Nederlands Instituut van Psychologen NIP

COTAN Beoordelingssysteem voor de kwaliteit van tests

geheel herziene versie, mei 2009; gewijzigde herdruk mei 2010

Arne Evers, Wouter Lucassen, Rob Meijer en Klaas Sijtsma

Inhoud

Woord vooraf 1

Inleiding 2

1 Uitgangspunten van de testconstructie 7

2 Kwaliteit van het testmateriaal 9

3 Kwaliteit van de handleiding 16

4 Normen 19

5 Betrouwbaarheid 31

6 Begripsvaliditeit 38

7 Criteriumvaliditeit 43

Literatuur 46

Auteurs

Arne Evers, Universiteit van Amsterdam, Programmagroep Arbeids- & Organisatiepsychologie

Wouter Lucassen, Meurs HRM, Woerden

Rob Meijer, Rijksuniversiteit Groningen, Heymans Instituut DPMG

Klaas Sijtsma, Universiteit van Tilburg, Departement Methoden en Technieken van Onderzoek, FSW

Woord vooraf

Voor u ligt de herziene versie van het COTAN Beoordelingssysteem voor de Kwaliteit van Tests. Deze versie bouwt voort op eerdere versies van het systeem, zoals gepubliceerd in de *Documentatie van Tests en Testresearch* van 1982 (Visser, van Vliet-Mulder, Evers & ter Laak) en 2000 (Evers, van Vliet-Mulder & Groot) en op de website van het NIP in 2004 (www.psynip.nl). Deze nieuwe versie is tot stand gekomen in een werkgroep van de Commissie Testaangelegenheden Nederland (COTAN), die bestaat uit de vier ondertekenaars van dit woord vooraf. Aan de discussie over de inhoud van deze nieuwe versie van het beoordelingssysteem hebben alle COTAN-leden een bijdrage geleverd en deze tekst is goedgekeurd in de COTAN-vergadering van 19 maart 2009. Tijdens de voltooiing van deze herziening bestond de COTAN uit de volgende leden: K. Sijtsma (voorzitter), J.B. Blok (secretaris), A. Evers (senior editor testbeoordelingen), R.M. Frima (stafmedewerker), R.H. van den Berg, M.Ph. Born, H.W. van Boxtel, M.E. Dinger, B.T. Hemker, P.P.M. Hurks, W.W. Kersten, W.I. Lucassen, R.R. Meijer, E.F.M. Pouw, W.C.M. Resing, J.D.L.M. Schutijser en T. van Strien.

Deze herziening betreft een ingrijpende aanpassing van het beoordelingssysteem in het licht van de ontwikkelingen die zich in de loop der jaren op het gebied van de testtheorie en de testconstructie hebben voorgedaan. Voorbeelden van deze ontwikkelingen zijn *Computer Based Tests*, *Item Respons Theorie* en *continue normering*. Hoewel de vorige versie van het beoordelingssysteem met enige flexibiliteit ook kon worden toegepast op tests die gebruikmaakten van deze nieuwe ontwikkelingen, was de tekst toch vooral gericht op tests die op de klassieke wijze waren ontwikkeld en op de klassieke wijze werden aangeboden. Aan deze nieuwe technieken en benaderingen wordt in deze herziening explicieter aandacht besteed. De COTAN hoopt daarmee de bruikbaarheid en toepasbaarheid van het beoordelingssysteem op een hoger peil te hebben gebracht. De complexiteit van de nieuwe ontwikkelingen, evenals de noodzaak om de gehanteerde procedures en resultaten voor testgebruikers en de COTAN op adequate wijze te rapporteren, stelt hoge eisen aan de kennis en vaardigheden van de testauteur. Bij de toepassing van deze meer geavanceerde technieken zullen testauteurs dan ook vaker dan voorheen een beroep moeten doen op de expertise van deskundigen.

Ook in dit nieuwe beoordelingssysteem worden tests beoordeeld op de bekende zeven criteria, die in de inleiding van deze tekst worden genoemd. Uiteraard heeft het voortschrijdende inzicht op het gebied van testtheorie, testconstructie en testtoepassing geleid tot aanpassingen in de beoordelingsgezichtspunten en de wijze waarop deze per criterium tot een beoordeling leiden. Een consequentie daarvan is dat bij één instrument de beoordelingsresultaten volgens het oude en het nieuwe systeem van elkaar zouden kunnen verschillen.

Is het beoordelingssysteem nu strenger geworden? Over het algemeen kan dat niet worden gezegd. Voor sommige kenmerken, waarover in het vorige systeem een hard oordeel werd geveld, wordt nu een nuancering in de beoordeling mogelijk gemaakt, terwijl op andere punten de eisen juist zijn aangescherpt. De belangrijkste verandering is echter dat de beoordelingsvoorschriften veel specifieker – en niet noodzakelijkerwijs soepeler of strenger – zijn geworden.

In deze publicatie worden bij de diverse criteria de kwaliteitseisen beknopt toegelicht. Deze beschouwingen hebben niet de pretentie van volledigheid, noch vormen zij een receptenboek voor testconstructie. Van testauteurs mag worden verwacht dat zij op de hoogte zijn van de fundamenten van de psychometrie en van actuele kwaliteitsmaatstaven in de testconstructie, zodat zij ten volle de verantwoordelijkheid voor de kwaliteit van hun instrument kunnen nemen.

Voor deze herziening zijn met instemming van de betreffende auteurs tekstdelen overgenomen uit publicaties van Keuning (2004) over *Computer Based Tests*, Wools, Sanders en Roelofs (2007) over absoluut normeren en Bechger, Hemker en Maris (2009) over continue normering. De COTAN is hun veel dank verschuldigd voor hun deskundige bijdragen.

In de tweede druk is het erratum in de tabel voor de vaststelling van het eindoordeel voor criterium 4 verwerkt. Verder zijn enkele kleine redactionele verbeteringen aangebracht.

Arne Evers, Wouter Lucassen, Rob Meijer en Klaas Sijtsma
April 2010

Inleiding

Inhoud van het beoordelingssysteem

Hoewel de COTAN zich al sinds haar oprichting in 1959 bezighoudt met het beschrijven van tests, werden pas in de *Documentatie van Tests en Testresearch* van 1969 voor het eerst beoordelingen van tests gepubliceerd. Het gehanteerde systeem leidde tot een globaal oordeel over tests. De beoordeling kon variëren van A (zeer goede test) tot F (slechte test of test in ontwikkeling). Mede vanuit de behoefte een meer gedifferentieerd oordeel over een test mogelijk te maken, werd in de *Documentatie* van 1982 een geheel nieuw systeem gebruikt. Elke test werd beoordeeld op vijf criteria: 1) uitgangspunten van de testconstructie, 2) kwaliteit van het testmateriaal en de handleiding, 3) normen, 4) betrouwbaarheid en 5) validiteit. Bij de voorbereiding van de *Documentatie van Tests en Testresearch* van 2000 werd het systeem herzien, waarbij de vragen, de toelichtingen en de wegingsvoorschriften werden aangepast. Het ging hierbij vooral om relatief kleine aanpassingen en verbeteringen. Een wezenlijke verandering betrof echter de splitsing van de criteria 2 en 5 in aparte beoordelingen voor kwaliteit van het testmateriaal, kwaliteit van de handleiding, begripsvaliditeit en criteriumvaliditeit. Zonder veel moeite konden destijds de beoordelingen op vijf criteria worden omgezet naar de meer gedifferentieerde beoordelingen op de zeven nieuwe criteria.

Bij de huidige herziening zijn de wijzigingen echter ingrijpender, hoewel het aantal van zeven criteria ongewijzigd is gebleven. Ook het principe van verscheidene vragen per criterium, waaronder een of meer basisvragen, is onveranderd gebleven. Nieuwe inzichten op het gebied van testtheorie en testconstructie hebben er echter toe geleid dat er vragen zijn toegevoegd, dat sommige vragen zijn onderverdeeld in subvragen en dat andere vragen zijn vervallen. Ook zijn er aanwijzingen gespecificeerd en scoringsregels veranderd. Als men een test volgens het oude en het nieuwe systeem zou beoordelen, kan dat uiteraard tot verschillen in beoordeling leiden. Vanaf april 2009 worden tests volgens het nieuwe systeem beoordeeld. Er zullen geen herbeoordelingen plaatsvinden van de circa 600 tests die met de 'oude' versie zijn beoordeeld. Alleen met betrekking tot de veroudering van normgegevens (zie vraag 4.2) wordt hierop een uitzondering gemaakt.

Hieronder wordt een globale omschrijving van het herziene beoordelingssysteem gegeven, inclusief de belangrijkste wijzigingen. De zeven criteria zijn:

1 ***Uitgangspunten van de testconstructie.***
Dit criterium wordt beoordeeld door middel van drie vragen, waarmee achtereenvolgens wordt vastgesteld of het gebruiksdoel, de theoretische achtergrond en de operationalisatie daarvan in de testinhoud zijn beschreven. De beoordeling op dit criterium is van invloed op de waardering van andere criteria, omdat de meetpretentie bepaalt welk type normerings-, betrouwbaarheids- en validiteitsonderzoek moet worden verricht. De grootste verandering ten opzichte van de vorige versie van het beoordelingssysteem is dat de eerste vraag is gesplitst in drie deelvragen waarin expliciet wordt gevraagd of de meetpretentie, de doelgroepen en de functie van de test worden beschreven.

2 ***Kwaliteit van het testmateriaal.***
Dit criterium wordt beoordeeld door middel van acht vragen. Bij dit criterium komt onder andere aan de orde of testopgaven, scoring en instructie zijn gestandaardiseerd en of er voldoende aanwijzingen voor de geteste worden gegeven. Ook wordt er een vraag gesteld over de voor specifieke bevolkingsgroepen mogelijk kwetsende inhoud van items. Nieuw bij dit criterium is dat er een vraag over de kwaliteit van de items wordt gesteld en dat er aparte series vragen zijn opgenomen voor afname met behulp van papier-en-potlood en afname via een computer.

3 ***Kwaliteit van de handleiding.***
Dit criterium wordt beoordeeld door middel van zeven (papier-en-potlood) of tien (computer) vragen. Bij dit criterium wordt gevraagd naar de informatie die wordt geboden ter ondersteuning van de testgebruiker bij afname en interpretatie van de test. De voornaamste wijziging is dat er voor afname via een computer drie additionele vragen worden gesteld.

4 ***Normen.***
Dit criterium wordt beoordeeld door middel van zeven vragen (normgerichte interpretatie) of vijf vragen (domeingerichte of criteriumgerichte interpretatie). Nieuw voor alle typen normen is dat bij de beoordeling rekening wordt gehouden met een verjaringstermijn. Bij normgerichte interpretatie wordt verder vastgesteld wat de kwaliteit is van de normen en van de erbij verstrekte informatie. Ook voor normen die via continue normering worden berekend, worden nu richtgetallen gegeven voor de gewenste grootte van de normgroepen. Nieuw zijn ook de vragen die betrekking hebben op domeingerichte en criteriumgerichte interpretatie.

5 ***Betrouwbaarheid.***
Dit criterium wordt beoordeeld door middel van drie vragen. Eerst wordt de hoogte van de betrouwbaarheidscoëfficiënten beoordeeld en vervolgens de kwaliteit van het uitgevoerde onderzoek naar de betrouwbaarheid. Als gevolg van nieuwe ontwikkelingen wordt er nu naar zes mogelijke betrouwbaarheidsmaten gevraagd (in vergelijking met vier in de vorige versie van het beoordelingssysteem).

6 ***Begripsvaliditeit.***
Dit criterium wordt beoordeeld door middel van drie vragen. Eerst worden de uitkomsten beoordeeld en vervolgens de kwaliteit van het uitgevoerde onderzoek naar de begripsvaliditeit. Nieuw is dat explicieter wordt aangegeven welke typen onderzoeksgegevens ter ondersteuning van de begripsvaliditeit kunnen dienen en welke typen gegevens voor een bepaalde beoordeling zijn vereist.

7 ***Criteriumvaliditeit.***
Dit criterium wordt eveneens beoordeeld door middel van drie vragen. Net als bij de begripsvaliditeit wordt eerst de hoogte van de uitkomsten beoordeeld en vervolgens worden deze geëvalueerd in het licht van de kwaliteit van de onderzoeksprocedure. Bij dit criterium hebben geen belangrijke wijzigingen plaatsgevonden.

Het oordeel voor elk van deze criteria kan 'onvoldoende', 'voldoende' of 'goed' zijn. De schaalpunten op de vragen zijn 1, 2 en 3; deze komen eveneens overeen met de betekenissen 'onvoldoende', 'voldoende' of 'goed'. Bij enkele basisvragen moet de score '1' worden geïnterpreteerd als 'nee', de score '2' als 'n.v.t.' en de score '3' als 'ja'. Een negatief oordeel op een basis(deel)vraag leidt direct tot het oordeel 'onvoldoende' voor het betreffende criterium. De inhoud van de vragen, de toelichtingen en de wegingsvoorschriften worden gegeven in de volgende zeven hoofdstukken. Bij de wegingsvoorschriften van sommige criteria moeten somscores van vragen worden berekend.

Het doel van de toelichtingen is een houvast te bieden bij de beoordeling en waar nodig de statistische of psychometrische beweegredenen te verduidelijken. De toelichtingen hebben uiteraard niet de pretentie een statistisch of psychometrisch leerboek te zijn. Bij onduidelijkheden kan men de referenties of andere literatuur op de gebieden van testconstructie en psychometrie raadplegen.

De betekenis van de beoordelingen

Over het algemeen kan men stellen dat een 'onvoldoende' voor een criterium op twee manieren tot stand kan komen: óf omdat de gevraagde informatie afwezig is, óf omdat de kwaliteit van de wél aanwezige informatie negatief wordt beoordeeld. Zo kan een 'onvoldoende' voor de betrouwbaarheid van een test betekenen dat de betrouwbaarheid niet is onderzocht óf dat deze wel is onderzocht, maar dat dit onderzoek heeft aangetoond dat de test onvoldoende betrouwbaar is. Afwezigheid van onderzoeksgegevens wordt dus op dezelfde wijze beoordeeld als wél beschikbare onderzoeksgegevens die tot een negatief resultaat leiden, omdat de COTAN meent dat het aan de auteur is om onderzoeksgegevens te verschaffen. Hiermee worden de wetenschappelijke mores gevolgd dat de bewijslast voor een uitspraak bij de onderzoeker ligt. In het hierboven beschreven voorbeeld betekent dit dat de test bij afwezigheid van gegevens als onvoldoende betrouwbaar wordt gezien tot het tegendeel is aangetoond. Voor de testgebruiker kan het zinnig zijn onderscheid te maken tussen deze situaties, omdat hij bijvoorbeeld graag een nieuw veelbelovend instrument het voordeel van de twijfel schenkt. Mede om dit onderscheid mogelijk te maken, maar ook als extra informatiebron voor testauteur en testgebruiker, wordt bij 'onvoldoendes' van tests die sinds 1992 zijn beoordeeld, in het kort de reden van de beoordeling gegeven. Overigens wordt hierbij nog een keer benadrukt dat het de verantwoordelijkheid van de testauteur is om te zijner tijd voldoende informatie te verschaffen. Daarbij zou het krediet dat gebruikers aan een onvoldoende onderbouwd instrument geven omgekeerd evenredig moeten zijn met de ouderdom ervan.

Een tweede nuancering ten aanzien van de beoordeling 'onvoldoende' is, dat een of meer 'onvoldoendes' niet per se betekent dat een instrument onbruikbaar is. Zo kan een 'onvoldoende' voor normen zijn gegeven omdat de representativiteit van de normgroep te wensen overlaat. De test kan echter zeer bruikbaar zijn als de gebruiker in staat is zelf geschikte normen te verzamelen. Ten aanzien van betrouwbaarheid en validiteit gelden soortgelijke overwegingen. Een of meer schalen of subtests van een vragenlijst of test kunnen onvoldoende betrouwbaar zijn; dit hoeft echter niet te betekenen dat de andere schalen of subtests of de totaalscore onbruikbaar zijn. Bij tests die worden gebruikt voor belangrijke beslissingen op individueel niveau worden hoge eisen gesteld aan de betrouwbaarheid (zie de toelichting bij criterium 5). Zo wordt de betrouwbaarheid van een dergelijke test als 'onvoldoende' beoordeeld als deze lager is dan .80. Toch kan een dergelijke test nuttige informatie opleveren, bijvoorbeeld in combinatie met andere instrumenten. Omdat met dit beoordelingssysteem slechts afzonderlijke tests worden beoordeeld, kan met een dergelijke wijze van gebruik geen rekening worden gehouden. Ook is het binnen dit beoordelingssysteem mogelijk dat van deze zelfde test met een 'onvoldoende' voor betrouwbaarheid, de begrips- of de criteriumvaliditeit als 'voldoende' of zelfs als 'goed' wordt beoordeeld, bijvoorbeeld omdat in selectiesituaties een validiteitscoëfficiënt van .40 als hoog wordt beoordeeld. Zelfs een test met een lage voorspellende waarde kan in sommige gevallen nuttige informatie opleveren, afhankelijk van bijvoorbeeld toevalskans, selectieratio en kosten-batenverhouding.

Een derde nuancering betreft de grenswaarden die in het beoordelingssysteem worden genoemd en waaraan tests moeten voldoen om een zo groot mogelijke objectiviteit bij de beoordeling te garanderen. Zo worden bij de criteria Normen en Betrouwbaarheid specifieke steekproefgroottes respectievelijk hoogtes van betrouwbaarheidscoëfficiënten genoemd waaraan moet worden voldaan voor een 'voldoende' of 'goed' beoordeling en die als ankerpunt fungeren voor de beoordelaar. Voor deze grenzen is echter geen sluitende wetenschappelijke argumentatie te leveren: ze zijn gebaseerd op in het algemeen min of meer internationaal geaccepteerde adviezen van vooraanstaande deskundigen (zie de betreffende hoofdstukken voor referenties). Hiermee hangt samen dat in ieder geval van waarden die in de buurt van deze grenzen liggen, nauwelijks is te beargumenteren waarom een bepaalde waarde net wel, en een andere waarde net niet 'voldoende' of 'goed' is. Op deze wijze kan echter beter worden gewaarborgd dat alle tests in principe op dezelfde wijze worden beoordeeld.

Met bovenstaande opmerkingen is bedoeld duidelijk te maken dat van de testgebruiker wordt verwacht dat hij met de in absolute termen gegeven beoordelingen op de juiste wijze kan omgaan. Voor de deskundige testgebruiker heeft het oordeel 'onvoldoende' (voor welk criterium dan ook) vooral de functie van waarschuwingssignaal; in zo'n geval moet de testgebruiker, in overeenstemming met artikel 3.2.e van *de Algemene Standaard Testgebruik* (Nederlands Instituut van Psychologen, 2004), expliciet beargumenteren waarom hij het betreffende instrument inzet. Voor de minder deskundige testgebruiker is de boodschap, vooral wanneer er meerdere onvoldoendes voor een test voorkomen: testgebruiker, gebruik deze test niet!

Beoordelingsprocedure

Er wordt wel eens gedacht dat de COTAN alleen dié tests beoordeelt die ter beoordeling worden aangeboden. Dit is niet juist. De COTAN voert een actief beleid en beoordeelt in principe alle tests die voor opname in de *Documentatie van Tests en Testresearch* in aanmerking komen. (Deze opnamecriteria staan vermeld op de website van NIP/COTAN onder het kopje Criteria, zie www.psynip.nl of www.cotan.nl.) Een eerste stap, althans als de COTAN zelf het initiatief neemt tot een testbeoordeling, is het verzamelen van materiaal en publicaties van en over een test (testboekje, sleutels, handleiding, software, artikelen, proefschrift, e.d.). Meestal worden deze materialen en publicaties spontaan of op verzoek door de auteurs of uitgevers beschikbaar gesteld. Een probleem vormen de auteurs die om wat voor reden dan ook (zie ook de volgende paragraaf) weigeren het testmateriaal voor beoordeling af te staan. Deze tests kunnen niet worden gedocumenteerd en beoordeeld.

Het verzamelde materiaal wordt naar twee onafhankelijk van elkaar werkende beoordelaars gestuurd. In navolging van het beleid van erkende internationale tijdschriften op het gebied van de psychologie zijn de beoordelaars van een specifieke test anoniem. Alle COTAN-leden en een groep van daartoe aangezochte externe deskundigen fungeren als beoordelaar. Een beoordelaar krijgt een test toegewezen op basis van zijn deskundigheid op een bepaald gebied. Verder wordt ernaar gestreefd dat van een koppel beoordelaars er altijd minstens één COTAN-lid is. Ook wordt er bij de toewijzing van beoordelaars op gelet dat zij nooit een door henzelf of door een directe collega geconstrueerde test beoordelen, noch die van een concurrerende organisatie. Bij discrepanties in de beoordeling wordt de beoordelaars gevraagd in onderling overleg tot consensus te komen. In uitzonderingsgevallen wordt een derde beoordelaar ingeschakeld. Principiële beoordelingskwesties worden in de tweemaandelijkse COTAN-vergaderingen besproken.

Beide beoordelaars leveren een inhoudelijke onderbouwing van hun beoordeling, die de *senior editor* integreert tot een toelichting die samen met het eindoordeel als feedback aan de testauteur ter beschikking wordt gesteld. Aan de testauteur wordt vervolgens de mogelijkheid geboden op de beoordeling te reageren. Reacties en commentaren van de auteur worden door de beoordelaars behandeld. Als daar aanleiding toe is, worden onderdelen van tests nog een keer beoordeeld, eventueel door een onafhankelijk werkende derde beoordelaar. Vervolgens wordt de beoordeling gepubliceerd. De digitale versie van de *Documentatie van Tests en Testresearch* wordt daartoe maandelijks bijgewerkt. Met de invoering van deze herziene versie van het beoordelingssysteem zal ook worden overgegaan tot publicatie van de (geïntegreerde) toelichting die de beoordelaars bij hun beoordeling leveren, omdat dit kan verhelderen waarom een bepaalde beoordeling is gegeven. Er zal naar worden gestreefd om ook voor tests die recentelijk nog met de 'oude' versie van het systeem zijn beoordeeld deze toelichting te publiceren.

Hierboven is de beoordelingsprocedure geschetst voor een nieuwe test. Wanneer nieuwe informatie over een al beoordeelde test verschijnt (bijvoorbeeld een herziene testhandleiding, een onderzoeksrapport of een complete testrevisie), kan een herbeoordeling van de test plaatsvinden. De gehele hierboven beschreven procedure wordt dan opnieuw doorlopen. Wel is hieraan de beperking gesteld dat sinds de vorige beoordeling van de test minstens een jaar moet zijn verstreken.

Vertrouwelijkheid

De beoordelingsprocedure kent twee vormen van vertrouwelijkheid. De eerste is, zoals gezegd, die van de anonimiteit van de beoordelaars. Voor de testauteur blijft onbekend wie de test heeft beoordeeld. Correspondentie verloopt via het *Documentatiecentrum* van de COTAN of via de door de COTAN aangewezen *senior editor* testbeoordelingen. Hiermee wordt vermeden dat discussies over een beoordeling zich op een persoonlijk vlak kunnen gaan afspelen. Deze procedure is in overeenstemming met de procedure zoals belangrijke psychologische tijdschriften die bij het beoordelen van manuscripten hanteren. Wel wordt eenmaal per jaar op de website van de COTAN een lijst van beoordelaars gepubliceerd met het aantal door hen beoordeelde tests.

De tweede vorm van vertrouwelijkheid betreft het aangeschafte of ter beschikking gestelde testmateriaal. Derden hebben geen toegang tot de informatie die zich in het documentatiecentrum van de COTAN bevindt en het materiaal wordt aan de beoordelaars toegezonden onder de conditie van strikte geheimhouding. De COTAN is zich er als geen ander van bewust dat op tests niet alleen auteursrechten rusten, maar ook dat een test veelal het eindproduct is van een kostbaar ontwikkelingsproces en in feite het werkkapitaal vormt van psychologen en adviesbureaus, waardoor deze zich mede kunnen onderscheiden van andere(n). Niet alleen wordt daarom de beoordelingsprocedure met grote zorgvuldigheid uitgevoerd, de testauteur kan er ook van verzekerd zijn dat het testmateriaal niet door anderen dan degenen die bij de beoordeling van de betreffende test zijn betrokken, wordt ingezien. Dit aspect wordt hier benadrukt, omdat testauteurs voornamelijk als reden voor het niet beschikbaar stellen van testmateriaal geven dat daarmee de test tot openbaar bezit wordt gemaakt. Deze vrees is ongegrond. Het enige wat van de test openbaar wordt gemaakt zijn de standaardbeschrijving in de Documentatie, de beoordelingen op de zeven criteria van het beoordelingssysteem en de toelichting van de beoordelaars.

Om een test te beschermen tegen ongeoorloofd kopieergedrag zijn sommige auteurs en uitgevers ertoe overgegaan om de informatie in de handleiding te beperken. Dit betreft vooral informatie over de inhoud van (sub)schalen en informatie over de normering. De testgebruiker weet daardoor niet welke items tot welke schaal behoren en/of kan zelf de normtabellen niet raadplegen. De scoring wordt in dergelijke gevallen door online software of op afstand door de scoringsservice van de uitgever verricht. Met nadruk wordt hier gesteld dat de COTAN geen voorstander is van deze praktijk, omdat het de

interpretatiemogelijkheden van de testgebruiker ernstig beperkt. Voor de beoordeling is het van belang dat de beoordelaars in ieder geval wél de beschikking hebben over alle informatie die een complete testbeoordeling mogelijk maakt. In voorkomende gevallen wordt de auteur/uitgever daarom verzocht om voor de beoordeling dergelijke informatie te verschaffen. Deze wordt uiteraard met dezelfde mate van vertrouwelijkheid behandeld als hierboven beschreven. Het ontbreken van deze informatie kan op een of meer criteria tot de beoordeling 'onvoldoende' leiden.

Vertalingen of bewerkingen van buitenlandse tests

Veel Nederlandse tests zijn vertalingen of bewerkingen van buitenlandse instrumenten. Bij de beoordeling van deze instrumenten kan de vraag worden gesteld in hoeverre met de buitenlandse versie verricht onderzoek van toepassing is op de Nederlandse situatie of betrokken kan worden bij de beoordeling van de Nederlandse versie. Het antwoord op deze vraag is deels afhankelijk van de mate van letterlijkheid van de vertaling. In sommige gevallen heeft men de bedoeling het buitenlandse instrument zo letterlijk mogelijk over te zetten in het Nederlands, zowel taalkundig als naar betekenisovereenkomst (van items en begrippen). Hiervoor geeft de *International Test Commission* een aantal regels in de *Test Adaptation Guidelines* (ITC, 2000; zie ook Van de Vijver & Hambleton, 1996, en Hambleton, Merenda, & Spielberger, 2005).

Deze regels betreffen in eerste instantie de wijze waarop de vertaling (inclusief terugvertaling) totstandkomt. Wezenlijk daarbij is dat de vertaling niet letterlijk hoeft te zijn, maar dat het veel meer van belang is dat de teksten natuurlijk overkomen, omdat daardoor wordt bevorderd dat de betekenis van datgene wat wordt gevraagd, in beide talen hetzelfde is. Vervolgens moet equivalentieonderzoek het 'bewijs' opleveren dat met de vertaalde versie hetzelfde begrip wordt gemeten als met de oorspronkelijke versie. Een voorbeeld daarvan is het vergelijken van de factorstructuur van de oorspronkelijke en de Nederlandse versie. Wanneer equivalentie is aangetoond, kunnen de validiteitsgegevens en eventuele test-hertestgegevens van de oorspronkelijke versie bij de beoordeling van de Nederlandse versie worden betrokken. Gegevens in de zin van interne consistentie kunnen berekend worden op de nieuw te verzamelen Nederlandse data voor de normering. De auteur van de Nederlandse versie moet een samenvatting van het relevant geachte buitenlandse onderzoek in de handleiding opnemen; een eenvoudige verwijzing is niet voldoende.

Wanneer de Nederlandse versie een duidelijke bewerking is van de oorspronkelijke buitenlandse versie, bijvoorbeeld omdat sommige items niet voldoen en zijn vervangen, omdat de test is uitgebreid, omdat de antwoordschaal is gewijzigd, of omdat aanwijzingen en instructies zijn gewijzigd, kunnen de onderzoeksgegevens die in het buitenland zijn verzameld niet meer in de plaats komen van Nederlands onderzoek. Als de meetpretentie niet is gewijzigd en globaal dezelfde begrippen worden gemeten, mag men wél aannemen dat dezelfde theoretische uitgangspunten aan het instrument ten grondslag liggen. Ook deze moeten echter in de Nederlandse handleiding worden beschreven.

In géén van bovenstaande gevallen kunnen normgegevens van het buitenlandse instrument naar de Nederlandse situatie worden gegeneraliseerd. Deze zullen hier opnieuw moeten worden verzameld.

Nederlandstalige Belgische tests

Er is in de Documentatie een aantal Nederlandstalige Belgische tests opgenomen. Deze oorsprong is steeds bij de beoordeling aangegeven. Deze tests zijn beoordeeld en in de Documentatie opgenomen, omdat ze in principe zonder vertaling of bewerking in Nederland zijn af te nemen. Uiteraard heeft de beoordeling voor gebruik in Nederland plaatsgevonden. Hierbij zijn de volgende regels gehanteerd:

- ***Uitgangspunten van de testconstructie.***
  De beoordeling op dit punt wijkt slechts dan af als de test is geconstrueerd voor een specifiek Belgische situatie, bijvoorbeeld een interessetest voor een schoolsysteem dat in Nederland niet bestaat. De beoordeling wordt dan 'onvoldoende'.
- ***Kwaliteit van het testmateriaal.***
  Het Vlaams kent andere woorden, uitdrukkingen en zinsconstructies dan het Nederlands. Aangezien de test voor gebruik in Nederland wordt beoordeeld, geldt als eis dat opgaven en instructie in gangbaar Nederlands zijn gesteld. Of hieraan wordt voldaan, kan met behulp van de vragen 2.6 en 2.14 worden beoordeeld.
- ***Normen.***
  Wanneer normen zijn verzameld met behulp van Nederlandse groepen, worden deze op de normale wijze beoordeeld. Normen die zijn gebaseerd op Belgische groepen worden als 'onvoldoende' beoordeeld, tenzij aannemelijk is gemaakt dat de scoreverdeling van deze groepen niet verschilt van die van vergelijkbare Nederlandse groepen.

Aan betrouwbaarheid en validiteit worden geen speciale eisen gesteld. Aangenomen wordt dat deze gegevens generaliseerbaar zijn naar de Nederlandse situatie, tenzij er expliciete redenen zijn om aan te nemen dat dit niet zo is.

In feite worden Nederlandstalige Belgische tests zo, althans gedeeltelijk, op dezelfde wijze behandeld als andere van oorsprong 'buitenlandse' tests. Deze beoordelingen zijn uiteraard niet geschikt voor een oordeel over de kwaliteit van de betreffende tests voor gebruik in België.

Samenvatting

Deze paragraaf geeft een puntsgewijze opsomming van de belangrijkste onderwerpen die in deze inleiding aan de orde zijn gekomen:

- De testbeoordeling betreft de volgende zeven criteria: 'Uitgangspunten van de testconstructie', 'Kwaliteit van het testmateriaal', 'Kwaliteit van de handleiding', 'Normen', 'Betrouwbaarheid', 'Begripsvaliditeit' en 'Criteriumvaliditeit'.
- De beoordeling op elk criterium kan 'onvoldoende', 'voldoende' of 'goed' zijn.
- De beoordelingen met deze herziene versie van het beoordelingssysteem zijn niet volledig vergelijkbaar met beoordelingen volgens het 'oude' systeem.
- De beoordeling 'onvoldoende' kan betekenen dat bepaalde informatie niet beschikbaar is.
- De COTAN garandeert volstrekte vertrouwelijkheid van het door de auteur/uitgever ter beschikking gestelde testmateriaal.
- Aan de COTAN-beoordelaar moet alle voor de beoordeling relevante informatie ter beschikking worden gesteld.
- Ook van bewerkingen of vertalingen van (al dan niet bekende) buitenlandse tests moet een beschrijving van de 'Uitgangspunten' in de handleiding zijn opgenomen.
- Generalisatie van onderzoeksbevindingen met buitenlandse tests naar de Nederlandse versie is alleen mogelijk als equivalentie van beide versies is aangetoond.
- Vlaams-Belgische tests worden op de criteria 'Uitgangspunten', 'Kwaliteit van het testmateriaal' en 'Normen' expliciet beoordeeld voor gebruik in de Nederlandse situatie; deze beoordelingen hebben geen geldigheid voor de Belgische situatie.

1 Uitgangspunten van de testconstructie

Testconstructie vergt een grondige voorbereiding. Men wil immers verantwoorde uitspraken doen over verschillen binnen personen (zoals bij leerlingvolgsystemen, waarbij verschillen in de tijd een rol spelen, of bij beroepskeuzebegeleiding bij een ipsatieve interessetest), tussen personen (zoals bij personeelsselectie), of tussen groepen van personen en/of situaties (zoals bij organisatie-onderzoek). Op basis van de informatie die de testauteur biedt moet de toekomstige gebruiker kunnen beoordelen of de test past bij het doel waarvoor hij een test zoekt. Er moet daarom een heldere omschrijving van de meetpretentie van de test worden gegeven en de keuze van de testinhoud en de wijze waarop de begrippen worden gemeten moeten omstandig worden verantwoord.

Bij dit criterium gaat het uitsluitend om de vraag of de uitgangspunten expliciet zijn aangegeven. Het gaat hier niet om de kwaliteit van de onderzoeksopzet en -uitvoering, want deze komen in de criteria 3 tot en met 7 aan de orde.

Aanwijzingen bij basisvraag 1.1: "Is er aangegeven wat het gebruiksdoel is van de test?"

a De meetpretentie van de test moet zijn aangegeven. Het moet dus duidelijk zijn welke constructen men met de test beoogt te meten. 'Construct' is hier breed bedoeld en kan dus ook verwijzen naar een domein van specifieke gedragingen, voorkeuren of stijlen; van belang is dat goed is gedefinieerd welke gedragingen etc. tot het te meten domein behoren. Het te meten construct kan dus bijvoorbeeld zijn: intelligentie, leesvaardigheid, prestatiemotivatie, beroepeninteresse of ADHD.

b Er moet zijn aangegeven voor welke doelgroep of -groepen de test is bedoeld, bijvoorbeeld met betrekking tot leeftijd, beroep, opleidingsniveau, relevante voorkennis of normaal versus klinisch. Hierbij geldt: hoe uitgebreider de pretentie van brede toepasbaarheid, des te groter de verplichtingen ten aanzien van het te leveren empirische materiaal, zoals normen of valideringsgegevens.

c Testconstructie begint met een bezinning op het gebruiksdoel. Wil men criteriumgedrag voorspellen? Is een test bedoeld voor voortgangscontrole of trainingsevaluatie? Gaat het om een niveaubepaling voor de plaatsing van leerlingen? Gaat het om diagnose voor een behandelingsplan?

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

Sluit de test aan bij een bestaande theorie of ontwikkelt de auteur een eigen theorie? Wordt deze theorie voldoende beschreven? Wanneer de test een vertaling/bewerking is van een buitenlands instrument, moet een beschrijving worden gegeven van de achtergronden van dat instrument en kan niet worden volstaan met een simpele literatuurverwijzing. Ook bij tests die algemeen bekende begrippen meten, zoals intelligentie, moet een omschrijving van het begrip worden gegeven, zodat duidelijk wordt wat wel en wat niet tot het te meten domein wordt gerekend. Als de test niet zozeer theoretisch maar eerder historisch is gefundeerd, dus aansluit bij een traditionele wijze van meten van een bepaald begrip, moet duidelijk worden gemaakt waarom juist de betreffende begrippen worden gemeten en wat de verschillen en overeenkomsten zijn met soortgelijke tests. Wat is de meerwaarde van het nieuwe instrument ten opzichte van bestaande instrumenten? Wanneer de test een variant is op al bestaande instrumenten of een bewerking voor computerafname van een papier-en-potloodversie, wordt er dan aangegeven of er verschillen zijn tussen beide versies en zo ja, wat deze verschillen zijn?

Vragen voor criterium 1
Uitgangspunten van de testconstructie

		onv.	vold.	goed
Basisvraag 1.1	Is er aangegeven wat het gebruiksdoel is van de test?			
	a. Is er aangegeven welk(e) construct(en) de test beoogt te meten?	1	2	3
	b. Is er aangegeven wat de doelgroep(en) is (zijn) van de test?	1	2	3
	c. Is er aangegeven wat de functie is van de test?	1	2	3
	Bij negatieve beoordeling (1) op een of meer van de subvragen a t/m c kan men de rest van de vragen van dit criterium overslaan en doorgaan met criterium 2.			
1.2	Is de herkomst van het constructie-idee beschreven en/of worden de te meten constructen gedefinieerd?	1	2	3
1.3	Wordt de relevantie van de testinhoud voor de te meten construct(en) aannemelijk gemaakt?	1	2	3

Aanwijzingen bij vraag 1.3: "Wordt de relevantie van de testinhoud voor de te meten construct(en) aannemelijk gemaakt?"

Bij deze vraag gaat het om de stap van meetpretentie naar operationalisatie. Is er daartoe een zodanige omschrijving van het itemdomein beschikbaar dat duidelijk is of een willekeurig item wel of niet tot de test zou kunnen behoren? Worden de te meten constructen op zodanige wijze (bijvoorbeeld met behulp van facetanalyse) geanalyseerd dat duidelijk wordt welke aspecten binnen de constructen kunnen worden onderscheiden? Worden eventueel, op basis van theoretische of inhoudelijke overwegingen, verschillende gewichten aan deze aspecten toegekend en wordt hiermee bij de keuze van de items rekening gehouden? Als bij de constructie of bewerking van een test items zijn afgevallen of gewijzigd, wordt dan aangegeven wat de gevolgen hiervan zijn voor de meting van het oorspronkelijk bedoelde begrip? Dus: is het inhoudsdomein nog volledig gedekt of is het vernauwd of verschoven?

Wordt bij adaptieve tests (waarbij de keuze van de items die aan de geteste persoon worden voorgelegd mede bepaald wordt door diens antwoordpatroon op de tot dan toe beantwoorde items) aangegeven hoe de testinhoud wordt gegarandeerd? Bij adaptieve tests krijgt elke kandidaat immers andere items gepresenteerd, waardoor het mogelijk is dat bepaalde onderwerpen onvoldoende in de test naar voren komen. Om die reden is het veelal noodzakelijk om een inhoudscontrole uit te voeren (bijvoorbeeld volgens de methode zoals voorgesteld door Kingsbury & Zara, 1991), zodat elke test overeenkomt met de specificatietabel.

Vaststelling eindoordeel voor criterium 1 Uitgangspunten van de testconstructie		
De som van de beoordelingen op de vragen 1.1.a t/m 1.1.c is 8 of 9.	Beide andere vragen worden minstens met '2' beoordeeld.	goed
	Een van beide andere vragen met '2' en de ander wordt met '1' beoordeeld.	voldoende
	Beide andere vragen worden met '1' beoordeeld.	onvoldoende
De som van de beoordelingen op de vragen 1.1.a t/m 1.1.c is 6 of 7 en geen van deze vragen wordt met '1' beoordeeld.	Beide andere vragen worden minstens met '2' beoordeeld.	voldoende
	Een of beide andere vragen wordt met '1' beoordeeld.	onvoldoende
Een of meer van de vragen 1.1.a t/m 1.1.c wordt met '1' beoordeeld.		onvoldoende

2 De kwaliteit van het testmateriaal

Bij de beoordeling van dit criterium wordt onderscheid gemaakt tussen tests die met behulp van papier-en-potlood of met behulp van de computer (Computer Based Tests = CBT) worden afgenomen. Bij deze laatste afnamewijze wordt geen onderscheid gemaakt tussen tests die via een lokale testomgeving of via het internet worden afgenomen, omdat de eisen in principe dezelfde zijn.

Er worden bij dit criterium, voor beide afnamewijzen, drie basisvragen gesteld. In het algemeen geldt dat om de score(s) op een test zinvol te kunnen interpreteren, de test zodanig moet zijn afgenomen en gescoord dat andere, niet-beoogde factoren geen invloed kunnen uitoefenen op de totstandkoming van de score(s). Zo moeten de afname en instructie dusdanig zijn gestandaardiseerd dat de invloed van variatie in instructie of verschil in proefleider (bij een papieren versie), van de afnamesituatie of van de beslisregel (bij een adaptieve test) op de testscore is geëlimineerd of in ieder geval binnen de grenzen van het mogelijke is beperkt. Ook moet de scoring zo objectief mogelijk zijn.

De score 2 voor de derde basisvraag naar mogelijke racistische of voor bepaalde bevolkingsgroepen kwetsende inhoud kan leiden tot een aparte aantekening bij de beoordeling, bijvoorbeeld: "In sterk beperkte mate toepasbaar voor allochtonen" (zie Hofstee e.a., 1990).

Voor een instrument dat is bedoeld voor papier-en-potloodafname moet men beginnen met vraag 2.1, voor een instrument dat via de computer wordt afgenomen met vraag 2.9. Als een instrument in zowel papier-en-potloodversie als CBT-vorm bestaat, moet de kwaliteit van het testmateriaal van beide vormen worden beoordeeld. Bij uiteenlopende beoordelingen kan dit met een voetnoot bij de beoordeling worden aangegeven. Hierbij wordt ervan uitgegaan dat de items en, voor zover mogelijk, de instructie van beide versies identiek zijn. Is dit niet het geval, dan heeft men in feite met twee verschillende tests te maken en moet de hele beoordeling voor elke versie apart worden verricht. De psychometrische gegevens zijn dan immers in ieder geval niet meer over versies generaliseerbaar.

Vragen voor criterium 2
Kwaliteit van het testmateriaal
Papier-en-potloodversie

		onv.	vold.	goed
Basisvraag 2.1	Zijn de testopgaven gestandaardiseerd? Bij negatieve beoordeling (1) van deze vraag kan men de rest van de vragen van dit criterium overslaan en doorgaan met criterium 3.	1		3
Basisvraag 2.2	a. Is er sprake van een objectief scoringssysteem, en/of, b. als de scoring door beoordelaars of observatoren gebeurt, is dan het beoordelings- of observatiesysteem volledig en duidelijk? Bij negatieve beoordeling (1) van deze vraag kan men de rest van de vragen van dit criterium overslaan en doorgaan met criterium 3.	1 1	2 2	3 3
Basisvraag 2.3	Zijn de items vrij van racistische, etnocentrische, seksistische en voor bepaalde bevolkingsgroepen kwetsende inhoud? Bij negatieve beoordeling (1) van deze vraag kan men de rest van de vragen van dit criterium overslaan en doorgaan met criterium 3.	1	2	3
2.4	Zijn items, testboekje, antwoordschalen en/of antwoordformulier zodanig ontworpen dat fouten bij de *invulling* kunnen worden vermeden?	1	2	3
2.5	Is de instructie voor de geteste volledig en duidelijk?	1	2	3
2.6	Zijn de items correct geformuleerd?	1	2	3
2.7	Hoe is de kwaliteit van het testmateriaal?	1	2	3
2.8	Is het scoringssysteem zodanig ontworpen en beschreven dat fouten bij de *scoring* kunnen worden vermeden?	1	2	3

Vragen voor criterium 2 Kwaliteit van het testmateriaal Afname via computer		onv.	vold.	goed
Basisvraag 2.9	Is de test gestandaardiseerd of worden bij adaptieve tests beslisregels geëxpliciteerd? Bij negatieve beoordeling (1) van deze vraag kan men de rest van de vragen van dit criterium overslaan en doorgaan met criterium 3.	1		3
Basisvraag 2.10	Is er sprake van een geautomatiseerd of objectief scoringssysteem? Bij negatieve beoordeling (1) van deze vraag kan men de rest van de vragen van dit criterium overslaan en doorgaan met criterium 3.	1	2	3
Basisvraag 2.11	Zijn de items vrij van racistische, etnocentrische, seksistische en voor bepaalde bevolkingsgroepen kwetsende inhoud? Bij negatieve beoordeling (1) van deze vraag kan men de rest van de vragen van dit criterium overslaan en doorgaan met criterium 3.	1	2	3
2.12	Is de software zodanig ontworpen dat fouten door onjuist gebruik kunnen worden vermeden?	1	2	3
2.13	Is de instructie voor de *geteste* volledig en duidelijk?	1	2	3
2.14	Zijn de items correct geformuleerd?	1	2	3
2.15	Hoe is de kwaliteit van de vormgeving van de gebruikersinterface?	1	2	3
2.16	Is de test voldoende beveiligd?	1	2	3

Papier-en-potloodversie

Aanwijzingen bij basisvraag 2.1: "Zijn de testopgaven gestandaardiseerd?"

Testopgaven zijn gestandaardiseerd als de opgaven wat betreft inhoud, vorm en volgorde voor iedereen hetzelfde zijn. Dit is belangrijk als men scores wil kunnen interpreteren en vergelijken. Er wordt een uitzondering gemaakt voor de eis van een uniforme inhoud en volgorde van testitems voor adaptieve tests (zie vraag 2.9). Hoewel adaptieve tests vrijwel altijd in CBT-vorm voorkomen, hebben sommige papier-en-potloodtests ook adaptieve kenmerken, bijvoorbeeld met instap- en afbreekregels.

Aanwijzingen bij basisvraag 2.2.a: "Is er sprake van een objectief scoringssysteem?"

Onder een objectief scoringssysteem wordt verstaan dat waarden die aan alle mogelijke antwoorden van proefpersonen/cliënten worden toegekend bij voorbaat zodanig vastliggen, dat elke testleider, afgezien van administratieve fouten die bij de handmatige of geautomatiseerde scoring kunnen worden gemaakt, tot dezelfde score zal komen. Dit geldt vooral voor schriftelijke capaciteitentests en vragenlijsten met meerkeuze-items.

Aanwijzingen bij basisvraag 2.2.b: "Als de scoring door beoordelaars of observatoren gebeurt, is dan het beoordelings- of observatiesysteem volledig en duidelijk?"

Bij bijvoorbeeld observatieschalen, projectieve tests, onderdelen van individueel af te nemen intelligentietests met open vragen en essayvragen is er geen strikt objectieve beoordeling mogelijk. Wel moeten er procedures zijn beschreven waardoor de objectiviteit zo goed mogelijk wordt gewaarborgd. Dit betekent dat er richtlijnen moeten worden gegeven voor de beoordeling/scoring, waaronder modelantwoorden, modelgedragingen, schaalankers, wegingsvoorschriften en dergelijke. Hieruit moet duidelijk worden wat er per se in een antwoord moet staan of welk gedrag moet zijn vertoond om bepaalde scores te kunnen toekennen. Indien van toepassing moeten ook de aard en de inhoud van de training die aan beoordelaars wordt gegeven, zijn omschreven.

Aanwijzingen bij basisvraag 2.3: "Zijn de items vrij van racistische, etnocentrische, seksistische en voor bepaalde bevolkingsgroepen kwetsende inhoud?"

In 1990 hebben het Landelijk Bureau Racismebestrijding en het NIP een commissie ingesteld die een twintigtal tests heeft gescreend op mogelijk racistische inhoud (Hofstee e.a., 1990). Hoewel in geen van de gescreende tests racistische inhoud werd aangetroffen, werden vooral bij de verbale tests wel etnocentrismen en niet noodzakelijke idiomatische uitdrukkingen aangetroffen. Aan tests waarin dit voorkwam werd door genoemde commissie de kwalificatie "in sterk beperkte mate toepasbaar voor allochtonen" verbonden. In dit beoordelingssysteem wordt deze strategie overgenomen. Het principe van beperkte toepasbaarheid kan ook voor andere groepen gelden (bijvoorbeeld een interessetest die met afbeeldingen werkt waarop alleen mannen zijn afgebeeld). In dergelijke gevallen moet bij deze vraag de score '2' worden toegekend. Expliciet racistische (of seksistische) inhoud leidt overigens tot een voor elke groep onbruikbare test (score '1'), behalve als deze inhoud deel uitmaakt van de meetpretentie (bijvoorbeeld de F-schaal of een androgynieschaal).

Met behulp van deze vraag wordt niet gevraagd te beoordelen óf er biasonderzoek is verricht (dat komt bij criterium 6 aan de orde), noch om de items op bias te beoordelen; het gaat hier louter om de a priori bruikbaarheid van een test voor een bepaalde groep.

Aanwijzingen bij vraag 2.4: "Zijn items, testboekje, antwoordschalen en/of antwoordformulier zodanig ontworpen dat fouten bij de invulling kunnen worden vermeden?"

Punten waarop gelet moet worden bij beoordeling van deze vraag zijn:

- de vragen/opgaven moeten begrijpelijk zijn (en bijvoorbeeld niet te moeilijk) voor de groep waarvoor de test is bedoeld;
- wanneer van aparte antwoordformulieren gebruik wordt gemaakt, moeten deze zo zijn ontworpen dat vergissingen (zoals een item overslaan) kunnen worden voorkomen en dat eventuele vergissingen door de geteste snel kunnen worden ontdekt.

Aanwijzingen bij vraag 2.5: "Is de instructie voor de geteste volledig en duidelijk?"

Binnen de instructie wordt onderscheid gemaakt tussen de instructie voor de geteste en de instructie voor de testleider. De kwaliteit van de instructies voor de geteste wordt in deze vraag beoordeeld; over de kwaliteit van de instructies voor de testleider wordt in vraag 3.2 een oordeel gevraagd.
De instructies of aanwijzingen voor de geteste zijn een onderdeel van het testmateriaal en vormen in het algemeen de eerste bladzijde(n) van het testboekje. De instructie moet zijn gestandaardiseerd en in gangbaar Nederlands zijn gesteld. De volgende aspecten moeten in de instructie zijn opgenomen:

- voorbeeldvragen;
- informatie over hoe antwoorden gegeven of genoteerd moeten worden;
- de te volgen strategie bij het niet-weten van het goede antwoord of bij alternatieven die even (on)aantrekkelijk of in gelijke mate van toepassing zijn;
- de beschikbare tijd.

Indien van toepassing moet in de instructie ook informatie worden gegeven over de anonimiteit van de testresultaten. Zo mogelijk moeten er oefenopgaven worden verstrekt voor personen die geen ervaring hebben met tests.

Aanwijzingen bij vraag 2.6: "Zijn de items correct geformuleerd?"

In de literatuur over toets- en vragenlijstconstructie treft men velerlei voorschriften aan voor de formulering van items. Hieronder volgt een – overigens niet-uitputtende – opsomming van regels waarop men bij het beoordelen van de items moet letten (grotendeels ontleend aan Erkens & Moelands, 1992, en Moelands, Noijons & Rem, 1992). Waar van toepassing, gelden de onderstaande voorschriften ook voor CBT-tests, zie vraag 2.14.

Open vragen

- Is het item grammaticaal juist geformuleerd?
- Bevat het item een te ingewikkelde zinsconstructie?
- Bevat het item onnodig moeilijke woorden?
- Bevat het item onnodige tussenvoegsels?
- Is het item onnodig negatief gesteld?
- Kan de formulering van het item aanleiding geven tot misverstanden?
- Bestaat er gevaar dat het item door klemtoonverschuiving duidelijk van betekenis verandert?
- Bevat het item voldoende informatie om het goede antwoord te kunnen geven?
- Geeft het item voldoende informatie over de gewenste lengte en vorm van het antwoord?
- Weet de geteste of een antwoord gemotiveerd moet worden?
- Zijn de informatie en de probleemstelling duidelijk te onderscheiden?

Gesloten vragen

- Zijn er misschien meerdere goede antwoorden?
- Wordt er niet naar twee of meer dingen tegelijk gevraagd?
- Bevat het item onduidelijkheden?
- Bevatten de alternatieven onduidelijkheden?
- Bevat de stam een duidelijke vraag of opdracht?
- Bevat de stam voldoende informatie om het item te kunnen beantwoorden?
- Is het niet mogelijk uit de itemkenmerken af te leiden wat het goede antwoord is?
- Bevat de stam geen overbodige informatie?
- Is de stam precies, beknopt en grammaticaal juist geformuleerd?
- Bevat de stam geen dubbele ontkenning?
- Als de stam een ontkenning bevat, is dat dan duidelijk zichtbaar gemaakt?
- Hebben alle afleiders enige plausibiliteit?

- Wordt er in het goede alternatief niet een term uit de stam herhaald?
- Staan er geen woorden als 'altijd' of 'nooit' in de afleiders?
- Ontstaat er geen dubbele ontkenning bij combinatie van de stam en een of meer alternatieven?
- Sluiten de alternatieven elkaar uit?
- Sluiten de alternatieven grammaticaal en inhoudelijk goed aan op de stam?
- Bevatten de alternatieven geen herhalingen uit de stam of van elkaar?
- Zijn de alternatieven logisch gerangschikt?
- Zijn de alternatieven voldoende van elkaar te onderscheiden?

Aanwijzingen bij vraag 2.7: "Hoe is de kwaliteit van het testmateriaal?"
Het gaat hierbij uitsluitend om allerlei praktische aspecten die men niet bij een van de andere vragen van dit criterium kan beoordelen, zoals:
- Is de tekst goed leesbaar?
- Is het test- of vragenlijstboekje overzichtelijk? Als er andere materialen (blokjes, apparaten, etc.) worden gebruikt: zijn deze hanteerbaar en functioneel?
- Is het kleurgebruik 'prettig' en functioneel (zie toelichting vraag 2.15, vierde aandachtspunt)?
- Zijn kleuren of symbolen (indien van toepassing) goed van elkaar te onderscheiden (vooral van belang voor kleurenblinden)?
- Is het testmateriaal duurzaam?

Aanwijzingen bij vraag 2.8: "Is het scoringssysteem zodanig ontworpen en beschreven dat fouten bij de *scoring* kunnen worden vermeden?"
Bij deze vraag moet onder andere worden gelet op de volgende punten:
- De scoringsprocedure moet duidelijk zijn omschreven.
- Als er scoringsmallen worden gebruikt, moet er zijn aangegeven hoe deze op de antwoordformulieren moeten worden gelegd; de mallen moeten bovendien goed passen op de antwoordformulieren.
- Als er scoringsmallen worden gebruikt, moet er op de mallen zijn aangegeven bij welke versie van de test ze horen (dit is vooral van belang wanneer de test is herzien).
- Er moet worden vermeld welke score aan overgeslagen items moet worden toegekend.
- Er moet worden aangegeven hoeveel items mogen worden overgeslagen zonder dat de test zijn waarde verliest.
- Als de test van beoordelaars/observatoren gebruikmaakt, moet zijn aangegeven hoe men met verschillen tussen beoordelaars/observatoren moet omgaan.

Over het algemeen geldt dat een apart antwoordformulier te verkiezen is boven het scoren van verschillende bladzijden in een testboekje, om mogelijke fouten in de scoring te voorkomen.
NB Bij tests die met behulp van papier-en-potlood worden afgenomen maar op een computer worden gescoord, moet de COTAN-beoordelaar de scoring kunnen controleren (zie vraag 2.10).

Afname per computer

Aanwijzingen bij basisvraag 2.9: "Is de test gestandaardiseerd of worden bij adaptieve tests beslisregels geëxpliciteerd?"
Bij een testafname via de computer gelden dezelfde eisen van standaardisatie waar het inhoud en vorm van de items betreft. Bij zulke testafnames verdient de standaardisatie van de testtijd extra aandacht: het is van belang dat de tijd die voor een item of voor de gehele test beschikbaar is niet afhankelijk is van het systeem waarop de applicatie draait.

Hoewel de standaardisatie-eis (zie vraag 2.1) in principe voor alle tests geldt, wordt hierop wat betreft iteminhoud en itemvolgorde een uitzondering gemaakt voor adaptieve tests. De theorie achter adaptief testen gaat er immers van uit dat een efficiëntere schatting van het vaardigheidsniveau van de geteste kan worden verkregen als het aanbod van items steeds wordt aangepast aan de antwoorden van de geteste op voorafgaande items. Bij dit type tests moeten echter de beslisregels of de algoritmes voor de samenstelling van de test zijn geëxpliciteerd. Hoe wordt de test gestart? Hoe wordt de keuze voor een volgend item gemaakt? En wanneer wordt de test beëindigd? Indien óf de startprocedure, óf de selectieprocedure, óf de stopprocedure niet is beschreven, dan is het oordeel op deze vraag 'onvoldoende' (1). Het oordeel 'goed' (3) kan alleen worden toegekend als de keuze voor een algoritme is onderbouwd en als de voor- en nadelen van de keuze uiteen zijn gezet.

Aanwijzingen bij basisvraag 2.10: "Is er sprake van een geautomatiseerd of objectief scoringssysteem?
Onder een objectief scoringssysteem wordt verstaan dat waarden die aan alle mogelijke antwoorden van personen worden toegekend, bij voorbaat zodanig vastliggen dat elke testleider, afgezien van administratieve fouten die bij de scoring kunnen worden gemaakt, tot dezelfde score zal komen. Als de scoring volledig is geautomatiseerd, dan is het scoringssysteem per definitie objectief. Het oordeel op deze vraag is in dit geval 'goed' (3). Dit oordeel kan echter alleen worden gegeven als de COTAN-beoordelaar over voldoende gegevens beschikt om de juistheid van de scoring te kunnen controleren. Onder 'scoring' wordt in dit geval verstaan: het toekennen van een score aan de items, het sommeren van de itemscores per (sub)test of (sub)schaal (eventueel met gebruikmaking van itemgewichten) en het omzetten van deze somscores in normscores met gebruikmaking van een normtabel. Deze controle-eis kan betekenen dat de testauteur voor de beoordeling extra informatie moet aanleveren die niet in de handleiding is opgenomen (sleutels, gewichten, normtabellen).

Ook als de testgebruiker niet over bovengenoemde extra informatie beschikt, is het wel wenselijk dat hij toegang heeft tot informatie die nuttig is om de resultaten van een geteste te kunnen interpreteren. Het gaat hierbij vooral om ruwe test- of schaalscores. Als deze niet in een automatisch gegenereerd rapport vermeld staan en ook in de applicatie niet anderszins opvraagbaar zijn, kan het oordeel op deze vraag niet hoger worden dan 'voldoende'.

Worden enkele of alle items handmatig gescoord, dan moeten er bij tests met open items antwoordmodellen, scoringsvoorschriften en een beoordelaarsinstructie worden meegeleverd. Hierdoor moet duidelijk worden wat er in een antwoord moet staan of welk gedrag moet zijn vertoond om bepaalde scores te kunnen toekennen.
Bij een test met gesloten items is alleen een scoringsvoorschrift noodzakelijk. Als er niets meegeleverd wordt, dan is het oordeel op deze vraag 'onvoldoende' (1). In andere gevallen ligt het vooral aan de volledigheid en duidelijkheid van het meegeleverde materiaal of het oordeel 'voldoende' of 'goed' moet worden gegeven.

Aanwijzingen bij basisvraag 2.11: "Zijn de items vrij van racistische, etnocentrische, seksistische en voor bepaalde bevolkingsgroepen kwetsende inhoud?"
Zie de toelichting bij vraag 2.3.

Aanwijzingen bij vraag 2.12: "Is de software zodanig ontworpen dat fouten door onjuist gebruik kunnen worden vermeden?"
Het mag niet kunnen gebeuren dat testresultaten negatief worden beïnvloed doordat een kandidaat de CBT-software onjuist gebruikt. Naast het aanbieden van een begrijpelijke instructie zijn er diverse manieren om 'fouten' door onjuist gebruik van de CBT-software te vermijden. Bij deze vraag is het van belang of de testauteur voldoende heeft gedaan om de kans op fouten door onjuist gebruik te minimaliseren. Hierbij kunnen diverse voorzorgsmaatregelen belangrijk zijn:

- het uitschakelen van overbodige functies en sneltoetsen;
- het afsluiten van de toegang tot de harde schijf;
- het onmogelijk maken andere (niet-bedoelde) software op te starten;
- het moeilijk maken de CBT-software voortijdig of zonder opslaan te verlaten.

Bij tests die via internet worden afgenomen en waarbij er gebruik-gemaakt wordt van een browser (*Internet Explorer, Firefox, Safari,* etc.) die de items aanbiedt en de antwoorden doorgeeft aan de server, is het veelal niet mogelijk de computer van de cliënt op bovengenoemde punten te beïnvloeden. In dat geval moet in de testhandleiding staan welke voorzorgen de testgebruiker moet nemen.

De vormgeving van de interface is ook van invloed op de kans op fouten. Bij deze vraag wordt niet gevraagd om te beoordelen of de gebruikersinterface naar behoren is vormgegeven, maar de vormgeving mag wel in overweging worden genomen bij het beoordelen van deze vraag. Als de gebruikersinterface dusdanig is vormgegeven (bijvoorbeeld extreem veel navigatiemogelijkheden, nagenoeg onleesbare teksten of een niet te begrijpen indeling), dan moet het oordeel 'onvoldoende' (1) worden toegekend. Als er bij het proberen van de CBT-software geen onoverkomelijke problemen optreden en de CBT-software reageert zoals verwacht, dan moet het oordeel 'voldoende' (2) worden toegekend. Het oordeel 'goed' (3) mag alleen worden toegekend als het ook daadwerkelijk lastig is om andere (niet bedoelde) software op te starten, niet-bedoelde toetsen of toetscombinaties te gebruiken of de CBT-software zonder opslaan te verlaten.

Bij een computergestuurde testafname, of dit nu een stand-alone-, een netwerk- of een internetapplicatie is, kan het overigens altijd voorkomen dat de testafname wordt onderbroken door een technische oorzaak waar noch de geteste, noch de CBT-software debet aan zijn. In zo'n geval moet een doorstart mogelijk zijn, waarbij de applicatie na een identificatie en een eventuele herhaling van de testinstructie de test bij het juiste item vervolgt met, indien van toepassing, inachtneming van de nog beschikbare testtijd.

Van de COTAN-beoordelaar wordt niet verwacht dat hij zelf een uitputtende controle op bovenstaande aspecten uitvoert. Wel moet hij beoordelen of de testauteur in de handleiding concrete verantwoording heeft afgelegd over de getroffen voorzorgen en over de wijze waarop deze in de praktijk zijn getoetst.

Aanwijzingen bij vraag 2.13: "Is de instructie voor de *geteste* volledig en duidelijk?"
Een duidelijke en volledige instructie is belangrijk opdat degene die de test maakt geen 'fouten' kan maken doordat hij niet weet hoe de CBT-software werkt. De volgende aspecten moeten in de instructie zijn opgenomen:

- voorbeeldvragen;
- de werking van de software (waaronder de wijze van antwoorden);
- de te volgen strategie bij het niet-weten van het goede antwoord of bij alternatieven die even (on)aantrekkelijk of in gelijke mate van toepassing zijn;
- de beschikbare tijd, per test of per item.

Tevens is van belang:

- als het om een adaptieve test gaat, moet er een uitleg over adaptief testen worden gegeven;
- indien van toepassing moet informatie worden gegeven over de anonimiteit van de testresultaten;
- voor personen die geen ervaring hebben met het betreffende type test, moeten er zo mogelijk oefenopgaven worden verstrekt die de cliënt daadwerkelijk moet maken en waarop hij feedback krijgt.

Een onduidelijke en onvolledige instructie of een te uitgebreide instructie (bijvoorbeeld wanneer er bij elk item instructie wordt gegeven over hoe het item moet worden beantwoord) leidt op deze vraag tot het oordeel 'onvoldoende' (1). Het oordeel 'goed' (3) mag alleen worden toegekend als ook tijdens het maken van de test de instructie kan worden geraadpleegd.

Aanwijzingen bij vraag 2.14: "Zijn de items correct geformuleerd?"
Zie de toelichting bij vraag 2.6. Daarnaast is van belang op te merken dat ook voor tests die via de computer worden afgenomen – en vooral voor adaptieve tests – geldt dat de COTAN-beoordelaar alle items moet kunnen bekijken. Dit kan betekenen dat de testauteur – alleen voor de beoordeling – een overzicht van alle items moet aanleveren.

Aanwijzingen bij vraag 2.15: "Hoe is de kwaliteit van de vormgeving van de gebruikersinterface?"

Hieronder worden aspecten genoemd waarop bij het beoordelen van de gebruikersinterface moet worden gelet. Deze aspecten moeten worden beoordeeld voor de aanbevolen standaardinstallatie en computeromgeving. Negatieve beoordeling van een van de genoemde aspecten kan al leiden tot het oordeel 'onvoldoende' (1), als deze de gebruikswaarde van het instrument ernstig beperken.

- Is de schermvormgeving consistent? Het gaat om de volgende kenmerken van de schermomgeving:
  - symbolen moeten steeds dezelfde functie hebben;
  - kleuren moeten consistent worden gebruikt en moeten steeds dezelfde functie hebben;
  - informatie (items, instructie, antwoordveld, etc.) moet steeds op dezelfde locatie weergegeven worden of er moet steeds op dezelfde manier onderscheid zijn gemaakt tussen soorten informatie;
  - er moet consistent gebruik zijn gemaakt van lettertypes en
  - grootte s.
- Is de schermindeling overzichtelijk? De overzichtelijkheid van een scherm wordt bepaald door verschillende factoren:
  - Zijn de verschillende typen informatie (instructie, item, antwoordveld, etc.) duidelijk van elkaar te onderscheiden?
  - Zijn de buttons duidelijk herkenbaar en is de functie van de buttons altijd duidelijk? Bijvoorbeeld bij de button <afsluiten>: wordt in dit geval de test afgesloten of alleen de instructie?
  - Zijn de items en de instructie zonder scrollen leesbaar?
  - Is bepaalde informatie (bijvoorbeeld de instructie) gemakkelijk te vinden?
  - Is altijd duidelijk waar men zich bevindt of welke handeling men moet verrichten om te komen waar men wil?

  Er moet bij dit aspect worden gecontroleerd of de bediening van de test zo vanzelfsprekend is dat iemand met geen enkele computerervaring toch in staat is om de test te maken zonder dat er bijvoorbeeld sprake is van testbias.
- Is de informatie op het scherm leesbaar? De leesbaarheid wordt bevorderd indien:
  - niet meer dan twee lettertypes worden gebruikt;
  - niet meer dan drie puntgroottes worden gebruikt;
  - woorden niet cursief worden afgebeeld;
  - woorden niet worden onderstreept als er geen sprake is van een hyperlink.
- Is het kleurgebruik 'prettig' en functioneel? Van belang is dat kleur op een dusdanige manier is toegepast dat het de overzichtelijkheid en leesbaarheid van het beeldscherm bevordert. Functioneel kleurgebruik betekent dat kleuren een bepaalde betekenis hebben of dat het scherm overzichtelijker wordt, door bijvoorbeeld de items of het antwoordveld een afwijkende kleur te geven. Het is zeker niet wenselijk om een groot aantal kleuren te gebruiken of om kleuren zonder enige reden toe te passen. Met 'prettig' kleurgebruik wordt de keuze voor bepaalde kleurencombinaties of het contrast tussen kleurnuances bedoeld. Bepaalde kleurencombinaties en slecht contrasterende kleuren zijn bijvoorbeeld moeilijk te onderscheiden. Bij het gebruik van kleuren moet er ook rekening mee zijn gehouden dat de test in het algemeen ook geschikt moet zijn voor kleurenblinden en dat het kleurgebruik voor deze groep geen nadelige gevolgen mag opleveren.
- Is het beeld- en geluidsmateriaal functioneel? Onder 'beeldmateriaal' wordt in dit verband verstaan: al het mogelijke beeldmateriaal zoals animaties, filmfragmenten en statische afbeeldingen. Van belang is dat zowel het beeldmateriaal als de geluidsfragmenten een duidelijke functie hebben en dat ze niet zijn opgenomen om de CBT-software te 'verfraaien'. Hierbij moet worden aangetekend dat de functionaliteit van het beeld- en geluidsmateriaal al in het geding is als het slecht leesbaar of verstaanbaar is.

Aanwijzingen bij vraag 2.16: "Is de test voldoende beveiligd?"

Een test is 'goed' beveiligd als al het mogelijke is gedaan om de toegang tot de test, het testmateriaal en de testresultaten te beveiligen:

- De beveiliging van de *toegang tot de test* is van belang om zeker te weten dat degene die de test maakt ook degene is die de test zou moeten maken. Een vorm van legitimatie is daarom belangrijk. Mogelijkheden zijn onder andere het gebruik van *passwords* en *usernames*, een verplichte legitimatie door middel van een identiteitskaart of rijbewijs aan de testleider of het gebruik van webcams.
- De beveiliging van het *testmateriaal* is ten eerste belangrijk omdat het uit het oogpunt van de validiteit niet wenselijk is dat degenen die de test maken de mogelijkheid hebben om items, informatie over de algoritmes of scoringsvoorschriften te kopiëren naar een andere computer of printer. Ten tweede is het belangrijk dat er geen informatie over de items gemakkelijk te verkrijgen is. Daarom zouden in het geval dat de items zijn opgenomen in een itembank, alleen geautoriseerden toegang moeten kunnen krijgen tot de itembank. Bij adaptieve tests kunnen items ook bekend raken doordat het ene item misschien veel vaker in de test opgenomen wordt dan een ander item. Daarom is het in sommige gevallen belangrijk dat de testauteur een mechanisme (bijvoorbeeld volgens de *Sympson-Hettermethode*, 1985; zie ook Stocking & Swanson, 1993) inbouwt waardoor op mogelijke over- of onderbenutting van de items wordt gecontroleerd.
- De beveiliging van de *testresultaten* is belangrijk om misbruik (bijvoorbeeld het ongeoorloofd aanbrengen van wijzigingen in de resultaten) te voorkomen en om de privacy en anonimiteit van de geteste voldoende te kunnen waarborgen.

Voor het oordeel 'goed' (3) moet de handleiding concrete informatie bevatten waaruit blijkt dat al het mogelijke is gedaan om alle drie bovengenoemde aspecten te beveiligen. Het oordeel is 'onvoldoende' (1) als hierover geen informatie wordt gegeven of als uit de gegeven informatie blijkt dat de beveiliging op een of meer van deze aspecten niet is geregeld. Het oordeel 'voldoende' (2) wordt gegeven als aan de beveiliging op alle drie de aspecten wel aandacht is besteed, maar deze technisch en/of procedureel voor verbetering vatbaar zijn.

Vaststelling eindoordeel voor criterium 2 Kwaliteit van het testmateriaal Papier-en-potloodversie		
Alle drie de basisvragen worden met '3' beoordeeld.	Somscore 2.4 t/m 2.8 ≥ 11	goed
	Somscore 2.4 t/m 2.8 = 9 of 10	voldoende
	Somscore 2.4 t/m 2.8 ≤ 8	onvoldoende
Basisvraag 2.2* en/of 2.3 wordt met '2' beoordeeld en de andere basisvragen worden niet met '1' beoordeeld.	Somscore 2.4 t/m 2.8 ≥ 11	voldoende
	Somscore 2.4 t/m 2.8 ≤ 10	onvoldoende
Minstens een van de drie basisvragen wordt met '1' beoordeeld.		onvoldoende
* Bij basisvraag 2.2 kunnen beide subvragen van toepassing zijn; in dat geval geeft de laagste beoordeling de doorslag.		

Vaststelling eindoordeel voor criterium 2 Kwaliteit van het testmateriaal Afname via computer		
Alle drie de basisvragen worden met '3' beoordeeld.	Somscore 2.12 t/m 2.16 ≥ 11	goed
	Somscore 2.12 t/m 2.16 = 9 of 10	voldoende
	Somscore 2.12 t/m 2.16 ≤ 8	onvoldoende
Basisvraag 2.10 en/of 2.11 wordt met '2' beoordeeld en de andere basisvragen worden niet met '1' beoordeeld.	Somscore 2.12 t/m 2.16 ≥ 11	voldoende
	Somscore 2.12 t/m 2.16 ≤ 10	onvoldoende
Minstens een van de drie basisvragen wordt met '1' beoordeeld.		onvoldoende

3 De kwaliteit van de handleiding

Bij dit criterium wordt gevraagd naar de volledigheid van de informatie die de handleiding biedt voor de gebruiker. Hierbij gaat het enerzijds om praktische aanwijzingen voor de afname, scoring en interpretatie (soms gebundeld in een aparte gebruikershandleiding) en anderzijds om informatie over onderzoek dat met de test is verricht (soms gebundeld in een aparte 'technische' handleiding). Beide soorten informatie zijn voor de gebruiker van belang om te kunnen beoordelen welke conclusies er aan een testscore kunnen worden verbonden. Deze informatie moet dan ook voor de gebruiker in overzichtelijke vorm beschikbaar zijn, op papier of digitaal. Voor tests die via de computer worden afgenomen, geldt dat er ook specifieke aanwijzingen moeten worden gegeven voor de installatie en/of het opstarten en het gebruik van de test (soms wordt deze informatie gebundeld in een aparte installatiehandleiding); zie vraag 3.8 tot en met 3.10.

Aanwijzingen bij basisvraag 3.1: "Is er een handleiding beschikbaar?"
Elke test hoort te zijn voorzien van een handleiding. Een proefschrift of een verzameling artikelen wordt niet beschouwd als handleiding.

Aanwijzingen bij vraag 3.2: "Zijn de aanwijzingen voor de testleider volledig en duidelijk?"
De aanwijzingen voor de testleider in de handleiding hebben als belangrijkste doel dat de testafname gestandaardiseerd plaatsvindt. Er moet zo veel mogelijk letterlijk zijn voorgeschreven wat de testleider wel en niet mag zeggen (zo is de aanbeveling "de testleider legt het doel van de test uit" onvoldoende) en welke handelingen de testleider moet verrichten (bijvoorbeeld het op een bepaalde manier rangschikken van het testmateriaal bij een vaardigheidsproef). Ook moet worden voorgeschreven hoe de testleider op

Vragen voor criterium 3 Kwaliteit van de handleiding				
		onv.	vold.	goed
Basisvraag 3.1	Is er een handleiding beschikbaar? Bij negatieve beoordeling (1) van deze vraag kan men de rest van de vragen van dit criterium overslaan en doorgaan met criterium 4.	1		3
3.2	Zijn de aanwijzingen voor de testleider volledig en duidelijk?	1	2	3
3.3	Wordt er informatie gegeven over de gebruiksmogelijkheden en beperkingen van de test?	1	2	3
3.4	Wordt er in de handleiding een samenvatting van de onderzoeksresultaten gegeven?	1	2	3
3.5	Wordt er met behulp van voorbeelden aangegeven hoe testscores kunnen worden geïnterpreteerd?	1	2	3
3.6	Wordt er gewezen op soorten informatie die bij de interpretatie van belang kunnen zijn	1	2	3
3.7	Wordt de mate van deskundigheid vermeld die vereist is voor afname en interpretatie van de test?	1	2	3
Extra vragen voor afname via computer				
3.8	Wordt er informatie gegeven over de installatie van de computersoftware?	1	2	3
3.9	Wordt er informatie gegeven over de bediening en mogelijkheden van de software?	1	2	3
3.10	Zijn er voldoende mogelijkheden voor technische ondersteuning?	1	2	3

vragen moet ingaan (er kunnen bijvoorbeeld standaardteksten worden gegeven voor de antwoorden op veelvoorkomende vragen), welke mate van ondersteuning mag worden geboden en welke hulpmiddelen de geteste mag gebruiken Als de afname via de computer plaatsvindt, moet zijn voorgeschreven over welke computervaardigheden de geteste moet beschikken, en onder welke omstandigheden de test moet worden afgenomen (bijvoorbeeld comfort, werkruimte, licht).

Aanwijzingen bij vraag 3.3: "Wordt er informatie gegeven over de gebruiksmogelijkheden en beperkingen van de test?"
Een handleiding moet volledig, nauwkeurig en duidelijk zijn over de gebruiksmogelijkheden en beperkingen van de test. Daarom moet het voor de (toekomstige) testgebruiker duidelijk zijn welke constructen met behulp van de test worden gemeten, voor welke doelgroep de test is bedoeld en wat de functie is (bijvoorbeeld classificeren, selecteren). Verder moeten de beperkingen van de test zijn beschreven. Afhankelijk van de situatie waarvoor de test is bedoeld kan dit tot verschillende suggesties leiden. Is er bijvoorbeeld bij classificatiebeslissingen in het onderwijs aangegeven dat de beslissing niet op één toetsafname mag worden gebaseerd? Wijst men in het geval van voortgangscontrole op de relatie tussen toetsscore en het verdere onderwijs/leerproces? Leidt interpretatie van testgegevens in de klinische situatie tot empirisch gefundeerde uitspraken of slechts tot werkhypothesen? Wordt er bij tests voor beroepskeuzebegeleiding op gewezen niet alleen op de testscores af te gaan bij het nemen van beslissingen? Wordt er bij tests voor selectie aangegeven voor welk type functies de test is bedoeld en wat de kritieke functie-inhoud is van deze functies?

Aanwijzingen bij vraag 3.4: "Wordt er in de handleiding een samenvatting van de onderzoeksresultaten gegeven?"
Zowel voor (toekomstige) gebruikers van een test als voor beoordelaars van de COTAN is de handleiding de belangrijkste informatiebron. Dissertaties, artikelen in buitenlandse tijdschriften, onderzoeksrapporten en andere publicaties zijn voor gebruikers vaak moeilijk te verkrijgen en door het technisch taalgebruik bovendien niet altijd even toegankelijk. Een samenvatting van de opzet en de resultaten van normerings-, betrouwbaarheids- en validiteitsonderzoek hoort daarom in de handleiding te zijn opgenomen, op een zodanig informatieve en gedegen wijze dat een (potentiële) gebruiker van de test zich een oordeel kan vormen of de test voor zijn doeleinden geschikt is en de vereiste kwaliteit heeft. Een beoordelaar van de COTAN zal incidenteel oorspronkelijke literatuur willen raadplegen en daarom moet daarnaar in de handleiding worden verwezen. Indien van toepassing hoort er ook een samenvatting van de opzet en de resultaten van de kalibratie- en simulatiestudie te zijn opgenomen. Als nieuw onderzoek belangrijke informatie heeft opgeleverd, moeten de gebruikers worden geïnformeerd via supplementen op de handleiding of een herziene versie van de handleiding. Internet biedt uitstekende mogelijkheden om gebruikers een handzaam addendum te verstrekken.

Het gaat er bij deze vraag alleen om of deze informatie is opgenomen in de handleiding. Er wordt hier niet om een waardering van de onderzoeksopzet en de resultaten gevraagd, omdat dit bij de criteria 4, 5, 6 en 7 aan de orde komt.

Bij zogenoemde researchinstrumenten bestaat er veelal geen handleiding. In dergelijke gevallen wordt deze vraag negatief beoordeeld, maar zullen wel oorspronkelijke artikelen, dissertaties, rapporten etc. bij de beoordeling van de andere criteria worden betrokken.

Aanwijzingen bij vraag 3.5: "Wordt er met behulp van voorbeelden aangegeven hoe testscores kunnen worden geïnterpreteerd?"
In een handleiding moeten enkele gevalsbeschrijvingen (cases) en rapportagevoorbeelden worden opgenomen.

Aanwijzingen bij vraag 3.6: "Wordt gewezen op soorten informatie die bij de interpretatie van belang kunnen zijn?"
Wordt er bijvoorbeeld aangegeven welke andere variabelen aan de voorspelling bijdragen? Wordt er vermeld wat de mogelijke invloed is van achtergrondvariabelen en (test)ervaring op de scores?

Aanwijzingen bij vraag 3.7: "Wordt de mate van deskundigheid vermeld die vereist is voor de afname en interpretatie van de test?"
In de handleiding dient aandacht te worden besteed aan de deskundigheid van de beoogde gebruikers. Er kan bijvoorbeeld worden omschreven welke soorten professionals vanuit hun opleiding of werkervaring geschikt worden geacht. Daarnaast kan er een adequate omschrijving worden gegeven van de kennis en vaardigheden die noodzakelijk worden geacht voor de afname en de interpretatie van de test.

Extra vragen voor afname via computer

Aanwijzingen bij vraag 3.8: "Wordt er informatie gegeven over de installatie van de computersoftware?"
Informatie over de benodigde hard- en software en over de manier waarop de CBT-software geïnstalleerd kan worden, is vereist. Wat betreft de hardware is het van belang dat de vereiste CPU, het minimaal vereiste geheugen, de benodigde schijfruimte, de vereiste monitor en videokaart, de benodigde input devices en de benodigde exchange devices (bijvoorbeeld cd-romspeler) worden vermeld. Daarnaast kan informatie over bijvoorbeeld de vereiste netwerkkaart of geluidskaart nodig zijn. Wat betreft de software is het van belang dat wordt vermeld onder welke besturingssystemen de test functioneert en welke andere software vereist is (bijvoorbeeld een browser of bepaalde plugins). De manier waarop de CBT-software geïnstalleerd kan worden, moet stapsgewijs en zo mogelijk met ondersteuning van screendumps zijn beschreven.

Als er een beschrijving van de benodigde hardware, óf de benodigde software, óf van de installatie van de CBT-software ontbreekt, dan is het oordeel 'onvoldoende'. De beschrijving van de installatie van de CBT-software moet hierbij als aanwezig worden beschouwd wanneer de CBT-software zichzelf automatisch installeert. Alleen een uitgebreide beschrijving van de benodigde hard- en software én een duidelijke beschrijving van de installatie van de CBT-software (automatische installatie uitgezonderd) kunnen leiden tot het oordeel 'goed' op deze vraag.

Aanwijzingen bij vraag 3.9: "Wordt er informatie gegeven over de bediening en mogelijkheden van de software?"

Bij elke CBT moet er informatie worden gegeven over de bediening van de software en de mogelijkheden die de software kent, bijvoorbeeld te kiezen instellingen, de mogelijkheid van groepsoverzichten, en analyse- en rapportageopties. Als er geen of onvoldoende informatie wordt gegeven over een van beide aspecten, dan is het oordeel 'onvoldoende'. In andere gevallen is de duidelijkheid en volledigheid van de verstrekte informatie doorslaggevend voor het oordeel 'voldoende' of 'goed'.

Aanwijzingen bij vraag 3.10: "Zijn er voldoende mogelijkheden voor technische ondersteuning?"

Wanneer de testgebruiker bepaalde vragen heeft over de CBT-software, of wanneer er storingen in de CBT-software optreden, dan moet er ondersteuning beschikbaar zijn. Dit kan zijn in de vorm van documentatie over veelvoorkomende problemen, in de vorm van een paragraaf die is gewijd aan 'veelgestelde vragen' of in de vorm van een helpdesk (waarvan de beschikbaarheid en de bereikbaarheid in de handleiding moet zijn aangegeven).
Deze vraag mag alleen met het oordeel 'goed' worden beoordeeld, als er naast schriftelijke of elektronische documentatie over het oplossen van problemen, ook de mogelijkheid is om terug te vallen op een helpdesk. Alleen wanneer er geen documentatie over het oplossen van problemen beschikbaar is én de testgebruiker niet terug kan vallen op een helpdesk, dan is het oordeel 'onvoldoende'. In alle andere gevallen is het oordeel 'voldoende'.

Vaststelling eindoordeel voor criterium 3 Kwaliteit van de handleiding Papier-en-potloodversie		
De basisvraag wordt met '3' beoordeeld.	Somscore 3.2 t/m 3.7 ≥ 13	goed
	Somscore 3.2 t/m 3.7 = 11 of 12	voldoende
	Somscore 3.2 t/m 3.7 ≤ 10	onvoldoende
De basisvraag wordt met '1' beoordeeld.		onvoldoende

Vaststelling eindoordeel voor criterium 3 Kwaliteit van de handleiding Afname via computer		
De basisvraag wordt met '3' beoordeeld.	Somscore 3.2 t/m 3.10 ≥ 19	goed
	Somscore 3.2 t/m 3.10 = 17 of 18	voldoende
	Somscore 3.2 t/m 3.10 ≤ 16	onvoldoende
De basisvraag wordt met '1' beoordeeld.		onvoldoende

4 Normen

Het scoren van een test levert een zogenoemde ruwe score op. Ruwe scores worden gedeeltelijk bepaald door de kenmerken van de test, zoals het aantal items, de keuze van de tijdslimiet, de moeilijkheidsgraad of de populariteit van de items en de omstandigheden waaronder de test is afgenomen. Dit zorgt ervoor dat ruwe scores moeilijk te interpreteren zijn. In het algemeen krijgt de ruwe score pas betekenis door deze te vergelijken met een norm.

Men kan twee typen normscores onderscheiden (APA, 1999). Bij het eerste type wordt de behaalde ruwe score vergeleken met die van anderen. Deze wijze van interpreteren wordt *normgerichte interpretatie* genoemd. Men vergelijkt de score met de scoreverdeling van een of meer referentiegroepen. Het doel is vast te stellen hoe de onderzochte scoort ten opzichte van andere personen waarmee een zinvolle (op basis van overeenkomsten in bijvoorbeeld leerjaar, leeftijd, functie) vergelijking kan worden gemaakt. Dit type normen wordt ook wel relatieve normen genoemd.

Bij het tweede type vergelijkt men het resultaat niet met dat van anderen, maar worden de testresultaten absoluut geïnterpreteerd, dat wil zeggen dat zij met een absolute norm worden vergeleken. Deze wijze van interpreteren wordt *domeingerichte* of *criterium-gerichte interpretatie* genoemd. Bij deze wijze van normeren worden bepaalde standaarden of grensscores vastgesteld. In het geval van domeingerichte interpretatie worden deze standaarden op de een of andere wijze door experts of beoordelaars bepaald. De norm kan direct zijn afgeleid van een omschrijving van het domein van vaardigheden of leerstof dat men moet beheersen. Dit type normen wordt ook wel absolute normen genoemd. Bij criteriumgerichte interpretatie worden de grensscores aan onderzoeksgegevens ontleend. Bij deze wijze van normeren moeten er naast testgegevens dan ook gegevens over het criterium zijn verzameld. Bij gebruik van een enkele grensscore kan deze het verschil tussen zakken en slagen of tussen afwijzen en toelaten aangeven. In het geval van meer dan één grensscore, kunnen bijvoorbeeld verschillende vaardigheidsniveaus worden onderscheiden.

Als er geen normen worden verstrekt, dan is het eindoordeel op dit criterium in principe 'onvoldoende'. Er kunnen echter beargumenteerde uitzonderingen voorkomen, bijvoorbeeld bij tests waar een louter intra-individuele vergelijking wordt aanbevolen en gerechtvaardigd is, zoals bij ipsatieve tests of tests die vorderingen in de tijd meten. In dat geval kan bij dit criterium 'n.v.t.' worden ingevuld.

Vragen voor criterium 4
Normen
Algemene basisvragen

		onv.	vold.	goed
Basisvraag 4.1	Worden er normen verstrekt? Bij negatieve beoordeling (1) van deze vraag kan men de rest van de vragen van dit criterium overslaan en doorgaan met criterium 5	1	n.v.t.	3
Basisvraag 4.2	Zijn de normen actueel? Bij negatieve beoordeling (1) van deze vraag kan men de rest van de vragen van dit criterium overslaan en doorgaan met criterium 5.	1	2	3

Wanneer er van normgerichte interpretatie sprake is, doorgaan met vraag 4.3.
Wanneer er van domeingerichte interpretatie sprake is, doorgaan met vraag 4.8.
Wanneer er van criteriumgerichte interpretatie sprake is, doorgaan met vraag 4.11.

Vragen voor criterium 4 Normen Normgerichte interpretatie					
		onv.	vold.	goed	
Basisvraag 4.3	Wat is de kwaliteit van de verstrekte normgroepen? a. Zijn de normgroepen groot genoeg? b. Zijn de normgroepen representatief? Bij negatieve beoordeling (1) van vraag 4.3.a of 4.3.b kan men de rest van de vragen van dit criterium overslaan en doorgaan met criterium 5.	 1 1	 2 2	 3 3	
4.4	Worden de betekenis en de beperkingen van de normschaal duidelijk gemaakt voor de gebruiker en is het type normschaal in overeenstemming met het doel van de test?	1	2	3	
4.5	Worden er gemiddelden, standaardafwijkingen en gegevens over de scoreverdeling vermeld?	1	2	3	
4.6	Worden er gegevens verstrekt over mogelijke verschillen tussen subgroepen (bijvoorbeeld allochtonen-autochtonen, vrouwen-mannen)?	1	2	3	
4.7	Worden er gegevens verstrekt over de nauwkeurigheid van de meting en de daarbij behorende intervallen? a. standaardmeetfout b. standaardschattingsfout c. testinformatiefunctie / standaardfout	 1 1 1	 2 2 2	 3 3 3	 n.v.t. n.v.t. n.v.t.
Domeingerichte interpretatie					
Als de grensscores met behulp van beoordelaars worden bepaald: wat is de kwaliteit van de standaardbepalingsprocedure?		onv.	vold.	goed	
Basisvraag 4.8	Is er voldoende overeenstemming tussen de beoordelaars?	1	2	3	n.v.t.
4.9	Zijn de procedures op grond waarvan de grensscores zijn bepaald correct?	1	2	3	n.v.t.
4.10	Zijn de beoordelaars naar behoren geselecteerd en getraind?	1	2	3	n.v.t.
Criteriumgerichte interpretatie					
Als de grensscores empirisch worden onderbouwd: wat is de uitkomst en de kwaliteit van dit onderzoek?		onv.	vold.	goed	
Basisvraag 4.11	Rechtvaardigen de onderzoeksresultaten het gebruik van grensscores?	1	2	3	n.v.t.
4.12	Is de onderzoeksgroep in overeenstemming met het bedoelde gebruik?	1	2	3	n.v.t.
4.13	Is de onderzoeksgroep groot genoeg?	1	2	3	n.v.t.

Aanwijzingen bij basisvraag 4.1: "Worden normen verstrekt?"
Normgegevens, hetzij bedoeld voor normgerichte interpretatie (zoals normtabellen), hetzij bedoeld voor domeingerichte of criteriumgerichte interpretatie (zoals grensscores of verwachtingstabellen), moeten beschikbaar zijn op het moment dat de test voor daadwerkelijk gebruik verkrijgbaar is. De volgende situaties kunnen leiden tot een negatief antwoord op deze vraag:

- De hier bedoelde gegevens worden niet verstrekt (er worden bijvoorbeeld alleen gemiddelden en standaarddeviaties van de onderzochte groepen vermeld).
- Bij tests die zijn bedoeld voor interpretatie op groepsniveau worden normtabellen verstrekt die zijn gebaseerd op individuele scores, en omgekeerd (zie ook ad. 4.3.a).
- Nadat de normen zijn verzameld, hebben er nog wijzigingen in de test zelf plaatsgevonden, bijvoorbeeld wijzigingen in de items of de instructie.
- De normen zijn verzameld met behulp van een papier-en-potloodversie, terwijl de te beoordelen versie een computerversie betreft (of vice versa). Voor vragenlijsten heeft dit in het algemeen weinig invloed op de waarde van de normen (Bartram, 2005; King & Miles, 1995; Mead & Drasgow, 1993). Voor capaciteiten- en vaardigheidstests en/of tests die gebonden zijn aan een tijdslimiet zullen echter nieuwe normen moeten worden verzameld.

Aanwijzingen bij basisvraag 4.2: "Zijn de normen actueel?"
Normen zijn aan slijtage onderhevig. Van de psychometrische kenmerken van een test zijn normen het meest gevoelig voor maatschappelijke veranderingen, veranderingen in het onderwijs en in de inhoud van functies. Daarom moet er van tijd tot tijd hernormering van de test plaatsvinden, of moet de auteur door onderzoek aantonen dat hernormering niet nodig is. Voor intelligentietests moet bijvoorbeeld rekening worden gehouden met het Flynn-effect, waardoor normen verouderen (zie bijvoorbeeld Resing & Drenth, 2007, p. 142-145). Dit effect wordt geschat op 3 IQ-punten per tien jaar, of 4.5 IQ-punten per vijftien jaar. Dit is gelijk aan circa één standaardmeetfout (bij een betrouwbaarheid van .91). Waarschijnlijk geldt een dergelijk effect ook voor verwante tests, zoals testbatterijen voor algemene of specifieke cognitieve capaciteiten. Voor persoonlijkheidstests is er niets bekend over dergelijke algemene effecten. Vergelijking tussen de gegevens in de handleidingen van enkele Nederlandse tests heeft het volgende opgeleverd. Voor de *Amsterdamse Beroepen Interessen Vragenlijst* (Evers, 1979, 1992) worden over een periode van zestien jaar voor sommige schalen verschillen gevonden die oplopen tot twee standaarddeviaties. Voor de *NPV* worden over een periode van ruim twintig jaar verschillen gevonden van maximaal 1.4 standaarddeviatie bij de normgroep selectie, van maximaal 1.2 standaarddeviatie bij de normgroep algemeen en van maximaal 0.5 standaarddeviatie bij de normgroep psychiatrische patiënten (Luteijn, Starren & van Dijk, 1985; Barelds, Luteijn, van Dijk & Starren, 2007). Hierbij moet overigens worden aangetekend dat in beide vragenlijsten ook items zijn gewijzigd. Amerikaans onderzoek (Twenge, 2000) laat zien dat over een periode van veertig jaar angst- en neuroticismescores in de VS met circa een hele standaardafwijking zijn toegenomen.

In het Duitse beoordelingssysteem voor de kwaliteit van tests (Kersting, 2006) wordt een periode van acht jaar voor hernormering aanbevolen, overigens zonder hier consequenties aan te verbinden. In de APA-Standards (APA, 1999, p. 59, Standard 4.18) wordt gesteld dat: "... so long as the test remains in print, it is the publisher's responsibility to assure that the test is renormed with sufficient frequency to permit continued accurate and appropriate score interpretations". De APA noemt hierbij geen termijn. Op basis van bovenstaande bevindingen en daarbij een afweging makend tussen wat praktisch haalbaar en wenselijk is, komt de COTAN tot de volgende regel. Om de gebruiker te attenderen op mogelijk versleten normen, wordt aan de beoordeling van tests waarvan hernormerings- of ijkingsonderzoek sinds vijftien jaar na het afsluiten van het normeringsonderzoek niet heeft plaatsgevonden, de kwalificatie "De normen zijn verouderd" toegevoegd. Na nog eens vijf jaar zonder dergelijk onderzoek wordt deze kwalificatie gewijzigd in: "Wegens veroudering zijn de normen niet meer bruikbaar" en wordt de beoordeling 'onvoldoende'. Eenmaal per jaar zullen alle testbeschrijvingen in de via internet te raadplegen *Documentatie van Tests en Testresearch* op dit punt worden aangepast. Om te kunnen beoordelen in hoeverre normen mogelijk zijn verouderd, is het vermelden van het jaar (of de periode) van gegevensverzameling van belang. Als dit niet wordt vermeld, wordt de beoordeling voor 'Normen' daarom 'onvoldoende'.

Normgerichte interpretatie

Aanwijzingen bij basisvraag 4.3: "Wat is de kwaliteit van de verstrekte normgroepen?"
In principe moet de testauteur normen verschaffen voor elk door hem genoemd gebruiksdoel (zie vraag 1.1). Het kan blijken dat de groepen waarvoor normen worden verschaft slechts gedeeltelijk de meetpretentie dekken. Wanneer een auteur bijvoorbeeld aangeeft dat een test is bedoeld voor keuzebegeleiding binnen het voorbereidend beroepsonderwijs én voor selectie voor functies op dit niveau, dan moeten voor beide situaties normen worden verstrekt. Het is echter irreëel te verwachten dat voor elke functie op dit niveau normen worden verschaft.

Wil een normgroep goed aan zijn doel kunnen beantwoorden (namelijk het vormen van een betrouwbare reeks van referentiepunten), dan moet de normgroep én van voldoende omvang te zijn én representatief zijn voor de bedoelde groep. Voor de beoordeling van beide aspecten worden hieronder aanwijzingen gegeven. De beoordeling voor vraag 4.3 kan alleen 'goed' worden wanneer beide aspecten (vragen 4.3.a en 4.3.b) als 'goed' worden beoordeeld. De beoordeling wordt 'onvoldoende' wanneer minstens een van beide aspecten 'onvoldoende' wordt beoordeeld. In alle andere gevallen wordt de beoordeling 'voldoende'.

Aanwijzingen bij vraag 4.3.a: "Zijn de normgroepen groot genoeg?"

In de literatuur komt men slechts spaarzaam aanbevelingen tegen over de gewenste grootte van normgroepen (Angoff, 1971; Campbell, 1971). Deze aanbevelingen zijn óf gebaseerd op de berekening van standaardfouten in parameters zoals gemiddelde en mediaan, óf op ervaringsgegevens met betrekking tot de stabiliteit van schaalwaarden. Een synthese van deze twee, gekoppeld aan het belang van de met de test te nemen beslissingen, heeft in de volgende beoordelingsregels geresulteerd:

Tests voor belangrijke* beslissingen op individueel niveau (bijvoorbeeld personeelsselectie, verwijzing naar speciaal onderwijs, opname/ontslag kliniek, certificering).	N ≥ 400 300 ≤ N < 400 N < 300	goed voldoende onvoldoende
Tests voor relatief minder belangrijke beslissingen op individueel niveau (bijvoorbeeld voortgangscontrole, in het algemeen beschrijvend gebruik, zoals bij beroepskeuzebegeleiding, therapie-indicatie).	N ≥ 300 200 ≤ N < 300 N < 200	goed voldoende onvoldoende
* Met belangrijke beslissingen wordt bedoeld: beslissingen die op basis van de testscores worden genomen, die in principe, of op korte termijn, onomkeerbaar zijn, en die voor een belangrijk deel buiten de geteste persoon om worden genomen.		

De eis van de steekproefgrootte geldt uiteraard *per normgroep* waarvoor wordt genormeerd. Bij bijvoorbeeld ontwikkelingstests die voor verschillende leeftijdsgroepen worden genormeerd, kan dit verwarring geven. Als de normering afzonderlijk per leeftijdsgroep (of leerjaar, of schooltype) wordt uitgevoerd, dan is de steekproefgrootte van elke subgroep afzonderlijk van belang. Als echter continue normering of fit-procedures worden toegepast, waarbij gelijktijdig van de informatie van alle leeftijdsgroepen gebruik wordt gemaakt, kan de grootte per leeftijdsgroep kleiner zijn, omdat deze procedure efficiëntere schatters oplevert dan klassieke normering. Een onderzoek naar equivalentie van de aantallen tussen klassieke en continue normen is gedaan door Bechger, Hemker en Maris (2009). Dat onderzoek laat zien dat in een speciaal geval de benodigde aantallen lager kunnen liggen. Zij nemen hiertoe de standaardfout van het gemiddelde als parameter die als indicator voor de nauwkeurigheid van normen kan worden gebruikt. Het uitgangspunt voor het vaststellen van de groepsgrootte bij continue normering is dat de nauwkeurigheid van de normen minstens hetzelfde niveau moet halen als bij klassieke normering (dat wil zeggen normering waarbij de normgegevens op elke groep afzonderlijk worden berekend).

Bij dit onderzoek en bij de door COTAN hieruit afgeleide richtlijnen moeten twee kanttekeningen worden gemaakt. Ten eerste wordt bij de berekeningen een aantal statistische vooronderstellingen gedaan, zoals de veronderstelling dat de varianties in de subgroepen gelijk zijn, dat de scores binnen elke subgroep normaal zijn verdeeld en dat de regressie van de testscore op leeftijd lineair is. Het niet voldoen aan deze vooronderstellingen kan tot grotere standaardfouten leiden en dus tot een groter aantal benodigde waarnemingen dan Bechger e.a. (2009) noemen. Ook bestaan er verschillende varianten van continue normering, waarvan in het genoemde onderzoek slechts één variant met acht groepen is doorgerekend. Ten tweede is in het geval van continue normering niet in iedere subgroep een gelijk aantal personen noodzakelijk om voor iedere subgroep tot een equivalente precisie te komen. In de middelste groepen zijn bij continue normering minder observaties nodig dan in de extreme groepen. In onderstaande richtlijnen heeft de COTAN er vanwege transparantie voor gekozen een gelijk aantal per subgroep te eisen. De consequentie is dat bij de gestelde aantallen in de extreme groepen enige meetnauwkeurigheid wordt ingeleverd, maar dat deze in de middelste groepen veel groter is, vergeleken met de aantallen bij klassieke normering. De onderstaande richtlijnen moeten dan ook worden gezien als een ondergrens van wat theoretisch wenselijk is.

Richtlijnen voor subgroepsgrootte bij continue normering met acht subgroepen		
Tests voor belangrijke* beslissingen op individueel niveau (bijvoorbeeld personeelsselectie, verwijzing naar speciaal onderwijs, opname/ontslag kliniek, certificering).	N ≥ 150 100 ≤ N < 150 N < 100	goed voldoende onvoldoende
Tests voor relatief minder belangrijke beslissingen op individueel niveau (bijvoorbeeld voortgangscontrole, in het algemeen beschrijvend gebruik, zoals bij beroepskeuzebegeleiding, therapie-indicatie).	N ≥ 100 70 ≤ N < 100 N < 70	goed voldoende onvoldoende
* Met belangrijke beslissingen wordt bedoeld: beslissingen die op basis van de testscores worden genomen, die in principe, of op korte termijn, onomkeerbaar zijn, en die voor een belangrijk deel buiten de geteste persoon om worden genomen.		

Voor een uitgebreide toelichting op de wijze waarop deze richtlijnen tot stand zijn gekomen, wordt verwezen naar het onderzoek van Bechger e.a. (2009). Hier wordt slechts een korte toelichting gegeven. Voor de standaardfout van het gemiddelde zijn de waarden berekend die bij de aantallen horen die als grenzen voor de kwalificaties 'onvoldoende', 'voldoende' en 'goed' worden gebruikt bij normering op afzonderlijke groepen, namelijk 400, 300 en 200 personen. Bij een standaarddeviatie van 15 (zoals bij intelligentietests gebruikelijk is) zijn de standaardfouten van het gemiddelde respectievelijk 0.75, 0.87 en 1.06. Uitgaande van de regressieaanpak bij continue normering kunnen vervolgens voor elk aantal subgroepen en voor elke steekproefomvang de standaardfouten van het gemiddelde voor elke subgroep worden berekend. Bij acht subgroepen (bijvoorbeeld de acht groepen in het basisonderwijs) en bij een grootte van 100 personen per groep, zijn deze in de groepen 4 en 5 gelijk aan circa .54, bij de groepen 3 en 6 aan .63, bij de groepen 2 en 7 aan .77 en bij de groepen 1 en 8 aan .96. Bij klassieke normering en bij een groepsgrootte van 300 personen is de standaardfout gelijk aan .87 in alle groepen. Wanneer bij continue normering over acht groepen elke groep uit 100 personen zou bestaan, zou de nauwkeurigheid in zes van de acht groepen dus verbeteren, en bij twee groepen verslechteren ten opzichte van klassieke normering. Hoewel dit laatste gegeven uiteraard onwenselijk is, is de mate van verslechtering beperkt en de 'winst' bij de middelste groepen groot. Daarom heeft de COTAN de richtlijn van 300 personen bij klassieke normering gelijk gesteld aan die van 100 personen bij continue normering (bij minimaal acht groepen). Op vergelijkbare wijze zijn de aantallen van 150 en 70 personen tot stand gekomen. Alleen bij de twee extreme groepen wordt de nauwkeurigheid dan iets slechter, maar bij de andere zes groepen beter, in vergelijking met respectievelijk 400 en 200 personen bij klassieke normering.

Bovenstaande richtlijnen gelden uitdrukkelijk als voorbeeld, want een eerste beperking is dat de richtlijnen uitsluitend betrekking hebben op de situatie waarin acht subgroepen worden onderscheiden. De tweede beperking betreft het feit dat in het voorbeeld uitsluitend aandacht wordt geschonken aan de standaardfout van het gemiddelde en dat andere momenten van de verdeling niet in het voorbeeld zijn opgenomen. De derde beperking is dat er in het voorbeeld van wordt uitgegaan dat aan de statistische vooronderstellingen wordt voldaan. Wanneer er gebruik wordt gemaakt van continue normering, is het daarom aan de testauteur om in zijn specifieke situatie aan te tonen aan welk aantal in de klassieke normering de gebruikte steekproefomvang equivalent is. Merk op dat niet alleen moet worden aangetoond dat de standaardfout van het gemiddelde equivalent is, maar dat dit ook aangetoond moet worden voor andere momenten van de verdeling, zoals de standaarddeviatie. Als bij het bepalen van die equivalentie gebruik is gemaakt van vooronderstellingen, dan moet worden aangegeven of aan die vooronderstellingen is voldaan.

Wanneer een test louter is bedoeld om uitspraken op groepsniveau mogelijk te maken, gelden er andere eisen voor van de steekproefgrootte, omdat de standaardfout van groepsgemiddelden in het algemeen veel kleiner is dan die van individuele scores. Zo zou volgens Angoff (1971) de spreiding van individuele scores bij toetsen voor schoolprestaties 2 tot 2.5 maal groter zijn dan de spreiding van groepsgemiddelden. Een combinatie van de regel voor individuele normen en dit gegeven leidt tot de beoordelingsregel die in de tabel hieronder staat.

Tests voor onderzoek op groepsniveau.	K ≥ 40 30 ≤ K < 40 K < 30	goed voldoende onvoldoende

Hierin is *K* het aantal groepen en moet elke groep uit minstens 25 geteste personen bestaan. Voorbeelden van dit type tests zijn: vragenlijsten voor schoolklimaat; arbeidstevredenheid; arbeidsomstandigheden en organisatiecultuur. Voor representativiteit en aanvullende statistische informatie gelden, mutatis mutandis, dezelfde eisen zoals geformuleerd in de vragen 4.3.b tot en met 4.7, tenzij anders is aangegeven.

Aanwijzingen bij vraag 4.3.b: "Zijn de normgroepen representatief?"

Een steekproef is representatief als de samenstelling ervan voor een aantal variabelen overeenkomt met die van de betreffende populatie, waarbij de steekproef wordt verkregen met behulp van een aselect steekproefmodel. In een aselecte steekproef heeft elk element in de populatie een even grote kans in de steekproef te worden opgenomen. Om te kunnen beoordelen of de normgroepen representatief zijn, moet er zowel een adequate omschrijving van de populatie als van de wijze van steekproeftrekking of dataverzameling worden gegeven. Deze eis geldt per groep waarvoor normen worden verstrekt, zowel bij klassieke als bij continue normering. Het komt regelmatig voor dat de geboden informatie zo beperkt is, dat het zelfs niet duidelijk is om welke populatie het gaat. Is de test bijvoorbeeld bedoeld voor landelijk, regionaal of lokaal gebruik? Gaat het om een doorsnee van de bevolking of om mensen die een bepaald kenmerk bezitten (bijvoorbeeld uitsluitend mensen die zich hebben aangemeld voor psychologische hulpverlening, mensen met een bepaalde opleiding)?

De samenstelling van een steekproef moet in ieder geval beschreven zijn met betrekking tot de variabelen leeftijd, sekse, etniciteit en regio, omdat de ervaring leert dat deze variabelen bij de meest uiteenlopende tests en vragenlijsten tot scoreverschillen tussen subgroepen leiden. Het kan van belang zijn, onder andere afhankelijk van de meetpretentie van de test, dat ook de samenstelling met betrekking tot variabelen zoals urbanisatiegraad, sociaaleconomische status, opleidingsniveau, doel van het testgebruik, functieniveau, bedrijfstak, wel/geen verwijzing of klinische diagnose, wordt beschreven. Om te kunnen vaststellen of de verdelingen in de gegeven steekproeven overeenkomen met die in de bijbehorende populaties moet er ook altijd een beschrijving van de betreffende populaties worden gegeven. Voor de meer algemene achtergrondvariabelen kan men hiervoor in het algemeen gebruikmaken van gegevens van het *Centraal Bureau voor de Statistiek*. Voor tekorten op bepaalde niveaus van de gebruikte variabelen in het steekproefmodel mag in beperkte mate door weging worden gecorrigeerd. Oververtegenwoordiging is geen probleem. Bij ondervertegenwoordiging is maximaal een factor 2 acceptabel.

Bij de registratie van etniciteit kan men als probleem tegenkomen dat de gegevens niet beschikbaar zijn of niet geregistreerd mogen worden. Ook bestaan er van etniciteit verschillende definities, die voor het verrichten van onderzoek niet altijd sluitend of inhoudelijk bevredigend zijn. Ten slotte is de samenstelling van de doelpopulatie in dit opzicht niet altijd bekend. Dit ontslaat de testauteur echter niet van de inspanningsverplichting om zo volledig mogelijke informatie aan te leveren. Normgroepen komen op basis van het bovenstaande alleen in aanmerking voor de beoordeling 'goed' wanneer er is uitgegaan van een aselect steekproefmodel waarmee wordt gestreefd naar landelijke representativiteit. Twee veelvoorkomende manieren van gegevensverzameling voldoen niet aan deze eis: 'regionale normen' en *samples of convenience*' (gelegenheidssteekproeven). Voor de beoordeling van deze twee types steekproeven worden hieronder aanwijzingen gegeven. Normen die gebaseerd zijn op dergelijke steekproeven kunnen maximaal de beoordeling 'voldoende' krijgen.

Regionale normen

Als de test is bedoeld voor landelijk gebruik, zullen ook landelijk goed gespreide normen moeten worden verzameld, omdat de scores op vele typen tests regionale verschillen vertonen. Regio ontleent zijn invloed voornamelijk aan het feit dat regio samenhangt met variabelen die deze scoreverschillen veroorzaken, zoals sociaal-economische status, opleidingsniveau en etniciteit. Wanneer de auteur kan aantonen dat de samenstelling van een regionale steekproef op de belangrijkste achtergrondvariabelen overeenkomt met de landelijke populatie, kan vraag 4.3.b voor een *regionale* steekproef – ten hoogste – tot een score 2 leiden. Welke achtergrondvariabelen belangrijk zijn, hangt af van de te verwachten correlatie tussen achtergrondvariabele en testscore. Van een test voor taalvaardigheid valt bijvoorbeeld te verwachten dat de scores zullen samenhangen met etniciteit. Wanneer een regionale steekproef is gebaseerd op een steekproef uit een grote stad in het westen van het land, zullen allochtone respondenten ten opzichte van de rest van Nederland zeer waarschijnlijk zijn oververtegenwoordigd, waardoor de gemiddelde testscore en de normen geen representatief beeld voor heel Nederland geven. In dit geval moeten er over de samenstelling, wat betreft etniciteit van steekproef en populatie, gegevens worden verstrekt, en eventueel moet er voor een afwijkende samenstelling worden gecorrigeerd. De score 2 op vraag 4.3.b heeft als consequentie dat regionale normen maximaal als 'voldoende' kunnen worden beoordeeld. Of de beoordeling 'voldoende' vervolgens daadwerkelijk wordt toegekend, is volgens de gebruikelijke wegingsregels afhankelijk van de beoordeling van de vragen 4.4 tot en met 4.7. Belangrijke variabelen in dit verband zijn sekse, leeftijd, en etniciteit, maar ook het doel van het onderzoek (bijvoorbeeld selectie of loopbaanbegeleiding) kan van belang zijn.

Samples of convenience

Bij de dataverzameling wordt nogal eens gebruikgemaakt van een sample of convenience (gelegenheidssteekproef), bijvoorbeeld leerlingen met keuzeproblemen die zich aanmelden voor hulpverlening, psychologiestudenten omdat die makkelijk beschikbaar zijn, de sollicitanten die door een bureau voor werving en selectie worden getest, enzovoort. In het algemeen zijn dit slechte normgroepen, omdat er hierbij niet wordt gecontroleerd voor variabelen die met de testscore kunnen samenhangen en deze groepen niet kunnen worden beschouwd als representatief voor de mogelijk bedoelde populaties (bijvoorbeeld brugklasleerlingen, studenten aan het hoger onderwijs, alle Nederlandse werknemers in een bepaald type functie of van een bepaald niveau).

Gelegenheidssteekproeven zijn eigenlijk geen steekproeven in de strikte zin van het woord: meestal is het de hele cliëntenpopulatie die in een bepaalde periode bij een bepaalde instantie een vragenlijst of een test invult. Er is geen garantie dat elk lid van de doelpopulatie evenveel kans heeft om in de steekproef terecht te

komen. Er ligt geen steekproefmodel aan de verzameling van de gegevens ten grondslag. Het probleem bij gelegenheidssteekproeven is dat men niet goed weet wat men precies verzamelt. Wanneer bijvoorbeeld een dergelijke steekproef is gebaseerd op de cliënten van een aantal beroepskeuzebureaus, kan men een dergelijke groep dan beschouwen als een doorsnede van de Nederlandse bevolking van dezelfde leeftijd en opleiding, of is er reden om aan te nemen dat mensen met een beroepskeuzeprobleem verschillen van de overige Nederlanders? Of: kan men de normgroep die is opgebouwd uit de sollicitanten die zijn getest door wervings- en selectiebureau X met drie vestigingen verspreid over Nederland gebruiken voor de sollicitanten bij bureau Y? In het algemeen zullen dergelijke normgroepen de beoordeling 'onvoldoende' krijgen, omdat hun samenstelling onbekend of oncontroleerbaar is. Door sommige testauteurs wordt de omvang van de normgroep wel als argument voor de representativiteit, geldigheid of toepasbaarheid van de normen beschouwd. De omvang van dergelijke normgroepen is inderdaad meestal geen probleem (aantallen van enkele duizenden zijn geen uitzondering), maar de omvang van de steekproef zegt op zich niets over de representativiteit, noch over de bruikbaarheid. Bijvoorbeeld: de opdrachtgevers van bureau X bestaan voornamelijk uit bedrijven in de sector ICT. De normen die gebaseerd zijn op deze sollicitantengroep zijn mogelijk wel geschikt voor andere wervings- en selectiebureaus die gespecialiseerd zijn in de ICT-branche, maar niet voor bureaus die hun opdrachten voornamelijk uit andere branches ontvangen.

Men kan op gelegenheidssteekproeven gebaseerde normen beschouwen als een bijzondere vorm van regionale normen (omgekeerd zullen regionale normen ook vaak een gelegenheidssteekproef zijn); men zou bijvoorbeeld de sollicitanten bij bureau X kunnen beschouwen als een steekproef van alle sollicitanten in Nederland. Net zoals bij regionale normen, kan de kwaliteit van een gelegenheidssteekproef als 'voldoende' worden beoordeeld wanneer er een uitputtende beschrijving van de normgroep over mogelijk relevante variabelen van de normgroep wordt gegeven. Een beschrijving en het aantonen van representativiteit voor de beoogde doelgroep in termen van sekse, leeftijd en etniciteit alléén is echter niet genoeg; het gaat daarnaast om variabelen die samenhangen met het gebruiksdoel van de test, zoals bedrijfstak en functie in selectiesituaties en type stoornis in klinische situaties. Of de beoordeling 'voldoende' vervolgens daadwerkelijk wordt toegekend, is volgens de gebruikelijke wegingsregels afhankelijk van de beoordeling van de vragen 4.4 tot en met 4.7. De motivering voor het geven van een 'voldoende' in deze gevallen is dat dergelijke normen goed bruikbaar zijn wanneer de gebruiker weet waarmee een cliënt of sollicitant kan worden vergeleken.

Aanwijzingen bij vraag 4.4: "Worden de betekenis en de beperkingen van de normschaal duidelijk gemaakt voor de gebruiker en is het type normschaal in overeenstemming met het doel van de test?"

Bij de omzetting van ruwe scores in afgeleide scores kan er een keuze worden gemaakt uit drie typen normen (Drenth & Sijtsma, 2006): verhoudingsnormen, normen gebaseerd op rangorde en normen gebaseerd op gemiddelde en spreiding.

Een bekend voorbeeld van verhoudingsnormen is het ouderwetse Intelligentie Quotiënt (IQ), waarbij mentale leeftijd wordt gedeeld door chronologische leeftijd. Tegenwoordig wordt deze wijze van IQ-berekening echter niet of nauwelijks meer gebruikt. Een ander voorbeeld is het Didactisch Leeftijds Equivalent (DLE), waarbij de verhouding wordt berekend tussen het niveau van de leerprestatie die een kind heeft geleverd in termen van aantal maanden onderwijs waarna deze prestatie mag worden verwacht, en het werkelijk aantal maanden gevolgd onderwijs. Tegen verhoudingsnormen bestaan vele praktische en theoretische bezwaren. Zie voor een uitgebreide bespreking hiervan Evers en Resing (2007). Het gebruik van verhoudingsnormen in het algemeen en DLE's in het bijzonder wordt door de COTAN afgekeurd. Dit heeft als consequentie dat tests die uitsluitend in termen van DLE's rapporteren een 'onvoldoende' krijgen voor normen. Ook bij tests die naast standaardscores of rangordescores DLE's rapporteren – maar de COTAN ziet rapportage in termen van DLE's het liefst geheel verdwijnen – moet de rapportage in standaardscores of rangordescores vooropstaan en nadrukkelijk worden aanbevolen, vergezeld van een gedegen uitleg van deze systemen voor de gebruikers. Daarnaast moet de gebruiker in elk geval in de handleiding uitdrukkelijk voor de beperkingen van DLE's worden gewaarschuwd.

Normen die gebaseerd zijn op rangorde zijn percentielen en daarvan afgeleide schaaltypen, zoals vigintielen, decielen en het A t/m E-systeem dat door het Cito wordt gebruikt. Voorbeelden van normtypen die gebaseerd zijn op gemiddelde en spreiding, hier deviatienormen genoemd, zijn stanines, C-scores, T-scores en deviatie-IQ's. Bij deviatienormen kan er verder nog onderscheid worden gemaakt tussen lineaire transformaties en genormaliseerde transformaties. In principe moet er van normaliserende transformaties gebruik zijn gemaakt, tenzij de ruwe scores al bij benadering normaal zijn verdeeld. Als er meer dan één normgroep wordt onderscheiden, bijvoorbeeld voor opeenvolgende leeftijdsgroepen of leerjaren, is het beter als er eerst fit-procedures zijn toegepast op de cumulatieve verdeling (Laros & Tellegen, 1991), omdat transformaties die zijn gebaseerd op het direct omzetten van de geobserveerde cumulatieve proporties sterk steekproefgevoelig zijn.

Wanneer leeftijds- of leerjaarnormen worden verstrekt, kan een te breed leeftijds- of leerjaarinterval ertoe leiden dat de prestaties aan het begin van het interval worden onderschat en aan het eind overschat. Vooral bij capaciteitentests voor jonge kinderen kan dit een rol spelen, omdat het verschil in score binnen het tijdsbestek van een jaar een tiental IQ-punten kan bedragen. Echter, ook in de eerste leerjaren van het voortgezet onderwijs kan het verschil tussen twee opeenvolgende leerjaren nog een halve standaarddeviatie bedragen. Dergelijke vertekeningen kunnen gemakkelijk worden voorkomen door het aantal normtabellen bij leeftijds- en leerjaarnormen uit te breiden, door zogenoemde continue normen te gebruiken of door te corrigeren voor de dag waarop de test is afgenomen (zogenoemde

leerdagcorrectie), wat in feite op hetzelfde neerkomt. Als een test alleen is bedoeld voor gebruik in een bepaalde periode van het leerjaar (zoals geldt voor sommige Cito-toetsen), dan moet dit duidelijk worden aangegeven en moeten de normgegevens in de overeenkomstige periode zijn verzameld. Bij leeftijds- en leerjaarnormen moet worden vermeld in welke periode van het jaar de normen zijn verzameld.

De diverse schaaltypen bij zowel rangordenormen als deviatienormen verschillen in het aantal scoreklassen dat wordt gehanteerd. Systemen met veel klassen, zoals percentielen en deviatie-IQ's, maken een fijnmaziger onderscheid mogelijk dan systemen met weinig klassen, zoals A-E-scores en stanines. De keuze voor een bepaald systeem is afhankelijk van het doel en de kenmerken van de test. Wanneer het doel van de test is om een ruime differentiatie tussen personen aan te brengen, zal men kiezen voor een fijn verdeeld systeem, maar uiteraard is dan wel een voorwaarde dat ook de range van mogelijke ruwe scores voldoende differentiatiemogelijkheden biedt. Het is bijvoorbeeld niet zinvol percentielen (die immers 100 scoreklassen tellen) te gebruiken als de minimum ruwe score op een test 0 en de maximum score 20 bedraagt (waarbij dus slechts 21 scoreklassen voorkomen). De keuze voor een grover systeem gaat ten koste van de differentiatie, maar kan de uitkomsten beter toegankelijk maken. Een dergelijke keuze verdient de voorkeur wanneer er slechts globale indicaties worden gevraagd. Welk systeem ook door de testauteur wordt verkozen, de kenmerken en de mogelijke voor- en nadelen van het systeem moeten altijd worden beschreven en de keuze ervan moet worden beargumenteerd.

Voor tests die zijn bedoeld voor onderzoek op groepsniveau zullen in het algemeen geen normtabellen worden geleverd, maar volstaat een vermelding van het gemiddelde en de standaarddeviatie van de normgroep(en).

Aanwijzingen bij vraag 4.5: "Worden er gemiddelden, standaardafwijkingen en gegevens over de scoreverdeling vermeld?"

Voor elke normgroep moeten gemiddelden, standaardafwijkingen en gegevens over de scoreverdeling worden vermeld. Van de verdeling zijn bijvoorbeeld scheefheid, kurtosis, eventueel bimodaliteit en dergelijke van belang en ook of een of meer van deze kenmerken verschillen per normgroep. Zo kan het zijn dat de scores op een vragenlijst in de ene groep min of meer normaal verdeeld zijn, maar dat in een andere groep 50% de laagste score haalt (het zogenoemde bodemeffect). Zo kunnen bij capaciteitentests niet alleen bij lagere opleidingsniveaus bodemeffecten optreden, maar bij hogere opleidingsniveaus ook plafondeffecten. Hierdoor discrimineert de test in die groepen minder goed.

Aanwijzingen bij vraag 4.6: "Worden er gegevens verstrekt over mogelijke verschillen tussen subgroepen (bijvoorbeeld allochtonen-autochtonen, vrouwen-mannen)?"

De gegevens zoals bedoeld in vraag 4.5 moeten ook worden vermeld voor mogelijke subgroepen. Het onderzoek naar en het rapporteren van verschillen tussen subgroepen is om verschillende redenen gewenst:

- Het vaststellen van een mogelijk discriminerend effect (of *adverse impact*).
- Het kan een extra reden vormen voor het uitvoeren van test- en/of itembiasonderzoek.
- Door het beschikbaar stellen van deze gegevens kan de testgebruiker zelf bepalen of hij hiermee bij de interpretatie rekening wil houden. Ook al blijken er (significante) verschillen tussen subgroepen te bestaan, dit betekent niet dat het altijd gewenst is om normtabellen per subgroep te hanteren. Voorbeeld: ADHD blijkt vaker voor te komen bij jongens dan meisjes (stel bij 15% van de jongens en bij 5% van de meisjes). Wanneer er bij een vragenlijst voor de signalering van ADHD normgroepen per sekse worden gebruikt en wanneer in beide groepen de grens wordt gelegd bij het 90e percentiel, dan worden er evenveel jongens als meisjes met ADHD geclassificeerd. Dit is een onderschatting bij jongens en een overschatting bij meisjes. In dit geval is een normtabel voor jongens en meisjes gezamenlijk een betere keus.

Het gaat hier niet om alle mogelijke subgroepen, maar om subgroepen die met het oog op de aard en het doel van de test van belang zijn. Voorbeelden zijn seksegroepen, leeftijdsgroepen en etnische groepen.

Aanwijzingen bij vraag 4.7: "Worden er gegevens verstrekt over de nauwkeurigheid van de meting en de daarbij behorende intervallen?"

Voor de interpretatie van testscores is informatie over de nauwkeurigheid van de meting en de daarbij horende betrouwbaarheidsintervallen van belang. Maten die informatie verschaffen over de nauwkeurigheid van de meting zijn de standaardmeetfout, de standaardschattingsfout en (wanneer het tests geconstrueerd volgens een item-responsmodel betreft) de testinformatiefunctie/standaardfout (Drenth & Sijtsma, 1990, 2006, pp. 226-235; Lord & Novick, 1968).

De *standaardmeetfout* wordt berekend uit de standaarddeviatie en de betrouwbaarheid en is gelijk aan $S_E = S_x \sqrt{1-r_{xx©}}$ en kan worden gebruikt om een betrouwbaarheidsinterval voor de betrouwbare score T te schatten. De betrouwbare score T is niet-observeerbaar en wordt hier eenvoudig geschat door middel van de geobserveerde score X (dus, $\hat{T} = X$). Verder wordt verondersteld dat meetfouten normaal verdeeld zijn. Het betrouwbaarheidsinterval voor T is symmetrisch ten opzichte van de geobserveerde score X, die functioneert als schatting van T. Een 95% betrouwbaarheidsinterval verkrijgt men door voor een gegeven geobserveerde score X te berekenen: $X \pm 1.96S_E$ (dus men verkrijgt de ondergrens van het interval door $1.96S_E$ van X af te trekken en de bovengrens door

$1.96S_E$ bij X op te tellen). Dit interval geeft een indruk van de nauwkeurigheid van de meting en kan worden gebruikt voor het toetsen van hypothesen over iemands betrouwbare score.

De *standaardschattingsfout* kan voor hetzelfde doel worden gebruikt als de standaardmeetfout, dus het schatten van een betrouwbaarheidsinterval voor de betrouwbare score. Het verschil is dat er hierbij een andere schatting van de betrouwbare score wordt gebruikt. De betrouwbare score wordt hier geschat door middel van lineaire regressie, wat resulteert in de formule $\hat{T} = r_{xx\circledcirc}X + (1-r_{xx\circledcirc})\overline{X}$.
Hierin is $\overline{X}$ het gemiddelde van de groep waarin de betrouwbaarheid is bepaald. Deze formule staat wel bekend als *Kelley's formula*. Omdat hij een lineaire regressieschatting geeft van de betrouwbare score, wordt nu de standaardschattingsfout, $S_{est} = S_x\sqrt{r_{xx\circledcirc}}\sqrt{1-r_{xx\circledcirc}}$, gebruikt om een (95%) betrouwbaarheidsinterval voor T te schatten. Dat gebeurt door te berekenen: $\hat{T} \pm 1.96S_{est}$, waarbij men dus de geschatte betrouwbare score uit Kelley's formula gebruikt. Het voordeel van Kelley's formula ten opzichte van de eerste methode is dat er meer informatie wordt gebruikt om de betrouwbare score te schatten. Het idee daarbij is dat naarmate de betrouwbaarheid geringer is, geobserveerde scores minder gewicht krijgen en het groepsgemiddelde juist meer. Omgekeerd geldt dat bij een hoge betrouwbaarheid het groepsgemiddelde nauwelijks meer een rol speelt en de schatting van T bijna volledig door de geobserveerde score X wordt bepaald. Zo spelen dus ook de betrouwbaarheid en de gemiddelde testscore een rol bij de schatting van T, en daarin ligt het verschil met de standaardmeetfout. Het gevolg van het gebruik van meer informatie is dat $S_{est} < S_E$. In woorden: de tweede methode is nauwkeuriger dan de eerste. De tweede formule voor het schatten van de betrouwbare score *(Kelley's formula)* wordt regelmatig in de praktijk gebruikt, maar ten onrechte wordt dan de 'standaardmeetfout' gerapporteerd. Hier moet men dus goed op letten.

Het betrouwbaarheidsinterval is belangrijk als men statistische toetsen wil uitvoeren. We noemen twee mogelijkheden: verschilt de score van persoon v van die van een andere persoon w of van een grensscore X_o?

De eerste mogelijkheid behelst de nulhypothese dat de betrouwbare scores van twee personen, zeg persoon v en persoon w, gelijk zijn; dus $H_0: T_v = T_w$. Getoetst wordt dan dus of het verschil gelijk is aan 0, en beschouwd wordt dan de standaardmeetfout van het verschil $D_{vw} = X_v - X_w$. Deze standaardmeetfout is gelijk aan $S_{E(D)} = \sqrt{2}S_E$
Een 95% betrouwbaarheidsinterval voor het ware verschil, dat wordt geschat met behulp van het geobserveerde verschil, is gelijk aan $D_{vw} \pm 1.96\sqrt{S_{E(D)}}$. Als de waarde 0 in dit interval ligt, wordt de nulhypothese geaccepteerd, maar als de waarde 0 buiten het interval ligt, wordt de nulhypothese verworpen.

Voor de tweede mogelijkheid, het vergelijken van een testscore met een drempelwaarde, geldt dat deze drempelwaarde perfect betrouwbaar is. Dan hoeft alleen te worden nagegaan of de drempelwaarde al dan niet in het betrouwbaarheidsinterval $X \pm 1.96S_E$ ligt. Werkt men met de standaardschattingsfout, dan moet er in de formules steeds S_{est} worden gebruikt in plaats van S_E.

Standaardmeetfouten en bijbehorende betrouwbaarheidsintervallen kan men ook voor andere scores schatten, zoals betrouwbare verschilscores. Overigens moet men er rekening mee houden dat verschilscores, zoals ze bijvoorbeeld in profielen worden gebruikt, notoir onbetrouwbaar zijn, en de betrouwbaarheidsintervallen dus erg lang. In de literatuur is er veel bekend over psychometrische problemen met verschilscores, maar het blijft een heikel probleem (zie bijvoorbeeld Allen & Yen, 1979, pp. 208-211; Drenth & Sijtsma, 2006, pp. 241-243; Murphy & Davidshofer, 1998, pp. 138-139).

Is de test of vragenlijst geconstrueerd met behulp van een item-responsmodel, dan kan de testconstructeur ervoor kiezen de testinformatiefunctie te geven of het omgekeerde hiervan, de standaardfout afhankelijk van de schaal. Met behulp van deze standaardfout kunnen wederom betrouwbaarheidsintervallen worden geschat voor het niveau van de respondent (in de context van de item-responstheorie is dit een andere parameter dan de betrouwbare score). Het verschil met betrouwbaarheidsintervallen op basis van de klassieke standaardmeetfout is dat de betrouwbaarheidsintervallen nu variëren over de schaalwaarden. Hiermee is duidelijk dat niet iedereen even nauwkeurig met dezelfde test wordt gemeten. Het verdient aanbeveling om deze lokale informatie over de meetnauwkeurigheid zowel grafisch (voor de inzichtelijkheid) als numeriek in de vorm van tabellen (om exacte waarden te kunnen aflezen) weer te geven.

Voor een positieve beoordeling op deze vraag zal een testconstructeur minstens een van de drie genoemde mogelijkheden moeten rapporteren: standaardmeetfout, standaardschattingsfout of testinformatiefunctie/voorwaardelijke standaardfout, inclusief een afdoende uitleg voor de testgebruiker over het gebruik van betrouwbaarheidsintervallen. Het wordt ook aanbevolen voor elke ruwe score of standaardscore deze intervallen in de handleiding op te nemen.

Domeingerichte of criteriumgerichte interpretatie

Bij een test die gebruik maakt van grensscores wordt met behulp van deze score(s) het hele scorebereik verdeeld in twee of meer categorieën. Deze categorieën kunnen louter voor beschrijvende doeleinden worden gebruikt, maar meestal zijn ze bedoeld om onderscheid te maken tussen groepen geteste personen die een verschillend programma of een verschillende behandeling krijgen aangeboden of voor wie een verschillende verwachting geldt.
Zo kan een werkgever een grensscore gebruiken om potentiële werknemers te screenen, kan een school grensscores gebruiken om groepen leerlingen verschillende instructieprogramma's aan te bieden, kunnen in de geestelijke gezondheidszorg grensscores worden gebruikt om te beslissen over therapie-indicatie of over de aanwezigheid van een bepaalde vorm van psychopathologie, of kunnen bij certificering door de verantwoordelijke instelling minimum slaagscores worden vastgesteld.

Voor de vaststelling van de grensscore(s) bestaan verschillende methoden, waaraan in de literatuur ook wel wordt gerefereerd als standaardbepalingsprocedures. In het algemeen kan men onderscheid maken tussen procedures die daarbij gebruikmaken van de oordelen van experts (vragen 4.8 tot en met 4.10) en procedures die gebruikmaken van feitelijke gegevens over de relatie van de testscore met een of ander criterium (vragen 4.11 tot en met 4.13). In voorkomende gevallen wordt de grensscore ook wel bepaald door rechtstreeks te verwijzen naar een percentage in de referentiegroep. Bijvoorbeeld: als men tot de 10% laagst scorenden op een schooltoets behoort, komt men in aanmerking voor extra onderwijsfaciliteiten. Of: als men tot de 20% hoogst scorenden op een vragenlijst voor psychopathologie behoort, wordt men in een behandelingsprogramma opgenomen. Aangezien dergelijke grensscores in feite zijn gebaseerd op vergelijking met een normgroep, gelden voor deze scores de eisen zoals die geformuleerd zijn in vraag 4.3.

Domeingerichte Interpretatie

Er worden in de literatuur (o.a. Berk, 1986; Cascio & Aguinis, 2005; Cizek, 1996; Hambleton, Jaeger & Plake, 2000; Livingston & Zieky, 1982; Vos & Knuver, 2000) diverse standaardbepalingsmethoden genoemd waarbij met behulp van een aantal beoordelaars (inhoudsdeskundigen) standaarden of normen worden vastgesteld. Er zijn onder andere *examinee-centered*-methoden en *test-centered*-methoden. In de eerste categorie methoden wordt van de beoordelaars gevraagd om voor ieder item uit een representatieve test aan te geven welk gedrag van een (denkbeeldige of bestaande) student op de grens voldoende/onvoldoende ('zesjesstudent'; Van Berkel, 1999) wordt verwacht. In de tweede groep methoden worden echter reële personen beoordeeld en wordt er een grensscore afgeleid uit de feitelijke scoreverdeling. Via een bepaalde standaardbepalingsmethode kan vervolgens een norm worden vastgesteld.

Aanwijzingen bij basisvraag 4.8: "Is er voldoende overeenstemming tussen de beoordelaars?"

Alleen op basis van een hoge beoordelaarsovereenstemming kan de verkregen norm worden gelegitimeerd. Bij het beoordelen van deze vraag moet erop gelet worden of de beoordelaarsovereenstemming en niet de beoordelaarsbetrouwbaarheid is vermeld. Beoordelaarsovereenstemming heeft betrekking op identieke oordelen van verschillende beoordelaars, terwijl de beoordelaarsbetrouwbaarheid betrekking heeft op relatief identieke oordelen van verschillende beoordelaars, waarbij het absolute niveau van de beoordelingen niet gelijk hoeft te zijn.
Afhankelijk van het meetniveau van de data kunnen verschillende coëfficiënten worden gebruikt. Voor data van nominaal niveau wordt veelal coëfficiënt kappa, κ, gebruikt en voor data van ordinaal niveau wordt veelal de gewogen coëfficiënt kappa, κ_w, gebruikt. Voor de interpretatie van de hoogte van coëfficiënt kappa worden er in de literatuur geen eenduidige normen genoemd, al wordt een κ van .60 doorgaans als minimum beschouwd om van een acceptabele beoordelaars-overeenstemming te kunnen spreken. Shrout (1998) benoemt kappa's in de range van .61 – .80 als *moderate* en in de range van .81 – 1.00 als *substantial*. Bij de interpretatie moet enige voorzichtigheid worden betracht, omdat de prevalentie of base rate de hoogte van kappa kan beïnvloeden.

Voor data van intervalniveau is de intraklassecorrelatiecoëfficiënt (ICC) een veelgebruikte maat. De ICC is een verhouding van variantiecomponenten, waarbij in het geval van de beoordelaarsovereenstemmingscoëfficiënt de foutenvariantie wordt gevormd door de variantie binnen gepaarde waarnemingen. Bij de rapportage van deze coëfficiënt moet uitgegaan worden van één beoordelaar. Shrout (1998) wijst op de vergelijkbaarheid van kappa en ICC en stelt dat net als bij kappa een ICC van > .80 als 'goed' kan worden geïnterpreteerd.

Op basis van het bovenstaande zijn voor de interpretatie van de hoogte van coëfficiënt kappa en de intraklassecorrelatiecoëfficiënt de volgende normen opgesteld:

Coëfficiënt kappa of intraklassecorrelatiecoëfficiënt.	κ of ICC ≥ .80 .60 ≤ κ of ICC < .80 κ of ICC < .60	goed voldoende onvoldoende

Aanwijzingen bij vraag 4.9: "Zijn de procedures op grond waarvan de grensscores zijn bepaald correct?"

Om te kunnen beoordelen of de normen te legitimeren zijn, is het van belang dat de testauteur de gevolgde procedures nauwkeurig beschrijft. Bij het beoordelen van deze vraag moet worden gecontroleerd of aan onderstaande aspecten is voldaan:

- Zijn alle gevolgde stappen en beslissingen in overeenstemming met de in de methode gedefinieerde definities en procedures?
- Zijn alle in de methode gedefinieerde stappen constant gehouden? Hieronder vallen ook zaken als instructies, materialen en beschikbaar gestelde statistische informatie over prestatie-verdelingen.

Als aan een van beide aspecten niet is voldaan, dan moet het oordeel 'onvoldoende' worden toegekend. Het oordeel 'goed' mag alleen worden toegekend als de testauteur beargumenteert waarom een bepaalde standaardbepalingsprocedure gekozen is, aangeeft hoe er tijdens de standaardbepalingsprocedure is omgegaan met mogelijke inconsistenties in beoordelingen, en als de gevolgde procedure voldoet aan bovengenoemde aspecten. In andere gevallen is het oordeel 'voldoende'.

Aanwijzingen bij vraag 4.10: "Zijn de beoordelaars naar behoren geselecteerd en getraind?"

Omdat de beoordelaars een prominente rol spelen bij een standaardbepalingsmethode, is het niet wenselijk om ieder willekeurig persoon erbij te betrekken. De potentiële beoordelaars moeten ten minste kennis hebben van het domein waarop de test betrekking heeft en het is wenselijk dat de beoordelaar training heeft gevolgd over het beoordelen van (werk van) getesten. Daarnaast is het belangrijk dat elke beoordelaar de standaardbepalingsmethode begrijpt die gevolgd gaat worden, zodat er geen verschillen in beoordelingen kunnen optreden doordat beoordelaars de methode anders toepassen. Ook dit kan worden bevorderd door een training aan te bieden. Om te kunnen beoordelen of de testauteur de beoordelaars zorgvuldig heeft gekozen is een beschrijving van de selectieprocedure en de aangeboden trainingen aan de beoordelaars noodzakelijk.

Criteriumgerichte interpretatie

Grensscores kunnen op velerlei manieren empirisch worden onderbouwd. Een gemeenschappelijk kenmerk is echter dat in al deze gevallen niet alleen de testscores van de getesten beschikbaar zijn, maar ook gegevens over het te voorspellen criterium en daarmee over de relatie test-criterium. In feite betreft dit onderzoek naar de criteriumvaliditeit, dat echter ook de functie heeft op empirische wijze een norm vast te stellen. In deze laatste functie wordt dit onderzoek hier beoordeeld. Enkele voorbeelden:

- Op basis van onderzoek waarin de relatie tussen testscores en arbeidsprestaties is vastgesteld, kunnen bij personeelsselectie aftestgrenzen worden bepaald en/of verwachtingstabellen worden geconstrueerd.
- In de klinische psychologie kunnen ROC-curves en sensitiviteits- en specificiteitswaarden gebaseerd op de relatie tussen testscores en onafhankelijk vastgestelde criteria worden gebruikt om de meest gunstige grensscores te bepalen.
- Bij het toekennen van licenties of diploma's kan de zak/slaaggrens worden bepaald door na te gaan bij welke testscore de gunstigste verhouding bestaat tussen deelnemers die in de praktijk succesvol en onsuccesvol blijken te zijn.

Aanwijzingen bij basisvraag 4.11: "Rechtvaardigen de onderzoeksresultaten het gebruik van grensscores?"

Wanneer grensscores empirisch worden onderbouwd, zal de onderzoeker het bewijs moeten leveren voor de bruikbaarheid van de gekozen grensscore. In bijvoorbeeld een selectiesituatie kunnen dit gegevens zijn over de succesratio en in een klinische situatie over de sensitiviteit en de specificiteit (zie ook de Aanwijzingen bij de vragen 7.1 en 7.2). Voor de gewenste hoogte van deze maten kunnen geen algemene aanwijzingen worden gegeven, niet alleen omdat 'wat hoog of laag is' per te voorspellen criterium kan verschillen, maar ook omdat de predictieresultaten worden beïnvloed door andere variabelen zoals de 'base-rate' of de prevalentie. Het wordt daarom aan de expertise van de beoordelaar overgelaten de verschillende factoren tegen elkaar af te wegen en een oordeel te geven over de hoogte van de gevonden resultaten.

Aanwijzingen bij vraag 4.12: "Is de onderzoeksgroep in overeenstemming met het bedoelde gebruik?"

Het onderzoek waarin de grensscore wordt bepaald, moet betrekking hebben op de populatie waarvoor de test wordt gebruikt. Wanneer de onderzoeksgroep heterogener van samenstelling is dan de populatie waarvoor de test zal worden gebruikt (en binnen welke populatie uiteindelijk beslissingen zullen worden genomen), zal dit niet alleen tot geflatteerde resultaten leiden, maar mogelijk ook tot andere grensscores, dan wanneer het onderzoek wel bij de juiste groep had plaatsgevonden. Om een en ander te kunnen beoordelen, moet de steekproef zijn beschreven met behulp van mogelijk relevante psychologische of demografische variabelen.

Aanwijzingen bij vraag 4.13: "Is de onderzoeksgroep groot genoeg?"

Grensscores zijn te beschouwen als 'normale' punten in een scoreverdeling waaraan een bijzondere betekenis wordt toegekend. Voor de nauwkeurigheid waarmee deze punten worden vastgesteld, gelden dezelfde eisen als die welke gelden voor normtabellen (waarbij de nauwkeurigheid voornamelijk wordt bepaald door de grootte van de groep). Wel is het zo dat het bij de bepaling van een of meerdere grensscores slechts gaat om een beperkt aantal punten, terwijl bij normtabellen de nauwkeurigheid van de hele scoreverdeling in het geding is. De eisen die gesteld worden aan de omvang van de onderzoeksgroep kunnen daarom worden versoepeld ten opzicht van de eisen zoals die gelden voor normgerichte interpretatie (zie de aanwijzingen bij vraag 4.3.a). Ervan uitgaand dat grensscores alleen worden bepaald in situaties waarin het gaat om 'belangrijke beslissingen op individueel niveau' (voor een omschrijving hiervan zie de aanwijzingen bij vraag 4.3.a), wordt een onderzoeksgroep bestaande uit minstens 300 personen als 'goed', een groep bestaande uit minstens 200 personen als 'voldoende' en een onderzoeksgroep bestaande uit minder dan 200 personen als 'onvoldoende' beoordeeld.

Vaststelling eindoordeel voor criterium 4 **Normen** **Normgerichte interpretatie**		
Alle drie de basisvragen (4.1, 4.2 en 4.3) worden met '3' beoordeeld.	Somscore 4.4 t/m 4.7* ≥ 9	goed
	Somscore 4.4 t/m 4.7* ≤ 8	voldoende
Basisvraag 4.1 en één van de basisvragen 4.2 en 4.3 wordt met '3' beoordeeld, waarbij de andere basisvraag met '2' wordt beoordeeld.	Somscore 4.4 t/m 4.7* ≥ 9	voldoende
	Somscore 4.4 t/m 4.7* ≤ 8	onvoldoende
Basisvraag 4.1 wordt met '3' beoordeeld; beide basisvragen 4.2 en 4.3 worden met '2' beoordeeld.		onvoldoende
Minstens een van de drie basisvragen wordt met '1' beoordeeld.		onvoldoende
* Voor vraag 4.7 a t/m c geldt de hoogst scorende subvraag.		
Domeingerichte interpretatie		
Alle drie de basisvragen (4.1, 4.2 en 4.8) worden met '3' beoordeeld.	Somscore 4.9 en 4.10 ≥ 5	goed
	Somscore 4.9 en 4.10 = 3 of 4	voldoende
	Somscore 4.9 en 4.10 = 2	onvoldoende
Basisvraag 4.1 wordt met '3' beoordeeld en één of beide basisvragen 4.2 en 4.8 worden met '2' beoordeeld.	Somscore 4.9 en 4.10 ≥ 5	voldoende
	Somscore 4.9 en 4.10 ≤ 4	onvoldoende
Minstens één van de drie basisvragen wordt met '1' beoordeeld.		onvoldoende
Criteriumgerichte interpretatie		
Alle drie de basisvragen (4.1, 4.2 en 4.11) worden met '3' beoordeeld.	Somscore 4.12 en 4.13 ≥ 5	goed
	Somscore 4.12 en 4.13 = 3 of 4	voldoende
	Somscore 4.12 en 4.13 = 2	onvoldoende
Basisvraag 4.1 wordt met '3' beoordeeld en één of beide basisvragen 4.2 en 4.11 worden met '2' beoordeeld.	Somscore 4.12 en 4.13 ≥ 5	voldoende
	Somscore 4.12 en 4.13 ≤ 4	onvoldoende
Minstens één van de drie basisvragen wordt met '1' beoordeeld.		onvoldoende

5 Betrouwbaarheid

De klassieke testtheorie veronderstelt dat een testscore (X) is opgebouwd uit een betrouwbaar deel, ook wel ware score of betrouwbare score (T) genoemd, en een deel dat te wijten is aan de invloed van toevallige meetfouten. Dit laatste deel wordt meetfout (E) genoemd. De testscore is de som van de betrouwbare score en de meetfout: $X = T + E$. Het zou ideaal zijn als alleen de betrouwbare score werd gemeten, maar de realiteit is dat testscores ook uit meetfouten bestaan. Doel van betrouwbaarheidsanalyse is om de invloed van meetfouten op de testscores te schatten.

De variantie van de testscores in een groep van personen (S_X^2) is opgebouwd uit betrouwbare variantie (S_T^2) en foutenvariantie (S_E^2), zodat $S_X^2 = S_T^2 + S_E^2$. In zijn meest basale vorm geeft de foutenvariantie de spreiding weer die het gevolg is van toevallige meetfouten, zodat de betrouwbare variantie alle systematische verschillen tussen respondenten weergeeft. De paralleltestbetrouwbaarheid is de verhouding van deze betrouwbare variantie en de variantie van de testscores.

Naast de interpretatie van meetfouten als toevallige scorecomponenten is er een andere interpretatie die zegt dat meetfouten alle onbedoelde componenten van de testscore bevatten, om te beginnen de toevallige, maar vervolgens ook de onbedoelde, systematische componenten. De betrouwbare variantie geeft in dit geval de spreiding van de bedoelde scorecomponenten, en de foutenvariantie de spreiding die het gevolg is van de onbedoelde componenten, inclusief de toevallige meetfouten. Hiervan wordt een schatting gekregen door gebruik te maken van technieken uit de generaliseerbaarheidstheorie, de item-responstheorie en de structurele vergelijkingsmodellen. In beide gevallen is het belangrijk (maar niet altijd strikt noodzakelijk) hiervoor speciaal verzamelde gegevens te analyseren.

Een voorbeeld van een test waarmee niet alleen de bedoelde eigenschap maar ook een andere eigenschap wordt gemeten, is een rekentest waarvan de spreiding in de testscores niet alleen van rekenvaardigheid (bedoeld) afhankelijk is, maar ook van taalvaardigheid en toeval (beide onbedoeld). De eerste vorm van betrouwbaarheid is gelijk aan de verhouding van de variantie als gevolg van verschillen tussen respondenten in rekenvaardigheid en taalvaardigheid samen, en de variantie van de testscores. De tweede vorm is gelijk aan de verhouding van de variantie van alleen de bedoelde rekenvaardigheid, en de variantie van de testscore.

De bronnen van de foutenvariantie kunnen zeer verschillend zijn en hoeven niet alleen betrekking te hebben op onbedoelde psychologische eigenschappen, zoals de taalvaardigheid in het voorbeeld. Een alternatieve mogelijkheid is dat men zich afvraagt in hoeverre een testscore herhaalbaar is over een bepaalde periode. Zo kan 'stemming' op hetzelfde tijdstip worden gemeten met twee, als parallelle instrumenten bedoelde vragenlijsten, en kan blijken dat de meting op dat tijdstip zeer betrouwbaar was. Ligt er tussen de twee afnamen (van dezelfde test, en niet twee verschillende testversies) echter een lange tussenpoos, dan kan blijken dat de correlatie tussen de twee testscores laag is. De conclusie is dan dat de verschillen tussen respondenten over een langere periode zijn gemeten, maar voor een klein deel systematisch zijn. Dus is de betrouwbaarheid – hier de test-hertestbetrouwbaarheid – te gering voor het generaliseren van de testscore over tijdsperiodes zoals in het onderzoek betracht.

De indices voor betrouwbaarheid met vermelding van de foutenbron maken het dus mogelijk over een voor een bepaald doel betrouwbare test te spreken. Met behulp van de traditionele betrouwbaarheidsmaten, zoals verwoord in vraag 5.2, wordt in feite de generaliseerbaarheid van scores over versies (de paralleltestbetrouwbaarheid; betrouwbaarheidsschattingen op basis van inter-itemrelaties geven hiervan een schatting, waarover straks meer), tijdstippen (de test-hertestbetrouwbaarheid) en beoordelaars (de interbeoordelaarsbetrouwbaarheid) vastgesteld. Uit deze opsomming wordt duidelijk – maar om misverstand te vermijden wordt het nog eens gezegd – dat dé betrouwbaarheid van een test niet bestaat: we onderscheiden vormen van betrouwbaarheid naar de aard van de variantiebron die in het betrouwbaarheidsonderzoek wordt geanalyseerd.

Ook is het van belang te onderkennen dat de uitkomsten van het betrouwbaarheidsonderzoek voor een bepaalde test afhankelijk zijn van de onderzochte groep. Meet de test in twee groepen dezelfde eigenschap, dan is de betrouwbaarheid het grootst in de groep met de grootste variantie in de testscores. Meet de test echter in de ene groep alleen de bedoelde eigenschap en in de andere groep behalve de bedoelde eigenschap ook nog een onbedoelde eigenschap – denk aan het voorbeeld met rekenvaardigheid en taalvaardigheid –, dan is de validiteit van de test in het geding, en is het af te raden de scores van personen uit de twee groepen met elkaar te vergelijken.

Hoewel een test vaak uit meer dan één onderdeel (schalen, subtests) bestaat, geeft de beoordelaar in het algemeen één beoordeling voor het criterium betrouwbaarheid, die een samenvatting geeft van de resultaten op de verschillende onderdelen. Dit is bijvoorbeeld het geval bij vragenlijsten die uit diverse schalen bestaan, zoals de *BIT*, de *EPPS* en de *NPV*, en bij testseries die uit verscheidene, in principe onafhankelijk af te nemen subtests bestaan, zoals de *DAT*, de *DVMH* en de *MCT-M*. In dergelijke gevallen geeft de laagste coëfficiënt de doorslag in de beoordeling. Wanneer het echter een duidelijke negatieve uitzondering betreft (bijvoorbeeld: op één na alle subtests 'goed' en één subtest 'onvoldoende'), mag de hogere beoordeling worden aangehouden (in dit voorbeeld: 'goed'), en kan als voetnoot bij de beoordeling de uitzondering worden vermeld. Een andere situatie kan ontstaan wanneer de scores op de subtests worden gesommeerd tot een totaalscore, zoals bij sommige intelligentietests het geval is. Hierbij kunnen drie mogelijkheden worden onderscheiden:

- Als slechts de interpretatie van de totaalscore van belang is, hoeft uiteraard slechts de betrouwbaarheid van deze score worden beoordeeld.
- Als door de testauteur wordt aangegeven dat de totaalscore

weliswaar het belangrijkst is, maar dat ook interpretatie van factor- of subtestscores mogelijk is, worden deze laatste beoordeeld met de beoordelingscriteria die voor één niveau lager gelden dan die voor de totaalscore (zie Aanwijzingen 5.2). Wanneer de totaalscore in de categorie 'belangrijk' valt, vallen factor- of subtestscores in de categorie 'minder belangrijk'. Hierbij zal het meestal voorkomen dat de scores op de subtests minder betrouwbaar zijn dan de totaalscore, maar kan de beoordeling voor beide gelijk zijn.

- Als door de testauteur geen onderscheid wordt gemaakt tussen het belang van de totaalscore en de factor- of subtestscores, worden beide op dezelfde wijze (als even belangrijk) beoordeeld.

Wanneer een en ander leidt tot een verschillend oordeel over de betrouwbaarheid van factoren/subtests en totaalscore, kan dit als voetnoot bij de beoordeling worden vermeld. Ook wanneer voor verschillende groepen de betrouwbaarheden worden gegeven én deze betrouwbaarheden verschillen, wordt slechts één beoordeling gegeven. Daarbij moet het resultaat van groepen die het belangrijkste gebruiksdoel vertegenwoordigen, zwaarder worden gewogen. Mutatis mutandis geldt tevens dezelfde regel als hierboven gegeven: de laagst gevonden betrouwbaarheid geeft de doorslag, behalve wanneer het een duidelijke uitzondering betreft.

Vragen voor criterium 5 Betrouwbaarheid					
		onv.	**vold.**	**goed**	
Basisvraag 5.1	Worden er gegevens over de betrouwbaarheid verstrekt? Bij negatieve beoordeling van deze vraag kan men direct doorgaan naar criterium 6.	1		3	
5.2	Zijn de resultaten voldoende, gelet op het beoogde type beslissingen dat met behulp van de test moet worden genomen?				
5.2.a	Paralleltestbetrouwbaarheid.	1	2	3	n.v.t.
5.2.b	Betrouwbaarheid op basis van inter-itemrelaties.	1	2	3	n.v.t.
5.2.c	Test-hertestbetrouwbaarheid.	1	2	3	n.v.t.
5.2.d	Interbeoordelaarsbetrouwbaarheid.	1	2	3	n.v.t.
5.2.e	Methoden op basis van item-responstheorie.	1	2	3	n.v.t.
5.2.f	Methoden op basis van generaliseerbaarheidstheorie of structurele vergelijkingsmodellen.	1	2	3	n.v.t.
5.3	Wat is de kwaliteit van het onderzoek naar de betrouwbaarheid?				
5.3.a	Zijn de procedures op basis waarvan de betrouwbaarheidsgegevens zijn berekend correct?	1	2	3	
5.3.b	Zijn de steekproeven op basis waarvan de betrouwbaarheidsgegevens zijn berekend in overeenstemming met het beoogde testgebruik?	1	2	3	
5.3.c	Maken de gegevens die worden verstrekt een gefundeerd oordeel over de betrouwbaarheid mogelijk?	1	2	3	

Aanwijzingen bij basisvraag 5.1: "Worden er gegevens over de betrouwbaarheid verstrekt?"
Hierbij valt te denken aan betrouwbaarheidscoëfficiënten en aan de resultaten van generaliseerbaarheidsonderzoeken. Ook kan er op basis van de item-responstheorie een betrouwbaarheidscoëfficiënt, een tabel of figuur met standaardfouten of een informatiefunctie worden gerapporteerd.

Aanwijzingen bij vraag 5.2: "Zijn de resultaten voldoende, gelet op het beoogde type beslissingen dat met behulp van de test moet worden genomen?"
Over de gewenste hoogte van een betrouwbaarheidscoëfficiënt of een vergelijkbare maat kan geen algemene uitspraak worden gedaan, omdat het doel van het testgebruik hierop van invloed is. Nunnally en Bernstein (1994, p. 265) geven aan dat een test die wordt gebruikt voor belangrijke beslissingen, een betrouwbaarheid van minstens .90 moet bezitten. Met belangrijke beslissingen wordt bedoeld: beslissingen die op basis van de testscores worden genomen, die in principe, of op korte termijn, onomkeerbaar zijn, en die voor een belangrijk deel buiten de geteste persoon om worden genomen. Met deze waarde als uitgangspunt zijn de regels die in onderstaande tabel staan, opgesteld.

Tests voor belangrijke beslissingen op individueel niveau (bijvoorbeeld personeelsselectie, verwijzing naar speciaal onderwijs, opname/ontslag kliniek).
goed: $r \geq .90$
onvoldoende: $r < .80$
voldoende: $.80 \leq r < .90$
Tests voor minder belangrijke beslissingen op individueel niveau (bijvoorbeeld voortgangscontrole, in het algemeen beschrijvend gebruik zoals bij beroepskeuzebegeleiding en therapie-indicatie).
goed: $r \geq .80$
onvoldoende: $r < .70$
voldoende: $.70 \leq r < .80$
Tests voor onderzoek op groepsniveau (bijvoorbeeld meting van teamtevredenheid, klimaat in de klas, of organisatiecultuur).
goed: $r \geq .70$
onvoldoende: $r < .60$
voldoende: $.60 \leq r < .70$

Wanneer er variantiecomponenten in een generaliseerbaarheidsonderzoek worden geschat, moeten er grenzen worden aangehouden die met bovenstaande waarden overeenkomen. Dit geldt eveneens voor een betrouwbaarheidscoëfficiënt op basis van de item-responstheorie en structurele vergelijkingsmodellen. Voor de standaardfouten en de informatiefunctie is het lastiger om eenvoudige vuistregels op te stellen. Hier wordt onder 5.2.e nog op teruggekomen.

Veelal zullen er bij een test meer dan een van de onder 5.2.a t/m f genoemde indices worden vermeld. Bij de bepaling van het eindoordeel van de betrouwbaarheid moet díé coëfficiënt het zwaarst worden gewogen die het meest in overeenstemming is met het doel waarvoor de test wordt gebruikt. Wanneer men bijvoorbeeld over tijd wil voorspellen, dan is in eerste instantie een index van stabiliteit nodig.

Aanwijzingen bij vraag 5.2.a: "Paralleltestbetrouwbaarheid."
De betrouwbaarheid als verhouding van alle systematische variantie en de variantie van de testscores kan worden geschat met behulp van de paralleltestbetrouwbaarheid. Tests zijn parallel wanneer hun testscores in dezelfde groep dezelfde gemiddelden, varianties, en correlaties met andere variabelen hebben. Zijn deze kenmerken aanwezig, dan is de correlatie tussen de testscores gelijk aan de betrouwbaarheid van de afzonderlijke tests. Als de testversies niet parallel zijn, geeft hun correlatie een onderschatting van de paralleltestbetrouwbaarheid. Deze correlatie kan dan ook worden opgevat als een maat voor de generaliseerbaarheid over verschillende, niet parallelle testversies.

De paralleltestbetrouwbaarheid kan van belang zijn bij pure *speedtests*. De correlatie tussen testhelften die gevormd wordt op basis van halve testtijd en/of halvering van het testmateriaal, kan als paralleltestbetrouwbaarheid worden opgevat, maar wel voor een halve test. Vervolgens kan de correctie voor testlengte worden toegepast om een schatting te verkrijgen van de betrouwbaarheid van de gehele test.

Aanwijzingen bij vraag 5.2.b: "Betrouwbaarheid op basis van inter-itemrelaties."
Cronbachs (1951) alfa is gebaseerd op de covarianties tussen de items in de test en wordt bijna standaard gebruikt om de betrouwbaarheid van de testscore te schatten. Daarbij zijn er drie zaken van belang. Ten eerste is alfa een ondergrens voor de paralleltestbetrouwbaarheid. De waarde van alfa is dus lager dan de echte betrouwbaarheid van de test (Novick & Lewis, 1967). Ten tweede zijn er veel alternatieve methoden die veel op alfa lijken en ook nog eens een schatting van de betrouwbaarheid geven die daar dichter bij ligt dan alfa. Een voorbeeld is Guttmans lambda2 (Guttman, 1945). Altijd ligt de waarde van Guttmans lambda2 dichter bij de betrouwbaarheid dan die van alfa, hoewel de verschillen wel klein zijn. Ten derde wordt er in een groot deel van de literatuur over testtheorie gerapporteerd dat alfa een maat is voor de interne consistentie van de test. Vandaar de veel gebruikte aanduiding van

interne-consistentiecoëfficiënt. Dit betekent dat een hogere waarde van alfa zou aangeven dat de items in hogere mate dezelfde eigenschap meten. In de psychometrische literatuur is echter bekend dat deze interpretatie onjuist is: een lage waarde van alfa kan behoren bij zowel een één-factoriële als een meer-factoriële test en hetzelfde kan gezegd worden van een hoge waarde van alfa (Drenth & Sijtsma, 2006). Wil men iets over de samenstelling van de test zeggen, dan zijn er technieken als factoranalyse en principale componentenanalyse aangewezen.

Ook wordt er voor de schatting van de betrouwbaarheid nog wel eens de split-halfcoëfficiënt of splitsingsbetrouwbaarheid gebruikt. Deze methode wordt afgeraden, omdat de uitkomsten afhankelijk zijn van de toevallige verdeling van items over testhelften. Bovendien is het gemiddelde van de split-halfcoëfficiënten gebaseerd op alle mogelijke indelingen van de test in twee helften precies gelijk aan Cronbachs alfa (Lord & Novick, 1968). Om die reden zou men dus in ieder geval beter alfa kunnen nemen.

Een methode die nog weinig wordt gebruikt, is de *greatest lower bound* (*glb*; grootste ondergrens; Ten Berge & Sočan, 2004). Deze methode zoekt onder de assumpties van de klassieke testtheorie, bij een gegeven tabel met inter-itemcovarianties, de laagst mogelijke waarde van de betrouwbaarheid. Het resultaat is een *worstcase-scenario* voor de betrouwbaarheid, dat overigens altijd groter blijkt te zijn dan Cronbachs alfa en resultaten van vergelijkbare methoden. Aangezien de *glb* dichterbij de echte betrouwbaarheid ligt, is hij hiervan een betere onderschatting dan de andere, hierboven genoemde methoden.

Bij tests met een snelheidskarakter en bij tests met zogenoemde heterogene schalen zijn maten voor 'interne consistentie', zoals Cronbachs alfa, niet zinvol. Hetzelfde geldt voor zogenoemde causale indicatoren (zie voor het onderscheid tussen causale en effect-indicatoren bijvoorbeeld Nunnally & Bernstein, 1994) en voor de hiermee vergelijkbare zogenoemde *emergent traits* (zie voor het onderscheid tussen *emergent traits* enerzijds, en *single* of *multi-faceted traits* anderzijds, bijvoorbeeld Schneider & Hough, 1995), waarbij de items eveneens niet onderling gecorreleerd hoeven te zijn. In al deze gevallen kan mogelijk een van de andere betrouwbaarheidsmaten uitkomst bieden. Voor pure speedtests kan de betrouwbaarheid worden vastgesteld met de paralleltestmethode (zie de opmerking daarover bij ad 5.2.a) of de test-hertestmethode. Ook veel tests met een *power*-karakter kennen echter een tijdslimiet. Vooral wanneer een bepaald percentage van de geteste personen niet aan een deel van de opgaven is toegekomen, mogen interne-consistentiematen niet zonder meer worden berekend, omdat ze een overschatting van de betrouwbaarheid kunnen opleveren. In deze gevallen kan er een schatting van de betrouwbaarheid worden verkregen door de test te splitsen in twee helften (bijvoorbeeld in even en oneven items), en de correlatie tussen de scores op deze helften (waarbij voor elke helft de halve testtijd wordt gegeven), te corrigeren voor testlengte. Bij een niet te sterke tempofactor (maximaal 30% van de geteste personen heeft het laatste item niet af) is een andere mogelijkheid het toepassen van een correctieformule voor de betrouwbaarheid (De Zeeuw, 1978). Ook kan de betrouwbaarheid in dat geval worden geschat over dat deel van de test dat door ten minste 90% van de geteste personen is gemaakt.

Voor heterogene schalen en causale indicatoren kan de test-hertestmethode worden gebruikt, maar voor een dergelijk type test geldt dat correlaties met externe variabelen de betrouwbaarheidsmaten kunnen vervangen. Vooral bij causale indicatoren is een goede omschrijving van het domein essentieel (zie 'Uitgangspunten van de testconstructie').

Ook voor tests die gebruikmaken van instap- en/of afbreekregels en meer in het algemeen voor adaptieve tests geldt dat methoden zoals Cronbachs alfa niet zonder meer kunnen worden gebruikt (ten onrechte wordt door sommige auteurs gesteld dat dit wel zou kunnen, omdat de adaptieve score bijvoorbeeld bij simulatie hoog correleert met de score op de gehele test, als die zou zijn voorgelegd). In dit geval zullen item-responsmodellen moeten worden toegepast of een methode zoals gebruikt door Laros en Tellegen (1991). Hierbij wordt de betrouwbaarheid geschat op basis van verschillende afbreekregels en de correlaties van deze scores met een criteriumvariabele.

Aanwijzingen bij vraag 5.2.c: "Test-hertestbetrouwbaarheid."
De generaliseerbaarheid over tijd wordt bepaald door test-hertestcorrelaties. Een test wordt herhaald bij dezelfde onderzoeksgroep, waarbij het tijdsinterval en eventueel relevante gebeurtenissen in dat interval nauwkeurig moeten worden vermeld. De lengte van het tijdsinterval en de hoogte van de correlatie bepalen de mate waarin de testprestatie over tijd kan worden gegeneraliseerd. Test-hertestcoëfficiënten zijn vooral gewenst wanneer de test is bedoeld voor voorspelling over tijd, maar bijvoorbeeld ook wanneer men mag verwachten dat het te meten construct gerelateerd is aan leeftijd (zoals bij intelligentietests voor kinderen).

Aanwijzingen bij vraag 5.2.d: "Interbeoordelaars-betrouwbaarheid."
Vooral bij observatie- en beoordelingsinstrumenten is het van belang of de scores over observatoren/beoordelaars kunnen worden gegeneraliseerd. Maten die hiervoor worden gebruikt, zijn overeenstemmingsindexen zoals Cohen's kappa (Cohen, 1960, 1969), de coëfficiënt van Gower (1971), de identiteitscoëfficiënt (Zegers & Ten Berge, 1985) en andere maten die rekening houden met verschillen tussen zowel gemiddelden als varianties van beoordelaars (voor een overzicht zie Zegers, 1989). Ook variantie- en factoranalytisch onderzoek naar de structuur van het observator-/beoordelaarsgedrag kan hier relevant zijn.

Het is bij de beoordeling van de vermelde waarden van belang te letten op het type coëfficiënt dat is gebruikt. Zo wordt er in de literatuur (Heuvelmans & Sanders, 1993) bijvoorbeeld onderscheid gemaakt tussen beoordelaars*overeenstemming* en beoordelaars-

betrouwbaarheid. Het verschil is dat in de noemer van de formule voor beoordelaarsbetrouwbaarheid de variantiecomponent voor beoordelaars is weggelaten. Deze coëfficiënt zal daardoor hoger uitvallen dan die voor beoordelaarsovereenstemming. Vergelijkbaar zijn de verschillen in transformaties die op de scores worden toegepast voor de diverse door Zegers (1989) genoemde coëfficiënten. Terzijde moet hier worden opgemerkt dat een hoge interbeoordelaarsbetrouwbaarheid wel een noodzakelijke voorwaarde is voor, maar niet hetzelfde is als een hoge testbetrouwbaarheid.

Aanwijzingen bij vraag 5.2.e: "Methoden op basis van item-responstheorie."

In de literatuur over item-responstheorie worden twee benaderingen genoemd voor het vaststellen van de nauwkeurigheid van een testscore. De eerste benadering sluit nauw aan bij de klassieke definitie. Er zijn twee methoden. De eerste methode geeft de betrouwbaarheid van de geschatte latente trek, die in de item-responstheorie de geschatte betrouwbare score (*true score*), dus de testscore, vervangt (Embretson & Reise, 2000). De tweede methode is bekend onder de naam *rho*, en is door Mokken (1971) voorgesteld. Deze methode is gebaseerd op informatie over de individuele items, en levert een schatting van de betrouwbaarheid van de testscore, wanneer er aan voorwaarden is voldaan die typisch zijn voor item-responsmodellen. De regels voor interpretatie van beide betrouwbaarheidsmethoden zijn dezelfde als de regels die bij de Aanwijzingen van vraag 5.2 werden genoemd.

De tweede benadering is principieel anders dan wat we onder de klassieke testtheorie gewend zijn, omdat er hierbij een schatting wordt gegeven van de nauwkeurigheid van de meting als functie van de schaal van de latente trek. Het resultaat is een functie in plaats van een coëfficiënt. Deze functie is de zogenoemde testinformatiefunctie. Hier zijn diverse varianten op te geven. Ook kan de testinformatiefunctie worden omgerekend naar een functie die de standaardfouten geeft die behoren bij de schatting van de latente-trekwaarden. Hiermee kunnen bij elke latente-trekwaarde betrouwbaarheidsintervallen voor de echte latente-trekwaarde worden berekend. Bij verschillende latente-trekwaarden horen betrouwbaarheidsintervallen met verschillende lengte, wat tot uitdrukking brengt dat niet elke schaalwaarde, en dus niet elke persoon, even nauwkeurig wordt gemeten. Als bijvoorbeeld de meeste items voor de groep die met de test wordt gemeten van gemiddelde moeilijkheid zijn, dan zijn de betrouwbaarheidintervallen in het midden van de schaal korter (daar wordt relatief nauwkeurig gemeten) dan aan de uiteinden. Hierin komt tot uitdrukking dat men mensen voor wie alle items moeilijk zijn (wat resulteert in een lage testscore), of juist gemakkelijk (hoge testscore), niet nauwkeurig kan meten. De items zijn voor deze personen eenvoudig niet geschikt. In de klassieke testtheorie wordt voor het bepalen van dergelijke betrouwbaarheidsintervallen de standaardmeetfout gebruikt, en verondersteld wordt dat deze voor ieder individu geldt. Daarmee zijn de betrouwbaarheidsintervallen dus voor iedereen even lang, ongeacht hun positie op de schaal.

Concrete aanwijzingen voor de hoogte van informatiefuncties of de lengte van betrouwbaarheidsintervallen zijn moeilijk te geven, omdat ze afhankelijk zijn van de toepassing van de test en de ernst van de beslissingen die genomen moeten worden met behulp van de testscores. Men kan hiervoor de literatuur over item-responstheorie raadplegen. Zie bijvoorbeeld Reise en Havilund (2005, p. 234), waar een illustratie wordt gegeven van het gebruik van de informatiefunctie, en Langenbucher et al. (2004), waarin duidelijk wordt gemaakt dat een schaal niet overal even betrouwbaar meet.

Aanwijzingen bij vraag 5.2.f: "Methoden op basis van generaliseerbaarheidstheorie of structurele vergelijkingsmodellen."

Tot slot wordt de mogelijkheid genoemd om de betrouwbaarheid te schatten met behulp van structurele vergelijkingsmodellen. Hierin spelen confirmatorische factormodellen een belangrijke rol. Deze methode wordt op dit moment nog maar weinig toegepast, maar zie bijvoorbeeld Raykov (1997) en Green en Yang (2009).

Aanwijzingen bij vraag 5.3.a: "Zijn de procedures op basis waarvan de betrouwbaarheidsgegevens zijn berekend correct?"

Voor de genoemde vormen van betrouwbaarheid worden hieronder enkele aandachtspunten genoemd.

- Als de parallellie van twee testversies niet aannemelijk kan worden gemaakt (de meest kritische eigenschap is hier het gelijke correlatiegedrag met andere variabelen), moeten de berekende coëfficiënten in feite als overeenstemming tussen soortgenoten (een vorm van begripsvalidering) worden beschouwd.
- Bij het construeren van een test of schaal wordt er veelal naar gestreefd een zo hoog mogelijke Cronbachs alfa te verkrijgen. Dit gebeurt vaak door items te gebruiken die qua inhoud homogeen zijn. Dit kan leiden tot een zeer specifieke testinhoud, waarbij het gemeten begrip veel smaller is dan het oorspronkelijk bedoelde begrip. Dit hoeft niet altijd een zinvolle test of schaal op te leveren. Als een subgroep van de items in een test onderling sterker samenhangt dan (met) de rest van de items, of zelfs het bestaan van meer van dergelijke subgroepen binnen een schaal, hoeft dit een hoge Cronbachs alfa of een hieraan gerelateerde coëfficiënt niet in de weg te staan. Integendeel: bij correlaties van een redelijk niveau met de andere items, verhogen dergelijke homogene subgroepen van items de waarde van deze schattingen van de betrouwbaarheid. Relatief hoge inter-itemcorrelaties binnen een subgroep van items kunnen ontstaan doordat deze items niet-bedoelde variantie met elkaar delen en niet met de andere items, bijvoorbeeld vanwege een vergelijkbare wijze van formulering of vanwege een gemeenschappelijk woord. Deze niet-bedoelde variantie draagt bij aan hogere schattingen van de betrouwbaarheid, omdat hiermee de variantie van de systematische verschillen tussen respondenten groter is. Niet-bedoelde vernauwing van het begrip kan worden vermeden door al tijdens de ontwikkelingsfase van de test te toetsen op ééndimensionaliteit, bijvoorbeeld met behulp van structurele vergelijkingsmodellen (met behulp van bijvoorbeeld de programma's *AMOS*, *Mplus* of *LISREL*) en op basis daarvan

maatregelen te nemen in het licht van de theoretische uitgangspunten. Deze laatste toevoeging is van belang omdat de testauteur expliciet kan aangeven of hij een smal of een meerdimensionaal begrip wil meten. De zinvolheid hiervan ('Uitgangspunten van de testconstructie') en de uitkomsten van het onderzoek naar de dimensionaliteit ('Validiteit') als zodanig, worden elders in dit beoordelingssysteem beoordeeld. Hier wordt van de COTAN-beoordelaar gevraagd overwegingen voor een hoge Cronbachs alfa te beoordelen in het licht van deze gegevens en extra op bovengenoemde effecten te letten, wanneer analyses op eendimensionaliteit (nog) niet hebben plaatsgevonden.

- Voor de lengte van het test-hertestinterval zijn er geen strikte richtlijnen te geven. Een zeer kort interval (tot enkele weken) is in het algemeen niet zinvol, vanwege herinneringseffecten, en een zeer lang interval (langer dan een jaar) evenmin, omdat dan externe gebeurtenissen een grote invloed op de persoon en daarmee op de hertestscore kunnen hebben. Hierdoor kan er eigenlijk niet meer van de betrouwbaarheid van het instrument worden gesproken. De genoemde grenzen zijn echter in hoge mate arbitrair en moeten tevens nog in relatie worden gezien tot de leeftijdsgroep en de aard van de test. Ook hoort het beoogde testgebruik bij de keuze van het test-hertestinterval een rol te spelen. Bij een test die is bedoeld voor voorspelling op lange termijn is het bijvoorbeeld zinvol een relatief lang test-hertestinterval te kiezen.
- Als de interbeoordelaarsbetrouwbaarheid wordt gebruikt als schatting voor de betrouwbaarheid van het oordeel van één beoordelaar, moeten de observaties/beoordelingen onafhankelijk hebben plaatsgevonden. Dit moet duidelijk uit de beschrijving van de onderzoeksopzet blijken.

Aanwijzingen bij vraag 5.3.b: "Zijn de steekproeven op basis waarvan de betrouwbaarheidsgegevens zijn berekend, in overeenstemming met het beoogde testgebruik?"

Betrouwbaarheidscoëfficiënten moeten worden berekend voor de groepen waarvoor de test wordt gebruikt. Dit betekent dat deze *per normgroep* moeten worden berekend, aangezien de scores van geteste personen met deze groepen worden vergeleken, en het om de betrouwbaarheid van de meting binnen deze vergelijkingsgroep gaat. Het is daarom onjuist, en zelfs misleidend, om betrouwbaarheidscoëfficiënten te berekenen over het totaal van alle groepen, of, wat ook wel gebeurt, over een selectie van extreme groepen. De hoogte van de gevonden betrouwbaarheidscoëfficiënt is immers mede afhankelijk van de spreiding van de scores en deze zal in de totale groep bijna altijd en bij extreme groepen zeker hoger zijn dan bij de te gebruiken normgroepen. Indien de coëfficiënten niet per normgroep zijn berekend, moet de beoordeling 'onvoldoende' worden gegeven.

Aanwijzingen bij vraag 5.3.c: "Maken de gegevens die worden verstrekt een gefundeerd oordeel over de betrouwbaarheid mogelijk?"

Hieronder staan enkele voorbeelden van informatie die beschikbaar moet zijn om de waarde van het betrouwbaarheidsonderzoek te kunnen beoordelen.

- Worden de standaarddeviaties van de scores bij test en hertestgroep vermeld?
- Wordt er bij tests met een tijdslimiet per item vermeld welk percentage van de geteste personen het item heeft beantwoord?
- Zijn de steekproeven waarover de betrouwbaarheidscoëfficiënten zijn berekend voldoende beschreven?
- Wordt er vermeld op hoeveel observatoren of beoordelaars de gerapporteerde betrouwbaarheidscoëfficiënt betrekking heeft?
- Observatoren/beoordelaars zullen meestal voor hun taak worden getraind. Deze training zal de kwaliteit van de beoordelingen en daarmee de hoogte van de interbeoordelaarsbetrouwbaarheid beïnvloeden. De beschrijving van het trainingsprogramma moet zodanig zijn, dat nieuwe gebruikers zich op dezelfde wijze kunnen voorbereiden (opdat de betrouwbaarheid van de beoordelingen generaliseerbaar is). Aannemelijk moet zijn dat nieuwe gebruikers zich eenzelfde mate van geoefendheid eigen kunnen maken. Ook is het van belang dat er wordt vermeld of de gerapporteerde betrouwbaarheidscoëfficiënt betrekking heeft op het oordeel van één observator/beoordelaar, of op het gemiddelde oordeel van meerdere observatoren/beoordelaars.

In het uiterste geval, wanneer er geen enkele beschrijvende informatie bij de gerapporteerde betrouwbaarheidscoëfficiënten wordt gegeven, kan er op basis van afwezigheid van deze gegevens een 'onvoldoende' worden gegeven. Meestal zal er echter wel enige informatie beschikbaar zijn, zodat de kwaliteit van het onderzoek is te beoordelen. Vooral bij grensgevallen ('onvoldoende'/'voldoende' of 'voldoende'/'goed') kan er om redenen van gebrekkige informatie voor de lagere beoordeling worden gekozen.

Vaststelling eindoordeel voor criterium 5 Betrouwbaarheid			
De basisvraag wordt met '3' beoordeeld.	Vraag 5.2 wordt met '3' beoordeeld.	Vraag 5.3 wordt met '3' beoordeeld.	goed
		Vraag 5.3 wordt met '2' beoordeeld.	voldoende
		Vraag 5.3 wordt met '1' beoordeeld.	onvoldoende
	Vraag 5.2 wordt met '2' beoordeeld.	Vraag 5.3 wordt met '3' beoordeeld.	voldoende
		Vraag 5.3 wordt met '2' beoordeeld.	onvoldoende
		Vraag 5.3 wordt met '1' beoordeeld.	onvoldoende
	Vraag 5.2 wordt met '1' beoordeeld.		onvoldoende
De basisvraag wordt met '1' beoordeeld.			onvoldoende
Bij positieve beoordeling van 5.1 levert vraag 5.2 met betrekking tot de hoogte van de betrouwbaarheidsmaat een voorlopig oordeel. Dit voorlopig oordeel kan naar beneden worden bijgesteld naar aanleiding van het antwoord op vraag 5.3 naar de kwaliteit van het uitgevoerde onderzoek.			

6 Begripsvaliditeit

Validiteit is de mate waarin een test aan zijn doel beantwoordt: kan men uit de testscores die conclusies trekken die men op het oog heeft? In de literatuur worden vele soorten validiteit onderscheiden; zo noemen Drenth en Sijtsma (2006, p. 334-340) acht verschillende vormen. De onderscheidingen hebben betrekking op het doel van het validiteitsonderzoek, of op het proces van validering door bepaalde data-analysetechnieken. In de laatste decennia van de vorige eeuw won de opvatting terrein dat verschillende vormen van validiteitsbepaling niet moesten worden beschouwd als verschillende vormen van validiteit, maar als verschillende manieren om informatie over de validiteit te verzamelen, en dat validiteit moest worden gezien als een ondeelbaar begrip (*a unitary concept*, zie *Standards for Educational and Psychological Testing*, 1999). Het is in deze opvatting van belang vooral díe validiteitsinformatie te verzamelen die past bij het doel van de test, bijvoorbeeld beschrijving, predictie of classificatie. Valideren verwijst zo naar de activiteit van het op wetenschappelijke wijze argumenteren om een bepaalde interpretatie van een test te ondersteunen, waarbij niet alle typen evidenties even belangrijk zijn voor het doel (Ter Laak & De Goede, 2003). Met andere woorden: het gaat niet om de eigenschap van een test, maar om een eigenschap van de interpretatie van testscores. Een recentere visie op het validiteitsbegrip komt van Borsboom, Mellenbergh en Van Heerden (2004), die stellen dat validiteit betrekking heeft op de vraag of het attribuut dat men meet, in staat is variatie in de uitkomsten van de meting te veroorzaken. Ook in deze opvatting is een onderscheid in typen validiteit niet aan de orde.

Welke validiteitsbenadering men ook kiest, voor een gestandaardiseerde beoordeling is het noodzakelijk toch enige structuur in het validiteitsconcept aan te brengen. Hiertoe wordt aangesloten bij de klassieke driedeling naar het doel van het validiteitsonderzoek, zoals die onder andere in de *Richtlijnen voor ontwikkeling en gebruik van psychologische tests en studietoetsen* (Evers e.a., 1988) wordt gehanteerd: inhoudsvaliditeit, begripsvaliditeit en criteriumvaliditeit. Van deze drie wordt validiteitsinformatie die betrekking heeft op de relevantie van de inhoud van een test (inhoudsvaliditeit) en op de betekenis van een testscore (begripsvaliditeit) voor alle typen tests van belang geacht, ongeacht het doel van de test. Dit geldt echter niet voor informatie over de voorspellende waarde van testscores (criteriumvaliditeit): voor tests die geen voorspellende pretentie hebben (bijvoorbeeld toetsen voor voortgangscontrole) is dit type informatie niet vereist. Aan de andere kant is het wel zo dat gegevens over de criteriumvaliditeit kunnen worden betrokken bij het oordeel over de begripsvaliditeit (bij vraag 6.2), omdat deze gegevens eveneens een bijdrage kunnen leveren aan de verheldering van dat wat door de test wordt gemeten. In dat geval maken gegevens met betrekking tot de criteriumvaliditeit in feite ook deel uit van het proces van begripsvalidering (zie bijvoorbeeld Anastasi, 1986; Messick, 1988).

Gegevens of argumenten over de inhoudsvaliditeit worden in dit beoordelingssysteem behandeld als onderdeel van het testontwikkelingsproces en zijn daarom al aan de orde gekomen bij het criterium 'Uitgangspunten van de testconstructie'. In dit hoofdstuk komt de begripsvaliditeit aan de orde. Hoofdstuk 7 gaat over de criteriumvaliditeit.

Bij begripsvaliditeit gaat het erom te toetsen of de test inderdaad de eigenschap meet die wordt verondersteld. Meet de test het bedoelde begrip of, gedeeltelijk of voornamelijk, iets anders? Veelgebruikte methoden of technieken voor het aantonen van de begripsvaliditeit zijn: factoranalyse voor het aantonen van de één-dimensionaliteit, het vergelijken van de gemiddelde scores van groepen waarvan men mag verwachten dat ze verschillen zullen vertonen en het berekenen van correlaties met tests die hetzelfde zouden moeten meten (zogenoemde soortgenoten). Dit zijn in principe vrij eenvoudig uit te voeren onderzoeken die een eerste aanwijzing kunnen opleveren voor de begripsvaliditeit, maar die elk op zich nog géén aanleiding geven tot een 'voldoende' beoordeling. Slechts de cumulatie van dergelijke aanwijzingen, of meer uitgebreid structuur- of zogenoemd *multi-trait-multi-method*-onderzoek (Campbell & Fiske, 1959), kan leiden tot de beoordeling 'voldoende' of 'goed'.

Vragen voor criterium 6 Begripsvaliditeit		onv.	vold.	goed
Basisvraag 6.1	Worden er gegevens over de begripsvaliditeit verstrekt? Bij negatieve beoordeling (1) van deze vraag kan men de rest van de vragen van dit criterium overslaan en doorgaan met criterium 7.	1		3
6.2	Maken de resultaten voldoende aannemelijk dat het begrip zoals bedoeld, wordt gemeten (of: maken de resultaten voldoende duidelijk wat wordt gemeten) op basis van gegevens over:			
	a. de dimensionaliteit van de scores?	1	2	3
	b. de psychometrische kwaliteit van de items?	1	2	3
	c. de invariantie van de factorstructuur en mogelijke itembias bij verschillende groepen?	1	2	3
	d. de convergente en de discriminante validiteit?	1	2	3
	e. verschillen tussen relevante groepen?	1	2	3
	f. op basis van overige gegevens?	1	2	3
6.3.a	Zijn de procedures op basis waarvan de begripsvaliditeitsgegevens zijn berekend correct?	1	2	3
6.3.b	Komen de steekproeven die in het begripsvalideringsonderzoek zijn gebruikt, overeen met groepen waarvoor de test is bedoeld?	1	2	3
6.3.c	Wat is de kwaliteit van de andere maten die in het begripsvalideringsonderzoek zijn gebruikt?	1	2	3
6.3.d	Is de kwaliteit van het onderzoek, zoals beoordeeld in de vragen 6.3.a tot en met 6.3.c, zodanig dat de beoordeling van de begripsvaliditeit, zoals gegeven in vraag 6.2, kan worden bevestigd?	1	2	3

Aanwijzingen bij basisvraag 6.1: "Worden er gegevens over de begripsvaliditeit verstrekt?"
Het gaat hier om onderzoek van de interne of de externe structuur. De interne structuur kan worden onderzocht door associatiematen te bepalen tussen (groepen) items, subtests en tussen subtests en de test als geheel. De externe structuur wordt gewoonlijk onderzocht door relaties met andere tests te bepalen (convergente en discriminante validiteit) en verschillen tussen relevante groepen te berekenen.

Aanwijzingen bij vraag 6.2: "Maken de resultaten voldoende aannemelijk dat het begrip zoals bedoeld, wordt gemeten (of: maken de resultaten voldoende duidelijk wat wordt gemeten)?"
Bij begripsvalidering gaat het vooral om de cumulatie van onderzoeksresultaten: begripsvalidering is nooit af. Het spreekt misschien voor zich, maar toch wordt er hier benadrukt dat het feit dát onderzoek naar de begripsvaliditeit is verricht, niet zonder meer tot een score '3' of '2' bij deze vraag leidt. Bij de beoordeling speelt uitsluitend de kwaliteit van de uitkomsten in het licht van de theoretische uitgangspunten een rol – uiteraard naast de kwaliteit van de gebruikte procedures en het onderzoeksdesign, maar zie daarvoor vraag 6.3. Voor de beoordeling van de begripsvaliditeit zijn de volgende zes typen onderzoeksgegevens relevant:

- gegevens over de dimensionaliteit van scores;
- gegevens over de psychometrische kwaliteit van de items;
- gegevens over de invariantie van de factorstructuur en mogelijke itembias bij verschillende groepen;
- gegevens over de convergente en de discriminante validiteit;
- gegevens over verschillen tussen relevante groepen;
- overige gegevens.

Hieronder volgt een toelichting over elk van deze onderzoeksgegevens.

Gegevens over de dimensionaliteit van scores
Hierbij is zowel de dimensionaliteit op test- als op subtestniveau van belang. De onderzoeksgegevens zouden antwoord moeten geven op de volgende vragen:

- Wanneer er op basis van theoretische overwegingen verschillende subbegrippen worden verondersteld, manifesteren deze zich dan ook als onafhankelijke factoren?
- Blijken de scores van de test (of indien van toepassing: op subtestniveau) unidimensioneel?
- Hoe hoog is de correlatie tussen subtests: worden er wel voldoende onderscheiden begrippen gemeten?

Gegevens over de psychometrische kwaliteit van de items
De kwaliteit van de items kan op diverse manieren worden beoordeeld. Het is gebruikelijk om de gemiddelden van de itemscores per groep te beschouwen en eveneens gegevens over de samenhang tussen items en (sub)test(s) te vermelden. Bij tests die gebaseerd zijn op de klassieke testtheorie gaat het dan om de correlaties van een item met de totaalscore op de overige items in dezelfde (sub) test, ook wel bekend als de item-restcorrelatie (in *SPSS* is dit de *corrected item-total correlation*), bij tests gebaseerd op de item-responstheorie om de fit van items binnen het gekozen model. De volgende gegevens moeten, deels afhankelijk van het gehanteerde model, worden verstrekt.

- ***Item-restcorrelaties***
  De hoogte van de correlatie geeft aan in hoeverre het betreffende item hetzelfde meet als de andere items, maar deze interpretatie is niet zonder risico. De reden hiervoor is dat moet zijn vastgesteld, bijvoorbeeld door middel van factoranalyse, of de items inderdaad alle hoog op dezelfde factor laden. Ook als de test inhoudelijk heterogeen is, kan een item nog steeds hoog correleren met de totaalscore op de andere items, maar omdat deze een 'mengsel' van verschillende eigenschappen representeren, is de interpretatie van de correlatie onduidelijk of discutabel. Een andere interpretatie van de item-restcorrelatie is die van discriminerend vermogen. Stel dat de totaalscore op de andere items een schaal voorstelt, dan betekent een hoge item-restcorrelatie dat personen met een lage itemscore veelal een lage schaalscore hebben en personen met een hoge itemscore een hoge schaalscore. Het item is dus goed in staat dit onderscheid te maken. Voor het beoordelen van r_{it}-waarden bij tests waar het om een hoge mate van interne consistentie gaat (zie voor uitzonderingen de aanwijzingen bij vraag 5.2.b), kan men de richtlijnen aanhouden (gebaseerd op Veldhuijzen, Goldebeld & Sanders, 1993) in de volgende tabel:

r_{it}-waarde	Beoordeling
0.30 en hoger	goed
0.20 – 0.29	voldoende
0.19 en lager	onvoldoende

Bedenk wel dat het in bovenstaande tabel om r_{it}-waarden (item-totaalcorrelaties) gaat; de meer gebruikelijke r_{rt}-waarden (item-rest correlaties) kunnen vooral bij korte tests wat lager uitvallen. Ook blijkt de lengte van een test de r_{it}-waarde te beïnvloeden: hoe langer de test, des te lager in het algemeen de gemiddelde r_{it}-waarde.

- ***Itemparameters volgens een item-responsmodel***
  Via item-responsmodellen worden vaak schattingen van item-moeilijkheden en itemdiscriminaties verkregen op schalen die sterk verschillend zijn van die van de meer vertrouwde item-gemiddelden en item-restcorrelaties. Als deze voor de item-responstheorie typische itemindices worden gerapporteerd, verdient het aanbeveling om daarnaast ook de bekendere, klassieke itemindices zoals itemgemiddelden en item-rest-correlaties te rapporteren. De nauwkeurigheid van de item-parameterschattingen kan in voorkomende gevallen ook beoordeeld worden door te kijken naar de relatie tussen de standaardfout van de moeilijkheidsparameter $se(b_i)$ en de standaarddeviatie van de vaardigheidsverdeling van de kalibratiepopulatie $sd(\theta)$. Hierbij zou moeten gelden dat: $se(b_i) < c*sd(\theta)$, waarbij c een constante is. Voor het beoordelen van de standaardfout van $se(b_i)$ gelden de richtlijnen in onderstaande tabel:

c	Beoordeling
$c \geq 0,5$	groot (= 'onvoldoende')
$0,3 \leq c \leq 0,4$	matig (= 'voldoende')
$c \leq 0.2$	klein (= 'goed')

- ***Omvang van de steekproef***
  De steekproef moet voldoende groot zijn om te voorkomen dat de itemparameters onnauwkeurig geschat worden. Er zijn hiervoor om twee redenen geen eenduidige richtlijnen op te stellen. Ten eerste is de minimale benodigde grootte van de steekproef afhankelijk van het gekozen item-responsmodel en ten tweede worden er in de literatuur weinig aanwijzingen gegeven met betrekking tot de gewenste grootte van de steekproef. Het is vaak een kwestie van 'ervaring'. In de literatuur worden nauwelijks richtlijnen genoemd voor de steekproefgrootte die nodig is bij de logistische modellen voor dichotome items. Op basis van Parshall, Davey, Spray en Kalohn (1998) zijn de richtlijnen in onderstaande tabel opgesteld:

Model	N
3-parameter	N > 700
2-parameter	N > 400
1-parameter	N > 200

- ***Passing statistisch model***
  Alle statistische methoden zijn gebaseerd op vooronderstellingen ('assumpties') over verdelingen van variabelen (bijvoorbeeld normaal) en relaties tussen variabelen (bijvoorbeeld lineair). Dit geldt bijvoorbeeld voor factormodellen en item-responsmodellen, maar ook voor de bekende product-momentcorrelatie. Allerlei uitspraken over de kwaliteit van tests die zijn gebaseerd op statistische berekeningen zijn alleen te vertrouwen als er voor de betreffende toepassing is aangetoond dat aan deze vooronderstellingen is voldaan. Het is ondoenlijk om hier aan te geven wat dit voor elke techniek inhoudt, maar van de testconstructeur mag worden verwacht dat hij de noodzakelijke informatie over modelpassing *(goodness-of-fit)* in de handleiding van de test of vragenlijst rapporteert.

Gegevens over de invariantie van de factorstructuur en mogelijke itembias bij verschillende groepen
Dit onderzoek kan plaatsvinden op basis van modellen en procedures die passen binnen de klassieke testtheorie of de item-responstheorie. Als er verschillen in factorstructuur zijn vastgesteld of als er itembias is aangetoond, moeten de consequenties worden aangegeven (bijvoorbeeld een schatting van het effect op de totale testscore). Bijkomend voordeel van onderzoek naar itembias is dat het informatie oplevert over de mogelijke meerdimensionaliteit van het gemeten begrip.

Gegevens over de convergente en de discriminante validiteit
Beide typen gegevens kunnen in één onderzoek worden verkregen via de multi-trait-multi-method-benadering. Gegevens over de convergente validiteit kunnen ook worden verkregen via correlatie met 'soortgenoot'-tests. Gegevens over de discriminante validiteit zijn van belang om te kunnen uitsluiten dat niet 'per ongeluk' een ander dan het bedoelde begrip wordt gemeten (meet men wel arbeidstevredenheid en niet bijvoorbeeld negatieve affectiviteit; meet men wel rekenvaardigheid en niet voor een belangrijk deel taalvaardigheid?).

Gegevens over verschillen tussen relevante groepen
Afhankelijk van de meetpretentie van een test en de kenmerken van bepaalde groepen kan men verschillen tussen deze groepen verwachten. Zo is te verwachten dat leerlingen in basisgroep 8 hoger zullen scoren op een toets voor rekenvaardigheid dan leerlingen in groep 6. Evenzo is het te verwachten dat kinderen die zijn gediagnosticeerd als ADHD hoger zullen scoren op een test voor hyperactiviteit dan 'normale' kinderen. Dergelijk groepsvergelijkend onderzoek is belangrijk, omdat het een eerste aanwijzing kan geven dat de test groepen kan onderscheiden zoals is bedoeld. Als er tegen de verwachting in geen verschillen zouden blijken te zijn, zou het bovendien zeer onwaarschijnlijk zijn dat de test het bedoelde begrip meet. Het omgekeerde is echter niet waar: als er verschillen tussen relevante groepen blijken te zijn, hoeft dit nog niet te betekenen dat de test nu werkelijk meet wat wordt bedoeld (de rekenvaardigheidstoets kan nog steeds taalvaardigheid meten en de test voor hyperactiviteit een of andere vorm van sociaal onwenselijk gedrag).

Overige gegevens
Dit kunnen bijvoorbeeld gegevens zijn over de criteriumvaliditeit die tevens informatie opleveren over de begripsvaliditeit.

De vraag naar de totaalscore kan met een score '2' worden beoordeeld, als er resultaten op ten minste twee van de bovengenoemde typen onderzoek worden gerapporteerd, als deze uitkomsten in het algemeen de gewenste structuur ondersteunen, en als deze op zowel de interne als de externe structuur betrekking hebben. De score '3' kan worden toegekend als resultaten op ten minste drie van de bovengenoemde typen onderzoek worden gerapporteerd, deze uitkomsten unaniem de gewenste structuur ondersteunen, en op zowel de interne als de externe structuur betrekking hebben.

Aanwijzingen bij vraag 6.3.a: "Zijn de procedures op basis waarvan de begripsvaliditeitsgegevens zijn berekend correct?"
De opzet van het onderzoek en de gebruikte analysetechnieken moeten voldoende duidelijk zijn beschreven. Onvoldoende informatie kan namelijk tot het oordeel '2' of zelfs '1' op deze vraag leiden. Gelet op de veelsoortigheid van dit type onderzoek kunnen hier verder nauwelijks algemene aanwijzingen worden gegeven, behalve dat de grootte van de onderzoeksgroep van belang is bij de waardering van de onderzoeksresultaten. Enkele specifieke aandachtspunten hierbij zijn:

- Wanneer de samenhang tussen items en (sub)test(s) wordt onderzocht, moet worden gecorrigeerd voor het aandeel van het item zelf in de (sub)testscore, omdat de berekende waarden

anders geflatteerd uitvallen (dat wil zeggen dat er zogenoemde item-restcorrelaties in plaats van item-totaalcorrelaties moeten worden vermeld).

- Bij onderzoek naar de convergente validiteit moet worden gewaarschuwd tegen de interpretatie van onderzoeksresultaten zonder specifieke verwachtingen vooraf. Dergelijk onderzoek krijgt al gauw het karakter van 'vissen': post hoc zal men altijd wel een aantal interpreteerbare verbanden vinden, wanneer men de test correleert met de scores op een groot aantal (toevallige beschikbaar zijnde) andere tests. Hierbij is het aannemelijk dat enkele van de significante correlaties op toeval zullen berusten. Deze kans op toevalscorrelaties wordt groter naarmate de te valideren test uit meer subtests of schalen bestaat.

Aanwijzingen bij vraag 6.3.b: "Komen de steekproeven die in het begripsvalideringsonderzoek zijn gebruikt overeen met groepen waarvoor de test is bedoeld?"

Het valideringsonderzoek moet betrekking hebben op de populatie waarvoor de test wordt gebruikt. Hierbij is vooral de variantie van de testscores in de onderzoeksgroep van belang. Omdat validiteitscoëfficiënten in het algemeen lager zullen uitvallen bij afnemende variantie, zal een validiteitsonderzoek dat is uitgevoerd op een heterogenere groep dan de groep waarvoor de test uiteindelijk is bedoeld, geflatteerde resultaten laten zien. Zo is het onjuist om een test die is bedoeld voor therapieselectie bij mensen die zich daarvoor hebben aangemeld, te valideren op een doorsnee van de 'normale' bevolking. Om een en ander te kunnen beoordelen, moet de onderzoeksgroep zijn beschreven met behulp van mogelijk relevante psychologische of demografische variabelen.

Wanneer de test volgens de handleiding is bedoeld voor gebruik in verschillende situaties en/of voor verschillende groepen, dan moet er onderzoek zijn verricht in meerdere van deze situaties en/of bij meerdere groepen.

Aanwijzingen bij vraag 6.3.c: "Wat is de kwaliteit van de andere maten die in het begripsvalideringsonderzoek zijn gebruikt?"

Van de gebruikte maten moet de betrouwbaarheid bekend zijn. Het spreekt bijna voor zich dat validering aan maten met een lage betrouwbaarheid (lager dan .60) minder zinvol is, omdat de resultaten in dat geval niet goed te interpreteren zijn. Validering aan soortgenoten heeft bovendien alleen zin als daarvoor instrumenten worden gebruikt waarvan de validiteit zelf voldoende is onderzocht.

Aanwijzingen bij vraag 6.3.d: "Is de kwaliteit van het onderzoek, zoals beoordeeld in de vragen 6.3.a tot en met 6.3.c, zodanig dat de beoordeling van de begripsvaliditeit, zoals gegeven in vraag 6.2, kan worden bevestigd?"

Negatieve beantwoording ('1') op een van de vragen 6.3.a tot en met 6.3.c leidt tot een score '1' bij vraag 6.3.d. Dit betekent dat het oordeel over de resultaten van het begripsvalideringsonderzoek zoals gegeven in vraag 6.2 naar beneden moet worden bijgesteld. Ook meerdere '2'-oordelen op de vragen 6.3.a tot en met 6.3.c kunnen betekenen dat het onderzoek zo veel manco's vertoont dat vraag 6.3.d negatief wordt beantwoord en op grond hiervan het oordeel van vraag 6.2 naar beneden wordt bijgesteld.

Vaststelling eindoordeel voor criterium 6 Begripsvaliditeit			
De basisvraag wordt met '3' beoordeeld.	Vraag 6.2 wordt met '3' beoordeeld.	Vraag 6.3.d wordt met '3' beoordeeld.	goed
		Vraag 6.3.d wordt met '2' beoordeeld.	voldoende
		Vraag 6.3.d wordt met '1' beoordeeld.	onvoldoende
	Vraag 6.2 wordt met '2' beoordeeld.	Vraag 6.3.d wordt met '3' beoordeeld.	voldoende
		Vraag 6.3.d wordt met '2' beoordeeld.	onvoldoende
		Vraag 6.3.d wordt met '1' beoordeeld.	onvoldoende
	Vraag 6.2 wordt met '1' beoordeeld.		onvoldoende
De basisvraag wordt met '1' beoordeeld.			onvoldoende

7 Criteriumvaliditeit

Bij criteriumvaliditeit onderzoekt men in hoeverre de testscore een goede voorspeller is van niet-testgedrag (retrospectief, gelijktijdig of predictief). Het is van belang dat er op basis van de pretentie van de test verwachtingen worden gespecificeerd over het type criteria waarmee relaties worden verondersteld. Dit is vooral van belang als een test uit verscheidene subtests of schalen bestaat; zie wat hierover bij de aanwijzing bij vraag 6.3.a over 'vissen' wordt gezegd. Overigens hoeft voor een 'voldoende' of 'goed' beoordeling niet de validiteit van alle subtests of schalen worden aangetoond, omdat een enkele zeer valide schaal de test al tot een waardevol instrument kan maken.
In principe is onderzoek naar de criteriumvaliditeit voor alle typen tests vereist, omdat het uiteindelijke doel van tests is om voorspellingen te doen. Wanneer er echter in de handleiding van een test expliciet wordt aangegeven dat een test geen voorspellende pretenties heeft, en dit ook aannemelijk is, zoals bij tests die worden gebruikt voor voortgangscontrole, dan kan criteriumvaliditeit 'niet van toepassing' worden verklaard. Bij de beoordeling wordt in dergelijke gevallen de volgende voetnoot opgenomen: "Deze test is volgens de auteur(s)/uitgever niet bedoeld voor voorspellend gebruik. Criteriumvaliditeit is daarom niet van toepassing. Wanneer deze test echter wordt ingezet in situaties waarin voorspelling wel aan de orde is, geldt het oordeel 'onvoldoende', omdat er geen onderzoek naar de criteriumvaliditeit is verricht".

Vragen voor criterium 7 Criteriumvaliditeit				
		onv.	**vold.**	**goed**
Basisvraag 7.1	Worden er gegevens verstrekt over het verband test-criterium? Bij negatieve beoordeling (1) van deze vraag, kan men de rest van de vragen overslaan.	1		3
7.2	Zijn de resultaten voldoende, gelet op het beoogde type beslissingen dat met de test moet worden genomen?	1	2	3
7.3.a	Zijn de procedures op grond waarvan de criteriumvaliditeitsgegevens zijn berekend correct?	1	2	3
7.3.b	Zijn de steekproeven op grond waarvan de criteriumvaliditeitsgegevens zijn berekend in overeenstemming met het beoogde testgebruik?	1	2	3
7.3.c	Wat is de kwaliteit van de criteriummaten?	1	2	3
7.3.d	Is de kwaliteit van het onderzoek, zoals beoordeeld in de vragen 7.3.a tot en met 7.3.c, zodanig dat de beoordeling van de criteriumvaliditeit, zoals gegeven in vraag 7.2, kan worden bevestigd?	1	2	3

Aanwijzingen bij basisvraag 7.1: "Worden er gegevens verstrekt over het verband test-criterium?"

Hierbij kan men bijvoorbeeld denken aan:

- De correlatie van de scores op een intelligentietest met schoolresultaten.
- De voorspellende waarde van een test die wordt gebruikt voor de selectie van werknemers voor het latere functioneren (bijvoorbeeld validiteitscoëfficiënten of succesratio's).
- Bij het stellen van een klinische diagnose: gegevens over de sensitiviteit (de verhouding tussen het aantal door de test geïdentificeerde personen met een stoornis en het werkelijk aantal personen met die stoornis) en specificiteit (de verhouding tussen het aantal door de test geïdentificeerde personen zonder stoornis en het werkelijk aantal personen zonder die stoornis) en/of gegevens over de ROC-curve.

Het is niet zo dat deze gegevens voor elke nieuwe situatie of elke nieuwe test opnieuw moeten worden verzameld. Er mag gebruik worden gemaakt van het principe van validiteitsgeneralisatie. In dat geval moeten bij vraag 7.2 de hoogte van de validiteitscoëfficienten van het (de) oorspronkelijke onderzoek(en) en bij vraag 7.3 de kwaliteit van het (de) oorspronkelijke onderzoek(en) worden beoordeeld.

Aanwijzingen bij vraag 7.2: "Zijn de resultaten voldoende, gelet op het beoogde type beslissingen dat met de test moet worden genomen?"

Of een of meer validiteitsstudies voldoende zijn, hangt van een aantal zaken af. Van belang zijn onder andere het doel van de test, de hoogte van validiteitscoëfficiënten of de waarden van ROC-curves, de betrouwbaarheidsintervallen van deze maten, de winst die de test oplevert ten aanzien van al aanwezige informatie, de selectieratio en de utiliteit. Verder kan een test in verschillende situaties of bij verschillende groepen andere coëfficiënten opleveren of gedeelten van een criterium goed voorspellen. Zo wordt in selectiesituaties een validiteitscoëfficiënt van .40 als goed gezien (zie bijvoorbeeld Schmidt & Hunter, 1998), terwijl in opleidingssituaties met gemak hogere coëfficiënten worden behaald. Swets (1988) geeft een overzicht van waarden van ROC-curves die op verschillende gebieden worden gevonden. Bij bepaalde vormen van medische diagnose blijken deze te tussen .81 en .97 te liggen, bij leugendetectie tussen .70 en .95 en bij het voorspellen van schoolresultaten (slagen/zakken) met behulp van capaciteitentests werden waarden tussen .71 en .94 gevonden. Naarmate de auteur explicieter is over het doel van de test, kan de beoordelaar beter uitmaken of de test daaraan een zinvolle bijdrage levert. Het wordt daarbij aan de expertise van de beoordelaar overgelaten een oordeel te geven over de hoogte van de gevonden resultaten. Wanneer daar aanleiding toe is (bijvoorbeeld als subgroepen verschillen in gemiddelde scores of als uit onderzoek bij vergelijkbare tests is gebleken dat de voorspellende waarde per subgroep kan verschillen), moet de auteur onderzoek uitvoeren naar mogelijke predictiebias voor de betreffende groepen.

Aanwijzingen bij vraag 7.3.a: "Zijn de procedures op grond waarvan de criteriumvaliditeitsgegevens zijn berekend correct?"

Enkele aspecten waarop de beoordelaar moet letten:

- Is er sprake van criteriumcontaminatie? Dus: zijn de scores op predictor en criterium onafhankelijk van elkaar tot stand gekomen? Dit is bijvoorbeeld niet het geval als degene die de criteriumbeoordeling vaststelt, kennis heeft van de testresultaten.
- Is de tijd tussen testafname en criteriummeting in overeenstemming met het beoogde testgebruik? Bij valideringsonderzoek neemt men nogal eens zijn toevlucht tot gelijktijdigheidsonderzoek (concurrent validity), omdat follow-upgegevens niet beschikbaar zijn, of omdat men hierop niet wil wachten. In de selectiesituatie noemt men dit de 'werknemersmethode' (Guion, 1991). In principe zijn op dergelijke wijze verkregen validiteitsgegevens minder geschikt, omdat op zijn minst onduidelijk is of zij een adequate schatting geven van de werkelijke validiteit van de test. De oorzaak hiervoor is dat de samenstelling (selectie, uitval), kennis (ervaring), motivatie en het invulgedrag (faking) van de onderzoeksgroepen tijdens een predictief en een gelijktijdig onderzoek kunnen verschillen. Overigens lijken de effecten van deze factoren elkaar in de selectiesituatie min of meer op te heffen, waardoor er in meta-analyses nauwelijks verschil in de hoogte van validiteitscoëfficiënten uit predictief en gelijktijdig onderzoek wordt gevonden (Cook, 2004). Niettemin blijft voorzichtigheid geboden bij de interpretatie van de uitkomsten van individuele onderzoeken.
- Heeft het valideringsonderzoek onder dezelfde testcondities plaatsgevonden als waarin de test wordt gebruikt?
- Wanneer er voor attenuatie of voor beperking in spreidingsbreedte is gecorrigeerd, worden dan ook de ongecorrigeerde coëfficiënten en andere mogelijk relevante gegevens vermeld? In bepaalde gevallen leveren deze correcties namelijk een onder- of overschatting op van de validiteitscoëfficiënt. Nadat de test het ontwikkelstadium is gepasseerd, moet in geen geval de correctie voor attenuatie worden toegepast voor onbetrouwbaarheid in de test zelf. In de praktijk voorspelt men immers met behulp van de testscore, niet met behulp van de ware score.
- Is er een kruisvalideringsonderzoek uitgevoerd? Dit is vooral van belang bij een beperkte groepsgrootte en bij bepaalde analysetechnieken die in sterkere mate kapitaliseren op het toeval; dit zijn vooral multivariate methoden als (logistische) regressie-analyse en discriminantanalyse.
- Wordt de omvang van de steekproef vermeld? Hoe kleiner de steekproeven, des te groter de betrouwbaarheidsintervallen van regressiegewichten en validiteitscoëfficiënten.
- Als er validiteitsgeneralisatie wordt gebruikt, zal de testauteur aannemelijk moeten maken dat de situaties of de tests waarover generalisatie wordt geclaimd, overeenkomen. Voor de vergelijkbaarheid van tests zal de auteur moeten aantonen dat hetzelfde begrip wordt gemeten bij een minstens gelijke betrouwbaarheid. Dit kan vooral van belang zijn voor een

Nederlandse vertaling van een buitenlandse test, waarvoor al veel buitenlandse onderzoeksgegevens beschikbaar zijn. Als de testauteur deze gegevens wil gebruiken om de validiteit van de Nederlandse versie te onderbouwen, dan zal in eerste instantie, via bijvoorbeeld confirmatieve factoranalyse, de equivalentie van beide versies moeten worden aangetoond. Bij een positieve uitkomst kunnen de buitenlandse validiteitsgegevens worden meegenomen in de beoordeling. Dit kan echter alleen als deze gegevens in de Nederlandse handleiding op afdoende wijze worden samengevat.

Aanwijzingen bij vraag 7.3.b: "Zijn de steekproeven op grond waarvan de criteriumvaliditeitsgegevens zijn berekend, in overeenstemming met het beoogde testgebruik?"
Het valideringsonderzoek moet betrekking hebben op de populatie waarvoor de test wordt gebruikt. Vooral de variantie van de testscore in de onderzoeksgroep is daarbij van belang. Het is bekend dat validiteitscoëfficiënten drastisch afnemen wanneer men van een heterogene groep naar een homogene groep generaliseert. Zo is het onjuist om een test die is bedoeld voor therapieselectie bij mensen die zich daarvoor hebben aangemeld (homogene groep) te valideren op een doorsnee van de 'normale' bevolking (heterogene groep), omdat dit geflatteerde resultaten te zien zal geven. Om een en ander te kunnen beoordelen, moet de steekproef zijn beschreven met behulp van mogelijk relevante psychologische of demografische variabelen.

Aanwijzingen bij vraag 7.3.c: "Wat is de kwaliteit van de criteriummaten?"
Soms ligt de keuze van een criterium voor de hand en is het makkelijk beschikbaar (slagen/zakken, een rapportcijfer). In andere gevallen moeten er criteriummaten apart worden geconstrueerd en verzameld. In beide gevallen moet het criterium zo volledig mogelijk worden beschreven en moet er zijn aangegeven welke relevante gedragsaspecten wel en niet in de criteriummaat zijn opgenomen. Hierbij is zowel construct underrepresentation (niet alle relevante onderdelen van het criterium worden gemeten) als construct overrepresentation (er worden gedragsaspecten gemeten die niet tot het criterium kunnen worden gerekend) van belang. Zo mogelijk moet de betrouwbaarheid van de criteriummaat worden vermeld. Dit geldt vooral voor samengestelde criteria. Als de onderlinge relaties van de afzonderlijke elementen van het criterium laag zijn, kunnen er beter afzonderlijke validiteitscoëfficiënten voor elk van de elementen worden gegeven.

Aanwijzingen bij vraag 7.3.d: "Is de kwaliteit van het onderzoek, zoals beoordeeld in de vragen 7.3.a tot en met 7.3.c, zodanig dat de beoordeling van de criteriumvaliditeit, zoals gegeven in vraag 7.2, kan worden bevestigd?"
Negatieve beantwoording op een van de vragen 7.3.a tot en met 7.3.c leidt tot een score '1' bij vraag 7.3.d. Dit betekent dat het oordeel over de resultaten van het begripsvalideringsonderzoek zoals gegeven in vraag 7.2 naar beneden moet worden bijgesteld. Ook meerdere '2'-oordelen op de vragen 7.3.a tot en met 7.3.c kunnen betekenen dat het onderzoek zo veel manco's vertoont, dat vraag 7.3.d negatief wordt beantwoord en op grond hiervan het oordeel van vraag 7.2 naar beneden wordt bijgesteld.

Vaststelling eindoordeel voor criterium 7 Criteriumvaliditeit			
De basisvraag wordt met '3' beoordeeld.	Vraag 7.2 wordt met '3' beoordeeld.	Vraag 7.3.d wordt met '3' beoordeeld.	goed
		Vraag 7.3.d wordt met '2' beoordeeld.	voldoende
		Vraag 7.3.d wordt met '1' beoordeeld.	onvoldoende
	Vraag 7.2 wordt met '2' beoordeeld.	Vraag 7.3.d wordt met '3' beoordeeld.	voldoende
		Vraag 7.3.d wordt met '2' beoordeeld.	onvoldoende
		Vraag 7.3.d wordt met '1' beoordeeld.	onvoldoende
	Vraag 7.2 wordt met '1' beoordeeld.		onvoldoende
De basisvraag wordt met '1' beoordeeld.			onvoldoende

Literatuur

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory.* Monterey, CA: Brooks/Cole.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology, 37,* 1-15.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington, D.C.: American Council on Education.

Barelds, D. P. H., Luteijn, F., Dijk, H. van, & Starren, H. (2007). *NPV-2. Nederlandse Persoonlijkheidsvragenlijst-2.* Amsterdam: Harcourt Test Publishers.

Bartram, D. (2005). Computer-Based Testing and the Internet. In A. Evers, N. Anderson & O. Voskuijl (Eds.), *The Blackwell handbook of personnel selection* (pp. 399-418). Oxford, UK: Blackwell.

Bechger, T., Hemker, B., & Maris, G. (2009). *Over het gebruik van continue normering.* Arnhem: Cito.

Berge, J. M. F. ten, & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika,* 69, 613-625.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56,* 137-172.

Borsboom, D., Mellenbergh, G. J., & Heerden, J. van (2004). The concept of validity. *Psychological Review, 111,* 1061-1071.

Campbell, D. P. (1971). *Handbook for the Strong Vocational Interest Blank.* Stanford, CA: Stanford University Press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Cascio, W. F., & Aguinis, H. (2005). *Applied psychology in human resource management* (6th ed.). Upper Saddle River, NJ: Pearson Prentice-Hall.

Cizek G. J. (1996). Standard-setting guidelines. *Educational Measurement: Issues and Practice, 15,* 13-21.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

Cohen, J. (1969). Rc: a profile similarity coefficient invariant over variable reflection. *Psychological Bulletin, 71,* 281-284.

Cook, M. (2004). Personnel selection. *Adding value through people (4th edition).* Chicester, UK: John Wiley & Sons.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Drenth, P. J. D., & Sijtsma, K. (1990). *Testtheorie. Inleiding in de theorie van de psychologische test en zijn toepassingen.* Houten/Antwerpen: Bohn Stafleu Van Loghum.

Drenth, P. J. D., & Sijtsma, K. (2006). *Testtheorie. Inleiding in de theorie van de psychologische test en zijn toepassingen* (4e herziene druk). Houten: Bohn Stafleu van Loghum.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Erkens, T. T. M. G., & Moelands, H. A. (1992). *Toetsen met open vragen: een handleiding voor het construeren van toetsen met open vragen.* Arnhem: Cito.

Evers, A. (1979). *Amsterdamse Beroepen Interessen Vragenlijst. ABIV. Handleiding.* Lisse: Swets & Zeitlinger.

Evers, A. (1992). *Amsterdamse Beroepen Interessen Vragenlijst. ABIV92. Handleiding.* Lisse: Swets Test Services.

Evers, A., Caminada, H., Koning, R., Laak, J. ter, Maesen de Sombreff, P. van der, & Starren, J. (1988). *Richtlijnen voor ontwikkeling en gebruik van psychologische tests en studietoetsen.* Amsterdam: NIP.

Evers, A., & Resing, W. C. M. (2007). Het drijfzand van didactische leeftijdsequivalenten. *De Psycholoog, 42,* 466-472.

Evers, A., Vliet-Mulder, J.C. van, & Groot, C. (2000). *Documentatie van Tests en Testresearch in Nederland, dl. 1 en 2.* Amsterdam/Assen: NIP/Van Gorcum.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. Biometrics, 27, 857-871.

Green, S. A., & Yang, Y. (2009). Commentary on coefficient alpha: a cautionary tale. *Psychometrika, 74*, 121-135.

Guion, R. M. (1991). Personnel assessment, selection, and placement. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of Industrial & Organizational Psychology* (Vol. 2, 2nd ed., pp. 327-397). Palo Alto, CA: Consulting Psychologists Press.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10,* 255-282.

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement, 24,* 355-366.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.). (2005). *Adapting educational and psychological tests for cross-cultural assessment.* Mahwah, NJ: Lawrence Erlbaum Associates.

Heuvelmans, A. P. J. M., & Sanders, P. F. (1993). Beoordelaarsovereenstemming. In T. J. H. M. Eggen & P. F. Sanders (Red.), *Psychometrie in de praktijk* (pp. 443-470). Arnhem: CITO.

Hofstee, W. K. B., Campbell, W. H., Eppink, A., Evers, A., Joe, R. C., Koppel, J. M. H. van de, Zweers, H., Choenni, C. E. S., & Zwan, T. J. van der (1990). *Toepasbaarheid van psychologische tests bij allochtonen. LBR-reeks nr.11.* Utrecht: LBR.

International Test Commission (2000). *ITC Test Adaptation Guidelines.* www.intestcom.org.

Kersting, M. (2006). *"DIN SCREEN". Leitfaden zur Kontrolle und Optimierung der Qualität vor Verfahren und deren Einsatz bei beruflichen Eignungsbeurteilungen.* Lengerich: Pabst Science Publishers.

Keuning, J. (2004). *De ontwikkeling van een beoordelingssysteem voor het beoordlen van "Computer Based Tests".* POK Memorandum 2004-1. Citogroep: Arnhem.

King, W. C., & Miles, E. W. (1995). A quasi-experimental assessment of the effects of computerized non cognitive paper-and-pencil measurements: A test of measurement equivalence. *Journal of Applied Psychology, 80,* 643-651.

Kingsbury, G. G., & Zara, A. R. (1991). A comparison of procedures for content sensitive item selection in computerized adaptive tests. *Applied Measurement in Education, 4,* 241-261.

Laak, J. J. F. ter, & Goede, M. P. M. de (2003). *Psychologische diagnostiek. Inhoudelijke en methodologische grondslagen.* Lisse: Swets & Zeitlinger.

Langenbucher, J. W., Labouvie, E., Martin, C. S., Suanjuan, P. M., Bavly, L., & Kirisci, L. (2004). An application of item response theory analysis to alcohol, cannabis, and cocoaine criteria in DSM-IV. *Journal of Abnormal Psychology, 113,* 72-80.

Laros, J. A., & Tellegen, P. J. (1991). *Construction and validation of the SON-R 51/2-17, the Snijders-Oomen non-verbal intelligence test.* Groningen: Wolters-Noordhoff.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance of educational and occupational tests.* Princeton, NJ: Educational Testing Service.

Lord, F. M., & Novick, M. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Luteijn, F., Starren, H. & Dijk, H. van. (1985). *Nederlandse Persoonlijkheidsvragenlijst. Handleiding (herziene uitgave).* Lisse: Swets & Zeitlinger.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114,* 449-458.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33-45). Hillsdale, NJ: Lawrence Erlbaum.

Moelands, H. A., Noijons, J., & Rem, J. (1992). *Toetsen met gesloten vragen: een handleiding voor het construeren van toetsen met meerkeuze vragen.* Arnhem: Cito.

Mokken, R. J. (1971). *A theory and procedure of scale analysis.* The Hague: Mouton.

Murphy, K. R., & Davidshofer, C. O. (1998). *Psychological testing. Principles and aplications.* Upper Saddle River, NJ: Prentice Hall.

Nederlands Instituut van Psychologen (2004). *Algemene Standaard Testgebruik (AST).* Amsterdam: NIP.

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika, 32,* 1-13.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21,* 173-184.

Resing, W., & Drenth, P. (2007). *Intelligentie. Weten en meten (2e editie).* Amsterdam: Uitgeverij Nieuwezijds.

Reise, S. P., & Havilund, M. G. (2005). Item response theory and the measurement of clinical change. *Journal of Personality Measurement, 84,* 228-238.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262-274.

Schneider, R. J., & Hough, L. M. (1995). Personality and industrial/organizational psychology. In C. L. Cooper & I. T. Robertson (Eds.), *International Review of Industrial and Organizational Psychology, 10,* 75-129.

Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research, 7,* 301-317.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17,* 277-292.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240,* 1285-1293.

Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. *Proceeding of the 27th annual meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

Twenge, J. M. (2000). The age of anxiety? Birth cohort change in anxiety and neuroticism, 1952-1993. *Journal of Personality and Social Psychology, 79,* 1007-1021.

Veldhuijzen, N. H., Goldebeld, P., & Sanders, P. F. (1993). Klassieke testtheorie en generaliseerbaarheidstheorie. In T. J. H. M.Eggen & P. F.Sanders (Red.), *Psychometrie in de praktijk* (pp. 33-82). Arnhem: CITO.

Verstralen, H. H. F. M. (1993). Schalen, normen en cijfers. In T. J. H. M.Eggen & P. F.Sanders (Red.), *Psychometrie in de praktijk* (pp. 471-509). Arnhem: CITO.

Vijver, F. van de, & Hambleton, R. K. (1996). Translating tests: some practical guidelines. *European Psychologist, 1,* 89-99.

Visser, R. S. H., Vliet-Mulder, J. C. van, Evers, A., & Laak, J. ter (1982). *Documentatie van tests en testresearch in Nederland.* Amsterdam: NIP.

Vos, H. J., & Knuver, J. W. M. (2000). Standaarden in onderwijs-evaluatie. In R. J. Bosker (Ed.), *Onderwijskundig lexicon (Editie III), Evalueren in het onderwijs* (pp. 59-76). Alphen aan de Rijn: Samsom.

Wools, S., Sanders, P., & Roelofs, E. (2007). *Beoordelingsinstrument: Kwaliteit van competentie assessment.* Cito: Arnhem.

Zeeuw, J. de (1978). *Algemene psychodiagnostiek II. Testtheorie.* Amsterdam: Swets & Zeitlinger.

Zegers, F. E. (1989). Het meten van overeenstemming. *Nederlands Tijdschrift voor de Psychologie, 44,* 145-156.

Zegers, F. E., & Ten Berge, J. M. F. (1985). A family of association coefficients for metric scales. *Psychometrika, 50,* 17-24.

NIP/COTAN

www.psynip.nl

Postbus 9921
1006 AP Amsterdam
telefoon (020) 410 62 22
cotan@psynip.nl

Redactie **Communicatie NIP**
Vormgeving **Link Design, Amsterdam**
Druk **Heijnis & Schipper, Zaandijk**

© NIP/COTAN, gewijzigde herdruk, mei 2010
Niets uit deze uitgave mag worden vervoeldvoudigd, overgenomen of gekopieerd zonder toestemming van het NIP

ISBN 978 90 6999 015 6