 [0,61 - 0,84]	82,5	0,47
Kunt u zelf uw eten snijden, zoals bijvoorbeeld vlees?	96,0	0,66 [0,47 - 0,85]	0,90 [0,82 - 0,97]	98,8	0,15
Kunt u de nagels van uw tenen knippen?	61,7	0,62 [0,52 - 0,71]	0,80 [0,72 - 0,89]	91,1	0,53
Kunt u, als u staat, bukken en iets van de grond oppakken?[4]	72,9	0,62 [0,52 - 0,72]	0,77 [0,68 - 0,87]	80,1	0,33
Kunt u een voorwerp van 5 kilo, bijvoorbeeld een volle boodschappentas, 10 meter dragen?	69,2	0,58 [0,48 - 0,68]	0,75 [0,65 - 0,86]	84,4	0,43
Kunt u zich zelf aan- en uitkleden?	93,8	0,46 [0,24 - 0,67]	0,52 [0,22 - 0,82]	97,7	0,09
Kunt u zelf in en uit bed stappen?	96,4	0,37 [0,12 - 0,61]	0,61 [0,26 - 0,96]	98,2	0,57
Kunt u zich op dezelfde verdieping van de ene naar de andere kamer verplaatsen?	97,3	0,56 [0,25 - 0,88]	0,58 [0,19 - 0,97]	99,0	0,01
Kunt u een trap van 15 treden op- en aflopen zonder stil te moeten staan?[5]	76,2	0,59 [0,49 - 0,69]	0,81 [0,73 - 0,89]	90,2	0,41
Kunt u 400 meter aan een stuk lopen, zonder stil te staan? (zo nodig met stok)[6]	76,9	0,57 [0,46 - 0,68]	0,81 [0,74 - 0,90]	91,2	0,39
Zou u 100 meter hard kunnen lopen?	21,7	0,60 [0,52 - 0,68]	0,85 [0,80 - 0,90]	63,4	0,48

- = Kappa's niet berekend omdat het aantal personen dat niet tweemaal 'ja, zonder moeite' antwoordt te klein is.

[1] CBS onderzoek: voeren i.p.v. volgen.
[2] CBS onderzoek: toevoeging is (eventueel met bril); contactlenzen worden niet vermeld.
[3] CBS onderzoek: kleine letters i.p.v. gewone kleine letters.
[4] CBS onderzoek: kunt u iets oprapen terwijl U de benen gestrekt houdt.
[5] CBS onderzoek: gesplitst in 2 afzonderlijke vragen: trap van 15 treden oplopen respectievelijk aflopen zonder te moeten rusten.
[6] CBS onderzoek: zonder de toevoeging (zo nodig met stok).

beeld niet wezenlijk, alleen blijkt nu de vraag naar 'een gesprek voeren met één persoon' wel meer overeenstemming te vertonen dan op grond van toeval wordt verwacht (betrouwbaarheidsinterval omvat niet langer de waarde 0).

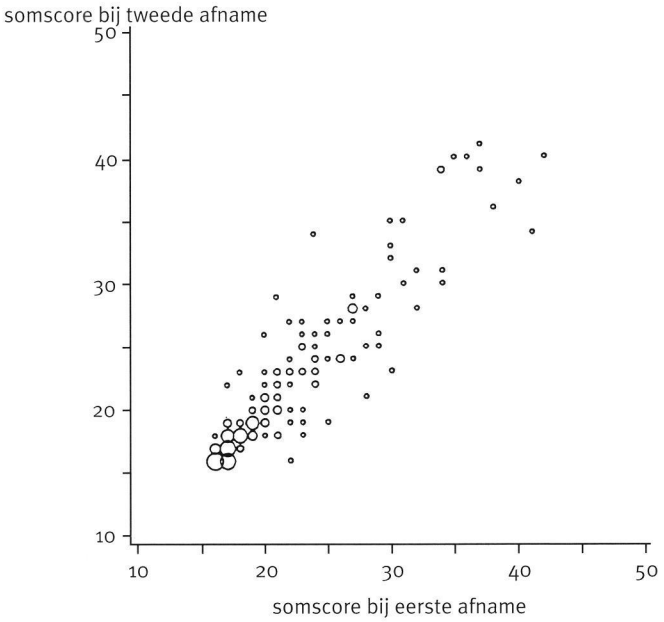
Tabel 4 geeft de kappa's en de gewogen kappa's voor de test-hertest met 18 weken tussenruimte, en de kappa's uit het eerdere onderzoek van het CBS, eveneens met gemiddeld 18 weken (4,5 maand) tussenruimte. De kappa's voor een herhaalde meting na 18 weken verschillen niet duidelijk van de kappa's voor een herhaalde meting na zes weken. Wel vallen nu hier en daar net andere vragen boven c.q. onder de grenzen van 0,40 en 0,75. Uit deze tabel blijkt niet duidelijk dat de overeenstemming met 4,5 maand tussenruimte minder goed is dan met een tussenperiode van zes weken. Dit is niet te verklaren doordat de analyses andere groepen respondenten betreffen: ook binnen de groep die de vragenlijst driemaal beantwoordt zijn geen duidelijke verschillen zichtbaar (gegevens niet getoond). De kappa's uit het huidige onderzoek blijken in de meeste gevallen hoger te zijn dan die uit het CBS onderzoek. *Tabel* 5 maakt aannemelijk dat de grotere overeenstemming in ons onderzoek vooral te danken is aan het beter overeenkomen van de antwoorden 'met enige moeite', 'met grote moeite' en 'kan niet', die in onze steekproef met een hogere frequentie voorkomen. In beide onderzoeken blijkt verder dat de ant-

Tabel 5 *Percentage vragen dat bij herhaalde afname van de vragenlijst na 18 weken hetzelfde wordt beantwoord, uitgesplitst naar bij eerste afname gegeven antwoord*

Antwoordcategorieën	% overeenkomende antwoorden	
	dit onderzoek	CBS-onderzoek
ja, zonder moeite	93,3	95,1
ja, met enige moeite	52,5	27,8
ja, met grote moeite	34,8	17,6
Nee, kan ik niet	76,8	60,7

woordcategorieën aan het uiteinde van de schaal stabieler worden beantwoord dan de middelste categorieën.

Tabel 6 geeft de correlatiecoëfficiënten tussen de somscores op de verschillende afnamemomenten. Deze zijn alle boven de 0,85. De test-hertest correlatie is hoger voor de totale somscore dan voor het aantal malen "minimaal met grote moeite"en het aantal malen "kan niet". Ook hier is de correlatie vrijwel identiek voor personen die geen bezoek aan een arts hadden gebracht. Om meer inzicht te krijgen in de betekenis van deze correlatiecoëfficiënten voor de meetkwaliteit van de somscores, worden in *figuur* 1 de gegevens weergegeven waarop de correlatiecoëfficiënt voor de totale somscore bij herhaalde meting na zes weken is gebaseerd. In deze figuur is de oppervlakte van de getekende cirkels evenredig met het aantal personen dat het betreft: de kleinste cirkels representeren één persoon; de grootste cirkel (voor tweemaal score 16 = geen enkele moeite met alle 16 handelingen) representeert 14 (7,3%) personen. In 32% van de gevallen blijkt de somscore exact overeen te komen, in 35%



Figuur 1 *Verband tussen de totaalscore bij eerste en bij tweede afname (met een tussenperiode van 6 weken). De oppervlakte van ieder circel geeft het aantal personen weer met de desbetreffende combinatie.*
Noot 1: De oppervlakte van iedere cirkel geeft het aantal personen weer met de desbetreffende combinatie

van de gevallen is er 1 punt verschil, in 13% van de gevallen 2 punten verschil, in 7% van de gevallen 3 punten verschil en in 13% van de gevallen meer dan 3 punten verschil.

De correlatiecoëfficiënten tussen de diverse somscores op de OECD vragenlijst op het eerste en derde meetmoment zijn meestal iets hoger, en slechts een enkele maal iets lager dan die tussen het eerste en het tweede meetmoment (*tabel* 6), ook wanneer alleen die respondenten in de analyse worden betrokken die op alle drie de meetmomenten de vragenlijst volledig invulden. Deze resultaten geven daarom geen aanleiding te veronderstellen dat de test-hertest betrouwbaarheid kleiner wordt in de loop van een periode van 4,5 maand.

Discussie

In dit onderzoek wordt voor de OECD vragenlijst een betere test-hertest betrouwbaarheid gevonden dan in eerder onderzoek van het CBS. De variërende tussenperiode in het CBS onderzoek (twee tot zes maanden) is hiervan waarschijnlijk niet de oorzaak. In dit onderzoek worden immers geen lagere kappa's gevonden voor een tussentijd van 18 weken in vergelijking met een tussentijd van zes weken.

De oorzaak van de betere reproduceerbaarheid in dit onderzoek is waarschijnlijk gelegen in de leeftijd van de onderzoekspopulatie: in dit onderzoek 55-89 jaar, in het CBS onderzoek vanaf 16 jaar. In de jongere CBS-populatie komen duidelijk minder beperkingen voor. Voor kappa's in twee-bij-twee tabellen is bekend dat wanneer de prevalentie de 0 of 100% nadert, de kappa (bij een zelfde percentage random misclassificatie) sterk daalt.[18-20] Aangenomen kan worden dat dit ook geldt voor kappa's in k-bij-k (k>2) tabellen. Ook hier geldt immers dat een relatief geringe misclassificatie bij een antwoord dat frequent wordt gegeven, een in verhouding groot deel van de positieve antwoorden veroorzaakt in de categorieën die relatief zelden worden aangekruist. Hierdoor wordt de kappa laag. Wanneer de andere antwoordcategorieën frequenter van toepassing zijn, neemt in deze categorieën het aantal echte positieve antwoorden relatief toe en daarmee ook de kappa. De resultaten in *tabel* 6 ondersteunen deze verklaring van de verschillen tussen ons onderzoek en dat van het CBS. De reproduceerbaarheid van het antwoord 'ja, zonder moeite' verschilt nauwelijks tussen beide onderzoeken, de verschillen zitten vooral in de overige antwoordcategorieën die door onze respondenten aanzienlijk frequenter worden gegeven. Een andere verklaring zou kunnen zijn dat jongeren vaker dan ouderen geplaagd worden door tijdelijke, in plaats van langdurige gebreken. De in onze studie gevonden reproduceerbaarheid lijkt eveneens groter dan die gevonden door Wilson en McNeil.[9] Dit zou goed kunnen komen doordat in ons geval, anders dan bij Wilson en McNeil, voor de opeenvolgende waarnemingen geheel gelijke vraagformuleringen zijn gebruikt.

Het gebruik van Cohen's kappa blijkt in het algemeen een veel negatiever beeld te geven dan gebruik van de gewogen kappa. Daaruit kan worden afgeleid dat verschillen vooral bestaan uit verschuivingen naar naastliggende categorieën. Op zich is dit begrijpelijk wanneer de schaal wordt

Tabel 6 *Correlatiecoëfficiënten [95% betrouwbaarheidsinterval] voor de somscores op de verschillende afnamemomenten*

	tussenperiode 6 weken				tussenperiode 18 weken	
	alle respondenten		alleen respondenten zonder artsbezoek in de laatste maand		alle respondenten (n=165)	
	iedereen (n=179)	alleen personen die ook na 18 weken de lijst invulden (n=165)	iedereen (n=105)	alleen personen die ook na 18 weken de lijst invulden (n=99)	iedereen (n=196)	alleen personen die ook na 6 weken de lijst invulden (n=165)
Totaal score	0,92 [0,89-0,94]	0,93 [0,91-0,95]	0,92 [0,89-0,95]	0,94 [0,91-0,96]	0,94 [0,92-0,95]	0,94 [0,92-0,96]
Aantal handelingen waar men minimaal grote moeite mee heeft	0,89 [0,85-0,91]	0,90 [0,86-0,92]	0,88[0,82-0,91]	0,88 [0,83-0,92]	0,92 [0,89-0,94]	0,92 [0,89-0,94]
Aantal handelingen dat niet uitgevoerd kan worden	0,86 [0,82-0,90]	0,87 [0,83-0,90]	0,85 [0,78-0,89]	0,86 [0,79-0,90]	0,88 [0,84-0,91]	0,86 [0,81-0,89]

gezien als een categorisering van een achterliggende continue variabele: soms zal de 'echte waarde' tussen de middens van twee categorieën in liggen en zal het moeilijk zijn te bepalen of men iets nu bijvoorbeeld 'met enige moeite' of 'met grote moeite' kan doen. Daardoor zal soms het ene, soms het andere antwoord worden geven. De grootte van dit effect zal overigens per antwoordcategorie kunnen verschillen (afhankelijk van welk deel van de personen in de desbetreffende categorie zich in het overgangsgebied met een naburige categorie bevindt).

Wanneer de gewogen kappa als maatstaf wordt aangehouden, blijkt de mate van overeenstemming over het algemeen bevredigend te zijn en voor ongeveer de helft van de vragen zelfs uitstekend. De test-hertest correlaties van verschillende somscores die kunnen worden berekend zijn eveneens uitstekend. Daarbij presteert de totale somscore (waarin alle informatie uit de afzonderlijk vragen is verwerkt) zoals te verwachten beter dan de twee somscores waarin naar het aantal beperkingen wordt gekeken.

We kunnen daarom concluderen dat de indicator als totaal in een oudere populatie een goede test-hertest betrouwbaarheid heeft, en dat de indicator binnen een termijn van 4,5 maand stabiel lijkt te zijn. De test-hertest betrouwbaarheid van diverse afzonderlijke vragen is echter minder goed. Bij het afzonderlijk gebruiken van deze vragen in onderzoek is daarom voorzichtigheid geboden. De resultaten doen wel de vraag rijzen of de vragen waarop de antwoorden het slechts reproduceerbaar zijn niet weggelaten, dan wel geherformuleerd zouden moeten worden. Weglaten van vragen kan de reproduceerbaarheid van de totale indicator weliswaar verbeteren, maar tast de inhoud van de indicator aan. Herformulering ligt daarom meer voor de hand. Daarbij moet worden opgemerkt dat de betrouwbaarheidsmarges rond de gevonden kappa's ruim zijn, en dat daarom de onderlinge rangorde zoals die uit dit onderzoek naar voren komt, in een volgend onderzoek ongetwijfeld anders zal liggen. De vraag over een gesprek kunnen voeren met één andere persoon lijkt echter zo slecht reproduceerbaar dat deze als eerste voor herformulering in aanmerking zou komen. Zolang echter geen vervangende vraag met een hogere reproduceerbaarheid beschikbaar is, lijkt het op inhoudelijke gronden beter de indicator als geheel te handhaven.

Uiteraard is de test-hertest betrouwbaarheid maar één aspect van meetkwaliteit van een instrument. Vragen over of de beantwoording systematisch verschilt tussen groepen (bij afwezigheid van werkelijk bestaande verschillen in beperkingen) kunnen met gegevens over test-hertest betrouwbaarheid niet worden beantwoord. Voor het beantwoorden van dergelijke vragen dient verdere validering plaats te vinden door bijvoorbeeld vergelijking van antwoorden met observaties van het daadwerkelijk functioneren.

Abstract

In 1981, the OECD published a questionnaire for measuring the prevalence of chronic disabilities in the general population. We estimated test-retest reliability of this questionnaire in an elderly population (55-85 years of age, n=356). The questionnaire was mailed three times to the participants (at t=0, 6 and 18 weeks). The first and second questionnaire were returned by 216 persons (61%); the first and the third questionnaire by 231 (65%). The test-retest reliability (measured by the weighted and unweighted kappa) did not differ markedly between the second and third replication of the questionnaire. The item 'able to have a conversation with one other person' was poorly reproducible (unweighted kappa 0.09; weighted kappa 0.22). The unweighted kappa of other items varied from 0.33 to 0.72; the weighted kappa's varied from 0.38 to 0.90. The correlation between the sum of all items of both measurement times was very high (0.92 and 0.94). Kappa's were not appreciably higher for respondents who had not contacted a physician in the month before the second measurement. For most items kappa's were higher than observed in an earlier study of the Netherlands Central Bureau of Statistics, presumably due to a higher prevalence of disability in our population. The reproducibility of

most individual items of the questionnaire is satisfactory; reproducibility of the sum score of all items is high.
Keywords: Test-retest reliability, long-term disability, OECD indicator, kappa.

LITERATUUR

1 *McWhinnie JR.* Disability assessment in population surveys: results of the O.E.C.D. common development effort. Rev Epidemiol Santé 1981;29:413-9.

2 *Foets M, Velden J van der.* Nationale studie naar ziekten en verrichtingen in de huisartspraktijk. Basisrapport: meetinstrumenten en procedures. NIVEL. Utrecht 1990.

3 *Meer JBW van der, Looman CWN, Mackenbach JP.* De longitudinale studie naar sociaal-economische verschillen in medische consumptie (LS-SEVM): enkele eerste resultaten. In: Mackenbach JP, red. De longitudinale studie naar sociaal-economische gezondheidsverschillen (LS-SEGV); opzet en enkele eerste resultaten. Rijkswijk: Ministerie van Welzijn, Volksgezondheid en Cultuur, 1994:111-27.

4 *Sonsbeek JLA van.* Methodische en inhoudelijke aspecten van de OESO-vragenlijst betreffende langdurige beperkingen in het lichamelijk functioneren. Mndber gezondheid (CBS) 1988;88/6:4-17.

5 *Sonsbeek JLA van.* Methodische en inhoudelijke aspecten van het meten van langdurige aandoeningen in enquête-onderzoek. In: Sonsbeek JLA van. Vertel me wat er aan scheelt. Betekenis en methodische aspecten van enquêtevragen naar de gezondheid. Proefschrift Katholieke Universiteit Nijmegen, 1996:141-90.

6 *Sonsbeek JLA van.* De bruikbaarheid van de OESO-indicator voor langdurige lichamelijke beperkingen in relatie tot arbeidsparticipatie. Mndber gezondheid (CBS) 1991;91/5:14-8.

7 *Berg J van den, Sonsbeek JLA van.* Experiences with the OECD long-term disability indicator: use in field-work and coding in IDH-categories. Mndber gezondheid (CBS) 1984;84/2:11-5.

8 *Köning-Zahn C, Furer JW, Tax B.* Het meten van de gezondheidstoestand: beschrijving en evaluatie van vragenlijsten. 2: Lichamelijke gezondheid, sociale gezondheid. Assen: Van Gorcum 1994.

9 *Wilson RW, McNeil JM.* Preliminary analysis of OECD disability on the pretest of the post census disability survey. Rev Epidemiol Santé 1981;29:469-75.

10 *Deeg DJH, Smit JH, Beekman ATF.* De dwang van de analyse methode bij het gebruik van longitudinale gegevens: Het geval van gezondheid en depressie. Tijdschr Soc Gezondheidz 1997;75:129-35.

11 *Fleiss JL.* Statistical methods for rates and proportions. New York, Wiley, 1981, formule 13.29 en 13.30.

12 *Maclure M, Willett WC.* Misinterpretation and misuse of the kappa statistica. Am J Epidemiol 1987;126 :161-9.

13 *Bloch DA, Kraemer HC.* 2 x 2 Kappa coefficients: measures of agreement of association. Biometrics 1989;45:269-87.

14 *Hale CA, Fleiss JL.* Interval estimation under two study designs for kappa with binary classifications. Biometrics 1993;49:523-34.

15 *Donner A, Eliasziw M.* A goodness-of-fit approach to inference procedures for the kappa statistic: confidence interval contruction, significance-testing and sample size estimation. Stat Med 1992;11:1511-9.

16 *Lee JJ, Tu ZN.* A better confidence interval for kappa on measuring agreement between two raters with binary outcomes. J Computational Graphic Stat 1994;3:301-21.

17 *Jonge H de.* Inleiding tot de Medische Statistiek. Deel 2: Klassieke Methoden. Leiden, Nederlands Instituut voor Praeventieve Geneeskunde, 1964.

18 *Guggenmoos-Holzman I.* How reliable are chance-corrected measures of agreement? Stat Med 1993;12:2191-205.

19 *Thompson WD, Walter SD.* A reappraisal of the kappa coefficient. J Clin Epidemiol 1988;41:949-58.

20 *Byrt T, Bishop J, Carlin JB.* Bias, prevalence and kappa. J Clin Epidemiol 1993;46:423-9.

CORRESPONDENTIEADRES

Dr. H.C. Boshuizen, RIVM, afdeling IMA, Postbus 1, 3720 BA Bilthoven, tel. 030 2742944. E-mail: Hendriek.Boshuizen@RIVM.nl

Voor publicatie geaccepteerd op 28 oktober 1999