Beoordelingssysteem

voor de

Kwaliteit van Tests

Deel 1

* Ingekort en aangepast voor de opdracht Diagnostiek en Testtheorie

INLEIDING

1. Inhoud van het beoordelingssyteem

Een test wordt beoordeeld op vijf categorieën. In deze inleiding worden de eerste drie categorieën besproken. Elke categorie bestaat uit meerdere vragen, waaronder één of meer basisvragen. Met behulp van de basisvragen wordt vastgesteld of aan bepaalde minimum vereisten is voldaan, zonder welke verdere beoordeling van de betreffende categorie overbodig wordt of niet mogelijk is.

Het eerste deel van de beoordeling van een test leidt tot een waardering op de volgende aspecten:

1. Uitgangspunten van de testconstructie. Deze categorie telt twee vragen, waarvan één basisvraag. Eerst wordt beoordeeld of het gebruiksdoel en de meetpretentie van de test is aangegeven. Voorts wordt de theoretische achtergrond en de operationalisatie daarvan in de testinhoud beoordeeld. De beoordeling van deze categorie is van invloed op de waardering van andere categorieën, omdat de meetpretentie bepaalt welk type normerings-, betrouwbaarheids- en validiteitsonderzoek moet worden verricht.

2a. Kwaliteit van het testmateriaal. Deze categorie telt twee basisvragen. In deze categorie komt aan de orde of testopgaven, scoring en instructie zijn gestandaardiseerd.

2b. Kwaliteit van de handleiding. Deze categorie telt vijf vragen, waarvan één basisvraag. In deze categorie wordt gevraagd naar de informatie die wordt geboden ter ondersteuning van de testgebruiker bij afname en interpretatie van de test.

3. Normen. Deze categorie bevat vijf vragen waarvan twee basisvragen. Er wordt vastgesteld of er normen worden verstrekt, of de gekozen normgroepen overeenkomen met het aangegeven gebruiksdoel van de test en wat de kwaliteit is van de normen en de erbij verstrekte informatie.

Het oordeel voor elk van deze categorieën kan zijn "onvoldoende", "voldoende" of "goed". De schaalpunten op de vragen zijn "-", "+/-" en "+". Een negatief oordeel op een basisvraag leidt direct tot het oordeel "onvoldoende" voor de betreffende categorie. De inhoud van de vragen, de toelichtingen en de wegingsvoorschriften worden gegeven in het volgende hoofdstuk. Bij de wegingsvoorschriften van sommige categorieën moeten somscores van items worden berekend. Hierbij wordt aan een negatief oordeel de score min één toegekend, aan een "+/-" oordeel de score nul en aan een positief oordeel de score plus één.

De functie van de toelichtingen is houvast te geven bij de beoordeling en niet om alle statistische of psychometrische achtergronden te verduidelijken. Bij onduidelijkheden dient men zelf de relevante literatuur te op te zoeken en te bestuderen.

2. De betekenis van de beoordelingen

In het algemeen kan men stellen dat een "onvoldoende" voor een categorie op twee manieren tot stand kan komen: òf omdat de gevraagde informatie afwezig is, òf omdat de kwaliteit van de wèl aanwezige informatie negatief wordt beoordeeld. Zo kan bijvoorbeeld een "onvoldoende" voor de kwaliteit van de aanwijzingen voor de testleider betekenen dat deze aanwijzingen niet worden vermeld, of dat deze wel worden vermeld, maar onvoldoende zijn. Afwezigheid van aanwijzingen voor de testleider wordt dus hetzelfde beoordeeld als wèl beschikbare aanwijzingen die tot een negatief resultaat leiden, omdat de COTAN meent dat de auteur nu eenmaal verplicht is aanwijzingen voor de testleider te verschaffen.
Een tweede nuancering ten aanzien van de beoordeling "onvoldoende" is, dat één of meer "onvoldoendes" niet per se betekent dat een instrument onbruikbaar is. Zo kan bijvoorbeeld een "onvoldoende" voor Normen zijn gegeven, omdat de representativiteit van de normgroep te wensen over laat. Echter, de test kan zeer bruikbaar zijn wanneer de gebruiker in staat is zelf geschikte normen te verzamelen.
Een derde nuancering betreft het feit dat in een beoordelingssysteem noodzakelijkerwijs grenswaarden worden genoemd waaraan tests moeten voldoen teneinde objectiviteit bij de beoordeling te garanderen. Zo worden bij de categorie Normen specifieke steekproefgroottes genoemd waaraan moet worden voldaan voor een "voldoende" of "goed" beoordeling. Voor deze grenzen is geen sluitende wetenschappelijke argumentatie te leveren, ze zijn gebaseerd op in het algemeen min of meer geaccepteerde adviezen van vooraanstaande deskundigen. 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. Wèl wordt met het stellen van deze grenzen bewerkstelligd dat deze waarden voor alle tests in principe hetzelfde worden beoordeeld.

1. UITGANGSPUNTEN VAN DE TESTCONSTRUCTIE

Basisvraag:

1.1. Is aangegeven wat het gebruiksdoel is van de test?

-	+ / -	+

(Bij negatieve beoordeling van vraag 1.1 kan men direct doorgaan naar categorie 2).

1.2. Is de herkomst van het constructie-idee beschreven en/of word(t)(en) het(de) te meten construct(en) gedefinieerd?

-	+ / -	+

TOELICHTING BIJ CATEGORIE 1

Algemeen

Testconstructie vergt 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 interessentest), tussen personen (zoals bij personeelsselectie), of tussen groepen van personen en/of tussen situaties (zoals bij organisatieonderzoek). Op grond van de informatie die de testauteur biedt moet de toekomstige gebruiker kunnen beoordelen of de test past bij het doel waarvoor hij/zij een test zoekt. Er moet derhalve een heldere omschrijving van de meetpretentie van de test worden gegeven en de keuze van de testinhoud en de wijze waarop het (de) begrip(pen) word(t)(en) gemeten moet omstandig worden verantwoord.
In deze categorie gaat het uitsluitend om de vraag of de uitgangspunten expliciet zijn aangegeven en gaat het niet om de kwaliteit van de onderzoeksopzet en -uitvoering; deze komen in de andere categorieën aan de orde.

Bij de vaststelling van het eindoordeel van categorie 1 worden de volgende mogelijkheden onderscheiden:

	EINDOORDEEL
• De basisvraag wordt positief beoordeeld:	
- de andere vraag wordt + beoordeeld	GOED
- de andere vraag wordt +/- beoordeeld	VOLD
- de andere vraag wordt negatief beoordeeld	ONVOLD
• De basisvraag wordt +/- beoordeeld:	
- de andere vraag worden minstens +/- beoordeeld	VOLD
- de andere vraag wordt negatief beoordeeld	ONVOLD
• De basisvraag wordt negatief beoordeeld:	ONVOLD

Aanwijzingen per vraag

Ad. 1.1.
Testconstructie begint met een bezinning op het gebruiksdoel.
Doelen kunnen veelzijdig zijn, zoals: het voorspellen van een criteriumgedrag, het beoordelen van vooruitgang of training, het stellen van een diagnose t.b.v. een behandelingsplan, etcetera.

- Welk(e) doel(en) worden in de testhandleiding genoemd?
- Voor welke groep(en) is de test bedoeld? (bijv. met betrekking tot leeftijd, beroep, niveau, normaal-klinisch, enz.). Hierbij geldt: hoe meer pretenties, des te groter de verplichtingen ten aanzien van het te leveren empirische materiaal zoals normen en valideringsgegevens.

Ad. 1.2.

- Sluit de test aan bij een bestaande theorie of ontwikkelt de auteur een eigen theorie?
- Wat is globaal de theorie achter de test?
- Wordt deze theorie voldoende beschreven?
- Is het construct (zijn de constructen) voldoende gedefinieerd?
- Gaan de testauteurs in op de meerwaarde die deze test zou hebben t.o.v. reeds bestaande tests? Zo ja, wat is de meerwaarde?

Toelichting:
Ook (of juist) van tests die zijn bedoeld voor de meting van algemeen bekende begrippen, zoals intelligentie, dient een omschrijving van het begrip te worden gegeven, zodat duidelijk wordt wat wel en wat niet tot het te meten domein wordt gerekend. Wanneer de test niet zozeer theoretisch maar eerder historisch is gefundeerd, dat wil zeggen aansluit bij een traditionele wijze van meten van een bepaald type begrippen, dient duidelijk te worden gemaakt waarom juist de betreffende begrippen worden gemeten en wat de verschillen en overeenkomsten zijn met soortgelijke tests.

2. DE KWALITEIT VAN HET TESTMATERIAAL EN DE HANDLEIDING

2A. TESTMATERIAAL

Basisvraag:

2.1. Zijn de testopgaven gestandaardiseerd? | - | | + |

Basisvraag:

2.2. Is er sprake van een objectief scoringssysteem? | - | + / - | + |

(Bij negatieve beoordeling van één van de basisvragen (2.1 en 2.2) kan men direct doorgaan naar categorie 2B)

2B. HANDLEIDING

Basisvraag:

2.3. Is een handleiding beschikbaar? | - | | + |

(Bij negatieve beoordeling van vraag 2.3 kan men direct doorgaan naar categorie 3).

2.4. Zijn de aanwijzingen voor de testleider volledig en duidelijk? | - | + / - | +

2.5. Wordt in de handleiding een samenvatting van de onderzoeksresultaten gegeven? | - | + / - | +

2.6. Wordt gewezen op soorten informatie die bij de interpretatie van belang kunnen zijn? | - | + / - | +

2.7. Wordt de mate van deskundigheid die vereist is voor afname en interpretatie van de test vermeld? | - | + / - | +

TOELICHTING BIJ CATEGORIE 2

Algemeen

Wil men de score(s) op een test kunnen interpreteren als een betrouwbare maat, dan dient deze zodanig te zijn afgenomen dat andere, niet-beoogde factoren geen invloed kunnen uitoefenen op de totstandkoming van de score(s). Zo dient bijvoorbeeld de afname en instructie dusdanig te zijn gestandaardiseerd dat de invloed van variatie in instructie of verschil in proefleider op de testscore is geëlimineerd of in ieder geval binnen de grenzen van het mogelijke is gehouden.
Categorie 2 is opgedeeld in twee subcategorieën: één met betrekking tot de kwaliteit van het testmateriaal en één met betrekking tot de kwaliteit van de handleiding. Voor de subcategorieën wordt apart een beoordeling gegeven.
In de eerste subcategorie wordt gevraagd naar het ontwerp, de inhoud en de vorm van het testmateriaal, de instructie en de scoring. Deze categorie bevat twee basisvragen.

Bij de vaststelling van het eindoordeel voor categorie 2A worden de volgende mogelijkheden onderscheiden:

	EINDOORDEEL
• Alle twee de basisvragen worden positief beoordeeld:	GOED
• Een of twee basisvragen word(t)(en) +/- beoordeeld:	VOLD
• Een of beide basisvragen word(t)(en) negatief beoordeeld:	ONVOLD

In de tweede subcategorie wordt gevraagd naar de volledigheid van de informatie die de handleiding biedt met betrekking tot gebruik en interpretatie van de test.

Bij de vaststelling van het eindoordeel voor categorie 2B worden de volgende mogelijkheden onderscheiden:

EINDOORDEEL

- De basisvraag wordt positief beoordeeld:
  - minstens twee van de vragen 2.4 t/m 2.7 worden positief beoordeeld èn minder dan twee van deze vragen worden negatief beoordeeld..... GOED
  - minstens twee van de vragen 2.4 t/m 2.7 worden negatief beoordeeld èn minder dan twee van deze vragen worden positief beoordeeld ..... ONVOLD
  - alle overige gevallen ..... VOLD
- De basisvraag wordt negatief beoordeeld: ..... ONVOLD

Aanwijzingen per vraag

Ad. 2.1.
Voor het beoordelen of de testopgaven gestandaardiseerd zijn is de volgende vraag essentieel:

- Zijn de testopgaven voor wat betreft inhoud, vorm en volgorde voor iedereen hetzelfde?

Toelichting:
Dit is belangrijk wil men scores kunnen interpreteren en vergelijken. Een uitzondering met betrekking tot de eis van een uniforme volgorde van testitems wordt gemaakt voor adaptieve tests. Bij dit type tests dienen evenwel de beslissingsregels met betrekking tot de keuze van elk volgend item geëxpliciteerd te zijn.

Ad. 2.2.
Voor het beoordelen of het gaat om een objectief scoringssysteem is het antwoord op de volgend vraag essentieel:

- Liggen de waarden die aan alle mogelijke antwoorden van proefpersonen/clienten worden toegekend bij voorbaat zodanig vast, dat elke testleider, afgezien van administratieve fouten die bij de scoring kunnen worden gemaakt, tot dezelfde score zal komen?

Dit geldt met name voor schriftelijke capaciteitentests en vragenlijsten met meerkeuze items.

Ad. 2.4.
Toelichting:
De aanwijzingen voor de testleider in de handleiding hebben als belangrijkste doel ervoor te zorgen dat de testafname gestandaardiseerd plaatsvindt. Er moet zoveel mogelijk letterlijk zijn voorgeschreven wat de testleider wel en niet mag zeggen (zo is bijvoorbeeld 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 op vragen moet worden ingegaan (er kunnen bijvoorbeeld standaardteksten worden gegeven voor de antwoorden op veel voorkomende vragen).

- Is er een duidelijke en uitgeschreven instructie voor de testleider?
- Wordt aangegeven welke vragen of problemen bij afname van de test kunnen voorkomen en hoe de testleider op deze vragen of problemen dient te antwoorden?

Ad. 2.5.
Voor (toekomstige) gebruikers van een test is de handleiding de belangrijkste informatiebron. Dissertaties, artikelen in buitenlandse tijdschriften, onderzoeksrapporten en andere publicaties zijn voor hen vaak moeilijk te

verkrijgen en deze publicaties zijn door het technisch taalgebruik bovendien niet altijd even toegankelijk. Een samenvatting van de opzet en de resultaten van normerings-, betrouwbaarheids- en validiteitsonderzoek hoort derhalve in de handleiding te zijn opgenomen. Wanneer nieuw onderzoek belangrijke informatie heeft opgeleverd, moeten de gebruikers worden geïnformeerd via supplementen op de handleiding, of een herziene versie van de handleiding.
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, dat komt in de andere hoofdstukken aan de orde. Indien de betreffende informatie niet in de handleiding is opgenomen, wordt op deze vraag een negatieve beoordeling gegeven. Ook als de informatie niet via de handleiding beschikbaar is, zal de beoordelaar de wèl aanwezige informatie (proefschriften e.d.) moeten beoordelen. Meestal zal hij/zij deze toch moeten raadplegen teneinde een gefundeerd oordeel te kunnen geven, omdat een samenvatting van de onderzoeksresulaten hiertoe zelden toereikend zal zijn.

- Is er in de handleiding een samenvatting van de opzet en resultaten van normerings-, betrouwbaarheids- en validiteitsonderzoek opgenomen?
- Geef aan welke paragrafen in de handleiding een uitwerking zijn van het normerings-, betrouwbaarheids-, en validiteitsonderzoek.

Ad.2.6.

- Wordt vermeld wat de mogelijke invloed is van achtergrondvariabelen en (test)ervaring op de scores?

Ad. 2.7.

- Wordt er een adequate omschrijving van de kennis en vaardigheden gegeven die noodzakelijk wordt geacht voor de afname en interpretatie van de test?

3. NORMEN

Basisvraag:

3.1. Worden normen (waaronder verwachtingstabellen of grensscores) verstrekt? | - | | + |

Basisvraag:

3.2. Wat is de kwaliteit van de verstrekte normgroep(en)? | - | + / - | + |

3.3. Worden de betekenis en de beperkingen van het gebruikte type normscore duidelijk gemaakt voor de gebruiker en is het type normscore in overeenstemming met het doel van de test? | - | + / - | + |

3.4. Worden gemiddelden, standaardafwijkingen en gegevens over de scoreverdeling vermeld? | - | + / - | + |

3.5. Worden gegevens verstrekt over mogelijke verschillen tussen subgroepen (bijvoorbeeld allochtonen-autochtonen, vrouwen-mannen)? | - | + / - | + |

TOELICHTING BIJ CATEGORIE 3

Algemeen

Het scoren van een test levert een zogenaamde ruwe score op. In het algemeen krijgt de ruwe score pas betekenis door deze te vergelijken met een norm. De norm kan direct zijn afgeleid van een omschrijving van het domein van vaardigheden of leerstof dat men dient te beheersen (domeingerichte interpretatie), of kan zijn gebaseerd op de scoreverdeling van een vergelijkingsgroep (normgerichte interpretatie).
Bij domeingerichte interpretatie kijkt men naar wat iedere geteste afzonderlijk goed heeft gemaakt en naar wat fout gaat. Men vergelijkt het resultaat niet met dat van anderen. Bij normgerichte interpretatie vergelijkt men de behaalde score juist wel met die van anderen. Hoe scoort de onderzochte ten opzichte van andere personen waarmee een zinvolle (op grond van overeenkomsten in bijvoorbeeld leerjaar, leeftijd, functie) vergelijking kan worden gemaakt?
In principe wordt voor alle tests een van beide typen normen geeist. Er kunnen echter uitzonderingen voorkomen, met name wanneer het ipsatieve tests betreft, waarbij louter intra-individuele vergelijking wordt aanbevolen. In dat geval kan bij deze categorie "n.v.t." worden ingevuld.

Normen zijn aan slijtage onderhevig. Van de psychometrische kenmerken van een test zijn normen het meest gevoelig voor maatschappelijke veranderingen, veranderingen in het onderwijs, in de inhoud van functies e.d. Derhalve zal van tijd tot tijd hernormering van de test moeten plaatsvinden, òf zal de auteur door middel van onderzoek moeten aantonen dat hernormering niet nodig is. Teneinde de gebruiker te attenderen op mogelijk versleten normen zal aan de beoordeling van tests waarvan hernormerings- of ijkingsonderzoek sinds 15 jaar niet heeft plaatsgevonden, de kwalificatie "De normen zijn verouderd" worden toegevoegd. Na nog eens vijf jaar zonder dergelijk onderzoek wordt deze kwalificatie gewijzigd in: "Wegens veroudering zijn de normen niet meer bruikbaar". Deze kwalificaties worden eenmaal per jaar gepubliceerd in de Aanvullingen en/of de eerstvolgende uitgave van de Documentatie.

Bij de vaststelling van het eindoordeel voor categorie 3 worden de volgende mogelijkheden onderscheiden:
EINDOORDEEL

- Beide basisvragen worden positief beoordeeld:
  - somscore andere vragen $\geq +1$ ..... GOED
  - somscore andere vragen ≤ 0..... VOLD
- Basisvraag 1 wordt positief en basisvraag 2 wordt +/- beoordeeld:
  - somscore andere vragen $\geq +1$ ..... VOLD
  - somscore andere vragen ≤ 0..... ONVOLD
- Eén of beide basisvragen word(t)(en) negatief beoordeeld ..... ONVOLD

Aanwijzingen per vraag

Ad. 3.1.
Normen moeten beschikbaar zijn op het moment dat de test voor daadwerkelijk gebruik verkrijgbaar is. Bij tests die bedoeld zijn voor interpretatie op groepsniveau zijn normtabellen die zijn gebaseerd op individuele scores niet bruikbaar, en omgekeerd. Ook zijn normen niet meer bruikbaar wanneer wijzigingen in de test zelf hebben plaatsgevonden, bijvoorbeeld bij wijzigingen in de items of instructie. Wanneer deze wijziging de omzetting van een potlood-en-papier versie in een computerversie betreft, heeft dit in het algemeen weinig invloed op de waarde van normen voor vragenlijsten. Voor capaciteiten- en vaardigheidstests en/of tests die gebonden zijn aan een tijdslimiet zullen echter nieuwe normen moeten worden verzameld.

- Worden normen verstrekt?
- Om wat voor normen gaat het? (domeingericht of normgericht).

Ad. 3.2.
In principe dient de testauteur voor elk door hem/haar genoemd gebruiksdoel (zie vraag 1.1) normen te verschaffen. Het kan blijken dat de groepen waarvoor normen worden verschaft slechts een gering deel van de meetpretentie dekken. Wanneer een auteur bijvoorbeeld aangeeft dat een test is bedoeld voor keuzebegeleiding binnen het voorbereidend beroeps-onderwijs èn voor selectie voor functies op dit niveau, dan dienen voor beide situaties normen te worden verstrekt. Het is evenwel irreëel te verwachten dat voor elke functie op dit niveau normen worden verschaft.

- Worden voor de verschillende gebruikersdoelen ook verschillende normen verstrekt?

Wil een normgroep goed aan zijn doel kunnen beantwoorden (namelijk het vormen van een betrouwbare reeks van referentiepunten), dan dient de normgroep representatief te zijn voor de bedoelde groep en tevens dient de normgroep van voldoende omvang te zijn.
Om te kunnen beoordelen of de normgroepen representatief zijn, dient zowel een adekwate omschrijving van de populatie als van de wijze van steekproeftrekking of dataverzameling te worden gegeven. Het komt regelmatig voor dat de geboden informatie zo beperkt is, dat zelfs niet duidelijk is om welke populatie het gaat. Is de test bijvoorbeeld bedoeld voor landelijk of regionaal of lokaal gebruik? Wanneer de test is bedoeld voor landelijk gebruik, dan zullen in het algemeen ook landelijke normen moeten worden verzameld, behalve wanneer de auteur kan aantonen dat hiermee samenhangende variabelen (bijvoorbeeld woonplaats, stad/platte-land) geen invloed hebben. Gaat het om een doorsnee van de bevolking of om mensen die een bepaald kenmerk bezitten (bijvoorbeeld uitsluitend mensen die zich hebben aangemeld voor psychische hulpverlening, mensen met een bepaalde opleiding enz.)?
Bij de dataverzameling wordt nogal eens gebruik gemaakt van een "sample of convenience", bijvoorbeeld leerlingen met keuzeproblemen die zich aanmelden voor hulpverlening, psycho-logiestudenten omdat die makkelijk beschikbaar zijn enz. In het algemeen zijn dit slechte normgroepen, omdat niet wordt gecontroleerd voor variabelen die met de testscore kunnen samenhangen en deze groepen niet kunnen worden beschouwd als representatief voor de mo-gelijk bedoelde populaties (bijvoorbeeld respectievelijk brugklasleerlingen en studenten aan het hoger onderwijs).
In de literatuur komt men slechts spaarzaam aanbevelingen met betrekking tot de gewenste grootte van normgroepen (Angoff, 1971; Campbell, 1971) tegen. Deze aanbevelingen zijn òf gebaseerd op de berekening van meetfouten in parameters zoals gemiddelde en mediaan òf op ervaringsgegevens met betrekking tot het stabiel worden van schaalwaarden. Een synthese van deze twee gekoppeld aan het belang van de met de test te nemen beslissingen heeft de volgende beoordelingsregels opgeleverd:

- tests voor belangrijke* beslissingen op individueel niveau (bijvoorbeeld personeelsselectie, verwijzing naar speciaal onderwijs, opname/ ontslag kliniek):
  onvoldoende $N < 300$ voldoende $300 \leq N < 400$ goed $N \geq 400$
- idem, maar minder belangrijke beslissingen (bijvoorbeeld voortgangscontrole, in het algemeen beschrijvend gebruik, zoals bij beroepskeuzebegeleiding, therapie-indicatie):
  onvoldoende $N < 200$ voldoende $200 \leq N < 300$ goed $N \geq 300$
- tests voor onderzoek op groepsniveau:
  onvoldoende $N < 100$ voldoende $100 \leq N < 200$ goed $N \geq 200$

De eis met betrekking tot de steekproefgrootte geldt uiteraard per normgroep waarvoor wordt genormeerd. Bij bijvoorbeeld ontwikkelingstests die voor verschillende leeftijdsgroepen worden genormeerd kan dit verwarring geven. Indien de normering afzonderlijk per leeftijdsgroep (of leerjaar, of schooltype) wordt uitgevoerd, dan is de steekproefgrootte van deze subgroep van belang. Indien echter fit-procedures worden toegepast, waarbij gelijktijdig van de informatie van alle leeftijfdsgroepen gebruik wordt gemaakt (zie ook ad. 3.3), dan is de N van de gecombineerde groepen relevanter.

- Is de normgroep representatief voor de populatie waarvoor de test kan worden gebruikt?
- Is de normgroep van voldoende omvang?

Ad. 3.3.
Bij de omzetting van ruwe scores in afgeleide scores kan een keuze worden gemaakt uit verschillende typen schaalscores. Zo kan bijvoorbeeld gekozen worden tussen op percentielen en op standaardscores gebaseerde systemen en tussen fijner verdeelde (een systeem met veel klassen) en grover verdeelde systemen (een systeem met weinig klassen). Ook kan een testauteur voor speciale gevallen een eigen normsysteem ontwerpen of een bestaand systeem aanpassen (Verstralen, 1993). De keuze voor een bepaald systeem dient verband te houden met het doel van de test en daarmee ook met de deskundigheid van de gebruiker. Wanneer het doel van de test in de categorie "belangrijk" valt (zie ad. 3.2) zal men kiezen voor een zo nauwkeurig mogelijke meting en dus voor een fijn verdeeld systeem. Voor het testgebruik in deze categorie is echter het gebruik van waarschijnlijkheidsintervallen gewenst (zie vraag 3.6). Mede om deze reden worden aan gebruikers van dit type tests hoge deskundigheidseisen gesteld.
De keuze voor een grover systeem gaat ten koste van de precisie, maar kan de uitkomsten beter toegankelijk maken. Een dergelijke keuze verdient de voorkeur wanneer slechts globale indicaties worden gevraagd en een lager deskundigheidsniveau is vereist. Welk systeem ook door de testauteur wordt verkozen, te allen tijde dienen de kenmerken en de mogelijke voor- en nadelen van het systeem te worden beschreven en de keuze ervan te worden beargumenteerd.

- Wat voor type normscore wordt gehanteerd?
- Wordt de betekenis en de beperkingen van het gebruikte type normscore duidelijk gemaakt?
- Is het type normscore in overeenstemming met het doel van de test?

Ad. 3.4.
Deze gegevens dienen voor elke normgroep te worden vermeld. Van de verdeling zijn bijvoorbeeld kurtosis, scheefheid e.d. van belang en tevens of dit verschilt per normgroep. Zo kan het zijn dat de scores op een vragenlijst in de ene groep min of meer normaal verdeeld zijn, maar kan in een andere groep 50% de laagste score halen. Bij intelligentietests kunnen bij lagere respectievelijk hogere opleidingsniveaus bodem- en plafond-effecten optreden, waardoor de test in die groepen minder discrimineert. Dergelijke gegevens heeft de gebruiker nodig bij de interpretatie van de testscores.

- Welke gegevens worden voor de normgroep vermeld, en welke worden niet gemeld en zijn toch gewenst? (geef bij elk type gegevens kort aan wat dit type gegevens betekent – bijv wat betekent standaardafwijking).

* 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 om worden genomen.

Ad. 3.5.
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 itembias onderzoek;
- door het beschikbaar stellen van deze gegevens kan de testgebruiker zelf bepalen of hij/zij hiermee bij de interpretatie rekening wil houden.

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 sexegroepen, leeftijdsgroepen en etnische groepen.

• Worden gegevens verstrekt over mogelijke verschillen tussen subgroepen? (zoals bijv. allochtonen-autochtonen, vrouwen-mannen).