

COI-beoordelingssysteem

Een systeem voor Consistente en Objectieve
Instrumentenevaluaties in Vlaanderen



Versie 30 september 2026

Referentie

Om te verwijzen naar deze publicatie: COI-werkgroep (2026). *COI-beoordelingssysteem*.

Inhoudstabel

1.	COI-werkgroep.....	3
2.	Bronnen.....	3
3.	Toelichting bij het COI- beoordelingssysteem	4
4.	Het COI-beoordelingssysteem	6
4.1	Beschrijving van het instrument.....	6
4.2	Criterium 1 – Normen of criteriumscores.....	8
4.3	Criterium 2 – Betrouwbaarheid	21
4.4	Criterium 3 – Validiteit.....	35
4.5	Criterium 4 – Kwaliteit van het testmateriaal en de handleiding.....	39
4.6	Criterium 5 – Fairness	42
4.7	Eindbeoordeling van het instrument	46

1. COI-werkgroep

Er is nood aan consistentie en objectiviteit bij instrumentenevaluaties in Vlaanderen. Deze doelen kunnen bereikt worden door een duurzame samenwerking tussen Vlaamse stakeholders, waarbij er gebruik wordt gemaakt van één overkoepelend, praktisch bruikbaar en wetenschappelijk onderbouwd beoordelingssysteem. De COI-werkgroep ontwikkelde het huidige beoordelingssysteem om deze doelen te bereiken.

Volgende organisaties hebben bijgedragen aan de ontwikkeling van het COI-beoordelingssysteem:

- Kwaliteitscentrum voor Diagnostiek
- Project protocolering diagnostiek (Prodia)
- Belgische Federatie Psychologen – Testcommissie
- Thomas More – Assessment Hub
- Vlaams Forum Diagnostiek
- Vlaamse Vereniging voor Klinische Psychologen – Divisie psychodiagnostiek

2. Bronnen

Het COI-beoordelingssysteem werd gebaseerd op al bestaande beoordelingssystemen en instrumentenevaluaties, zijnde:

- [Het EFPA-test review model](#) (versie 4.2.6 en versie 4.5), CC BY-SA 4.0; European Federation of Psychologists' Associations AISBL (EFPA, 2025). Model for the Review, Description and Evaluation of Psychological and Educational Tests (Test Review Model): <https://www.efpa.eu/resource/efpa-test-review-model-version-2025/>
- [Het COTAN beoordelingssysteem](#) (versie 2010) en addenda; Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2010). COTAN Beoordelingssysteem voor de kwaliteit van tests. <https://nip.nl/cotan/>
- [De Test Quality Scan](#); Briers, V., Joris, S., Magez, W., Schittekatte, M., & Tierens, M. (2020). TQ-scan, een Vlaams beoordelingssysteem voor de kwaliteit van tests. <https://bfp-fbp.be/testcommissie>
- De instrumentenfiches van [PRODIA](#), het [Kwaliteitscentrum voor Diagnostiek](#) en [Thomas More](#)

3. Toelichting bij het COI-beoordelingssysteem

Het COI-beoordelingssysteem start met een **korte beschrijving van het instrument**. Daaronder komt een samenvatting van de belangrijkste sterktes en beperkingen van het instrument op basis van de beoordeling. Op de volgende bladzijde krijgen de beoordelaars de ruimte om algemene opmerkingen of aandachtspunten bij het instrument te noteren. Dit gaat om punten die niet meegenomen zijn in de beoordeling, maar wel belangrijk zijn voor de gebruiker.

Vervolgens wordt aan de beoordelaars gevraagd om het instrument te beoordelen op **vijf criteria**, zijnde:

- De normen of criteriumscores
- De betrouwbaarheid van het instrument
- De validiteit van het instrument
- De kwaliteit van het testmateriaal en de handleiding
- Fairness

Om deze criteria te beoordelen, baseren de beoordelaars zich op **al het beschikbare materiaal** over het instrument:

- Het door de testontwikkelaar beschikbaar gestelde materiaal: testboekje, sleutels, handleiding, software, artikelen, ...
- Eventueel ander, onafhankelijk onderzoek (niet uitgevoerd door de testontwikkelaar) dat werd gevonden door de procesbewaker tijdens de voorbereidingsfase.

Per criterium worden een aantal vragen gesteld die de beoordelaars dienen te scoren. Er werd gekozen om **maximaal vijf vragen per criterium** te gebruiken zodat het COI-beoordelingssysteem een vlotte beoordeling toelaat voor de beoordelaars en leidt tot overzichtelijke en bondige instrumentenfiches voor hulpverleners. Om de vragen te selecteren werd een onderscheid gemaakt tussen 'nice to know' informatie en 'need to know' informatie; welke informatie is essentieel voor gebruikers van het instrument en welke informatie is nuttig maar minder essentieel? Bij de huidige versie van het COI-beoordelingssysteem dienen maximaal 18 vragen beantwoord te worden (afhankelijk van het feit of bepaalde vragen overgeslagen kunnen worden). Bij toekomstige versies kan het aantal vragen opgedreven worden tot maximaal 25 vragen (vijf vragen per criterium).

Per vraag wordt een korte **toelichting** gegeven. Dit betreft theoretische uitleg over de aard van de vraag en richtlijnen om de vraag goed te kunnen scoren. Hoewel de richtlijnen bij de vragen vaak vrij uitgebreid zijn, zal dit de beoordelaars net helpen om tot een snellere en objectievere scoring te komen.

Om de vraag te kunnen **scoren** dient de beoordelaar te kiezen uit drie 'scores':

- Goed
- Voldoende
- Onvoldoende

Ook is er bij sommige vragen de mogelijkheid om de optie 'niet van toepassing' te kiezen. Er werd gekozen om de **schaal beperkt** te houden (3-puntsschaal). Een 5-puntschaal kan bijvoorbeeld voor meer twijfel en uiteenlopende beoordelingen zorgen bij de beoordelaars. De betekenis en het onderscheid tussen de verschillende scores kan dan immers minder duidelijk worden. Meer twijfel en minder consensus kunnen er bovendien voor zorgen dat een beoordeling langer duurt.

Om de beoordelaars te helpen bij het toewijzen van een score wordt voor de meerderheid van de vragen een '**scoringssleutel**' voorzien. Hierbij wordt per score (of voor sommige scores) voorgeschreven in welke scenario('s) de score van toepassing is. Een scoringssleutel zorgt er opnieuw voor dat een beoordeling relatief snel en objectief kan gebeuren, wat op zijn beurt zal zorgen voor meer consensus tussen de beoordelaars. Voor sommige vragen werd geen scoringssleutel voorzien omdat er geen eenduidige scenario's toe te wijzen waren aan de verschillende scores (bijvoorbeeld omdat bepaalde factoren meer kunnen doorwegen dan anderen bij specifieke instrumenten). In dit geval dient de beoordelaar zelf een score toe te wijzen, maar de toelichtingen en richtlijnen bij de vragen geven de beoordelaars steeds voldoende handvaten om de vragen op een adequate en zo objectief mogelijke manier te scoren. Bij elke scoringssleutel wordt aangegeven hoe strikt de scoringssleutel gevolgd dient te worden. Bij sommige vragen dient de scoringssleutel strikter toegepast te worden dan bij andere. Indien er wordt afgeweken van de scoringssleutel, dient de beoordelaar deze afwijking te onderbouwen.

Naast een score per vraag wordt er bij elke vraag de mogelijkheid gegeven om de **gekozen score toe te lichten of om extra aandachtspunten of opmerkingen te noteren**. Bij de meeste vragen is het toelichten van de gekozen score verplicht omdat de testgebruiker bijkomende informatie nodig heeft om het instrument ten volle te kunnen beoordelen en de instrumentscores goed te kunnen interpreteren. Om de uiteindelijke instrumentenfiches beperkt te houden, wordt aanbevolen om maximaal 100 woorden te gebruiken. Dit is een richtlijn. Het is vooral belangrijk om een goed evenwicht te vinden tussen voldoende maar duidelijke en toegankelijke info geven. Het is belangrijk om te komen tot een beschrijving die duidelijk is en niet te veel technische details omvat.

Ook werd besloten om ook een '**eindbeoordeling**' of **kwaliteitslabel** aan het instrument toe te kennen. Om beoordelaars te helpen bij het toekennen van een eindbeoordeling of kwaliteitslabel, wordt ook hiervoor een scoringssleutel beschreven. Enkele belangrijke kanttekeningen bij deze eindbeoordeling en scoringssleutel:

- Een kwaliteitslabel kan nooit op zich staan en moet onderbouwd worden met de nodige argumenten. De COI-werkgroep heeft daarom besloten dat het niet is toegestaan om een toegekend kwaliteitslabel te delen of verspreiden zonder de volledige COI-instrumentenfiche waarin de onderbouwing voor het kwaliteitslabel beschreven staat.
- De scoringssleutel dient niet strikt gevolgd te worden en dient vooral als hulpmiddel voor de beoordelaar. De beoordelaars kunnen om verschillende redenen afwijken van de

scoringssleutel. Afwijkingen dienen echter goed onderbouwd en toegelicht te worden.

Bovendien dienen beiden beoordelaars het eens te zijn over het finale kwaliteitslabel.

De eindbeoordeling bestaat eveneens uit drie mogelijke 'scores', namelijk 'goed', 'voldoende' of 'onvoldoende'.

4. Het COI-beoordelingssysteem

4.1 Beschrijving van het instrument

Naam van het instrument

Indien de afkorting van het instrument meer gekend is in de praktijk dan de volledige naam, wordt enkel de afkorting opgegeven en is het niet nodig om de volledige naam op te geven (bv. 'ABAS' in plaats van 'Adaptive Behavior Assessment System'). Indien de testontwikkelaars de versie en/of taal van het instrument in de naam of afkorting opgeven, dient dit ook te worden opgenomen op de instrumentenfiche (Bv. ABAS-3-NL). De COI-beoordeling dient dan wel degelijk enkel betrekking te hebben op deze versie en/of vertaling van het instrument.

Datum beoordeling

Hier wordt de publicatiedatum van de instrumentenfiche opgegeven. Het is vooral belangrijk dat de maand en het jaartal correct zijn, de dag maakt minder uit.

COI-beoordelingssysteem versie

Hier wordt de datum van de versie van het COI-beoordelingssysteem opgegeven dat werd gebruikt voor de beoordeling van het instrument. De versiedatum staat op het voorblad van het huidige document (bv. Versie 12 november 2025)

Sterktes en beperkingen

De beoordelaar geeft hier een kort overzicht op van de belangrijkste sterktes en beperkingen van het instrument op basis van de beoordeelde criteria. Dit overzicht van sterktes en beperkingen dient vooral om het eindlabel te ondersteunen en om de gebruiker van de fiche snel te informeren over de kwaliteit het instrument. De sterktes en beperkingen worden dus pas aangevuld nadat de beoordelaar alle criteria heeft beoordeeld. De sterktes en beperkingen worden opgelijst in bullet points en dienen zo kort en duidelijk mogelijk te worden beschreven.

Eindlabel

Hier wordt het uiteindelijke eindlabel van het instrument met diens betekenis opgegeven. De mogelijke eindlabels, hun betekenis en richtlijnen voor het toekennen van het eindlabel staan beschreven onder punt 4.7 *Eindbeoordeling van het instrument*. Dit label kan pas worden ingevuld nadat de beoordelaar alle criteria heeft beoordeeld.

Algemene opmerkingen en aandachtspunten

De beoordelaar kan hier algemene opmerkingen of aandachtspunten bij het gebruik van de beoordeling of het instrument neerschrijven. Let op: hier worden geen opmerkingen of aandachtspunten neergeschreven in verband met een van de vijf te beoordelen criteria. De beoordelaar kan hier onder meer volgende zaken aangeven:

- Of de beoordeling gaat over een specifieke versie (bv. pen-en-papier, ouderversie, ...) van het instrument (indien dit het geval is dient duidelijk geëxpliceerd te worden voor welke versie(s) de beoordeling niet geldig is);
- Of de beoordeling gaat over een specifieke subschaal of subtest van het instrument (indien dit het geval is dient duidelijk geëxpliceerd te worden voor welke subschalen of -testen de beoordeling niet geldig is);
- Of er wordt aangeraden om het instrument enkel in specifieke contexten of voor specifieke doeleinden, te gebruiken;
- Of er wordt afgeraden om het instrument in specifieke contexten of voor specifieke doeleinden te gebruiken (bv. wanneer de testontwikkelaars aangeven dat de meetpretentie of het doel van het instrument X is, maar dit wordt niet bevestigd door onderzoek).
- Ook meer praktische aspecten zoals de gemiddelde duur van de testafname, vereisten aan de doelgroep, vereiste training om het instrument af te nemen, ... kunnen hier worden neergeschreven als de beoordelaar dit relevant vindt.

Doelgroep

De beoordelaar dient hier een korte beschrijving te geven van de doelgroep zoals die wordt beschreven door de testontwikkelaars in de handleiding van het instrument. Voor welke groep is het instrument bedoeld op bijvoorbeeld vlak van leeftijd, schooltype, opleidingsniveau, ontwikkelingsniveau, ...? Let op: hier dient enkel informatie te worden gegeven over de beoogde doelgroep en nog niet over de normgroep en het bijhorende normeringsonderzoek.

Indien hier geen informatie over wordt gegeven of er geen handleiding beschikbaar is, baseert de beoordelaar zich op andere bronnen zoals literatuuronderzoek of ongepubliceerde documentatie van de testontwikkelaars. Belangrijk is dat deze zaken niet door de beoordelaar mogen ingevuld worden; als de testontwikkelaars nergens een beschrijving van doelgroep hebben opgegeven, kunnen de beoordelaars schrijven 'Dit wordt nergens beschreven door de testontwikkelaars.'

Meetpretentie en doel van het instrument

De beoordelaar dient hier een korte beschrijving te geven van de meetpretentie en het doel van het instrument zoals deze woorden beschreven door de testontwikkelaars in de handleiding van het instrument.

- Meetpretentie: Welk(e) construct(en) beoogt men te meten?

Doel: Is de test bedoeld voor screening, diagnostiek, voorspelling prestatiebeoordeling, evaluatie, ...?.

Indien hier geen informatie over wordt gegeven of er geen handleiding beschikbaar is, baseert de beoordelaar zich op andere bronnen zoals literatuuronderzoek of ongepubliceerde documentatie van de testontwikkelaars. Belangrijk is dat deze zaken niet door de beoordelaar mogen ingevuld worden; als de testontwikkelaars nergens een beschrijving van de meetpretentie en het doel van het instrument hebben opgegeven, kunnen de beoordelaars schrijven 'Dit wordt nergens beschreven door de testontwikkelaars.'

4.2 Criterium 1 – Normen of criteriumscores

Om de ruwe scores van een instrument op een zinvolle manier te interpreteren worden voor een instrument (vaak) normen ontwikkeld. Hierbij kan uitgegaan worden van een normgerichte, criteriumgerichte of domeingerichte interpretatie.

- Normgerichte interpretatie : bij deze benadering worden de ruwe scores geïnterpreteerd door ze te vergelijken met de scoreverdeling van een of meer referentiegroepen waarmee een zinvolle vergelijking (op basis van overeenkomsten in bijvoorbeeld leeftijd) kan worden gemaakt. Hierbij kan het gaan om standaardscores (bv. T-scores), percentielscores, leeftijdsequivalenten, enzoverder.
- Criteriumgerichte interpretatie : bij deze benadering worden de ruwe scores geïnterpreteerd door ze te vergelijken met een 'afkappunt' of criteriumscore. De criteriumscore bepaalt de grens voor het onderscheid tussen twee of meer groepen en bepaalt daarmee in welke groep iemand valt (bv. geslaagd of niet geslaagd). De criteriumscore wordt bepaald op basis van data over het instrument en over het te voorspellen criterium.
- Domeingerichte interpretatie: bij deze benadering worden de ruwe scores ook geïnterpreteerd door ze te vergelijken met een 'afkappunt' of criteriumscore, maar dit afkappunt wordt bepaald doormiddel van consensus tussen verschillende beoordelaars.

De beoordelaar dient na te gaan voor welke benadering werd gekozen bij het instrument in kwestie. Indien er bijvoorbeeld voor een normgerichte interpretatie werd gekozen, dan kan de beoordelaar bij alle vragen van sectie 2.2. 'criteriumgerichte interpretatie' en sectie 2.3. 'domeingerichte interpretatie' het antwoord 'niet van toepassing' kiezen.

2.1. Normgerichte interpretatie

Vraag 1. Zijn de normen representatief voor de beoogde doelgroep en het gebruik van het instrument in de praktijk?

- Toelichting bij vraag 1:

- *Ten eerste dient de normeringssteekproef representatief te zijn voor de doelpopulatie van de test. Voor testgebruik in Vlaanderen is het onder andere belangrijk dat er gebruik werd gemaakt van een Vlaamse steekproef. Andere mogelijkheden zijn het rekruteren van een ‘gecombineerde’ steekproef (bv. Vlaams-Nederlands) of het uitvoeren van equivalentieonderzoek (voor meer informatie verwijzen we naar het [EFPA-beoordelingsmodel](#)). Daarnaast dient de steekproef representatief te zijn op vlak van een aantal andere variabelen zijnde leeftijd, sekse, etniciteit en regio. Afhankelijk van bijvoorbeeld de meetpretentie of de beoogde doelgroep van het instrument, zal de steekproef uiteraard ook representatief dienen te zijn voor een aantal andere variabelen. Er wordt echter ten sterkste aangeraden dat de steekproef ten minste representatief is voor leeftijd, sekse, etniciteit/nationaliteit en regio. Voor instrumenten voor kinderen en jongeren is het ten sterkste aangeraden dat de steekproef daarnaast ook representatief is voor de sociaal-economische status en/of het opleidingsniveau van de moeder. Voor instrumenten voor volwassenen is het ten sterkste aangeraden dat de steekproef daarnaast ook representatief is voor de eigen sociaal-economische status en/of het eigen opleidingsniveau. Voor meer richtlijnen in verband met deze criteria verwijzen we naar het de ‘[Richtlijnen met betrekking tot stratificatiecriteria voor normeringsonderzoek van psychodiagnostische testen in Vlaanderen](#)’ van het Kwaliteitscentrum voor Diagnostiek (2020). Merk op dat de representativiteit van de steekproef op vlak van andere variabelen met het oog op inclusiviteit (bv. genderidentiteit) wordt beoordeeld bij Criterium 5 ‘fairness’. De mate waarin een steekproef representatief is zal afhangen van de manier waarop de steekproef gerekruteerd werd. Hieronder worden enkele mogelijke methodes voor het bekomen van een representatieve steekproef beschreven.*
 - *Aselecte steekproef: bij deze steekproef heeft elke persoon uit de populatie evenveel kans om opgenomen te worden in de steekproef. De selectie van personen gebeurt volledig ‘at random’, bijvoorbeeld door elke 10e persoon te selecteren uit een lijst van alle personen uit de populatie. Representativiteit van de steekproef is hier meestal (zeer) goed.*
 - *Gestratificeerde steekproef: Aangezien het verkrijgen van een volledig aselecte steekproef meestal niet mogelijk is (bv., omwille van een tekort aan investeringsmiddelen en privacywetgeving), maken veel testontwikkelaars gebruik van een gestratificeerde stochastische steekproef. Een dergelijke steekproef is gestratificeerd voor enkele*

belangrijke variabelen om de representativiteit te verhogen. Er worden op voorhand voor deze variabelen (bv. geslacht) beoogde aantallen vastgesteld om op deze manier representativiteit te garanderen voor deze variabelen. De representativiteit van de steekproef kan hier goed zijn, maar hangt sterk af van de procedures die gevolgd werden en de variabelen die geselecteerd werden.

- *Gelegenheidssteekproef: bij dit type wordt elke geteste persoon simpelweg toegevoegd aan de steekproef, zonder duidelijk plan. Met andere woorden, wanneer de gelegenheid zich voordoet. Denk hierbij bijvoorbeeld aan alle cliënten die zich aanmelden voor hulp bij een instantie, in een bepaalde periode. De steekproef zal hier vaak minder of niet representatief zijn.*
- *Ten tweede dienen de normen representatief te zijn voor de manier waarop de test wordt gebruikt in de praktijk.*
 - *Bijvoorbeeld: als de normgroep het instrument invulde op pen-en-papier, maar in de praktijk is er ook een digitale versie van het instrument, dan zullen de normen mogelijk niet representatief zijn voor de digitale versie van het instrument. Hierbij dient de beoordelaar een kritische beoordeling te maken. Voor vragenlijsten zal er vaak minder of weinig invloed zijn op de waarde van de normen, maar voor capaciteits- en vaardigheidsinstrumenten of tijdsgebonden instrumenten zullen de normen niet representatief zijn. Er kan ook expliciet vermeld worden – bij de beschrijving van het instrument – dat deze beoordeling niet geldt voor de digitale versie van het instrument. In dat geval weegt dit criterium niet mee voor de beoordeling van deze vraag.*
 - *Bijvoorbeeld: voor instrumenten die bedoeld zijn voor interpretatie op groepsniveau worden normtabellen voorzien die zijn gebaseerd op individuele scores, en omgekeerd.*
 - *Bijvoorbeeld: nadat de normen zijn verzameld, zijn er wijzigingen aangebracht in het instrument (bv. de items of de handleiding) dat effectief in de praktijk wordt gebruikt.*
- *Scoringssleutel bij vraag 1: Er wordt enkel een scoringssleutel voorzien voor de scores 'niet van toepassing' en 'goed' omdat er duidelijke criteria zijn bij deze scores. Er kan niet afgeweken worden van deze scoringssleutels. Voor de scores 'voldoende' en 'onvoldoende' zijn de criteria minder duidelijk omdat het onderscheid tussen beide scores zal afhangen van verschillende factoren die van toepassing zijn voor het*

instrument in kwestie. De beoordelaar dient dus zelf een afweging te maken wanneer er geen 'goed' gescoord kan worden voor het instrument. Als de beoordelaar niet over (voldoende) informatie beschikt om in te schatten hoe representatief de normen zijn (bv. wijze van steekproeftrekking, een beschrijving van de populatie), kan er een 'onvoldoende' worden gescoord, maar ook in andere gevallen kan een 'onvoldoende' aangewezen zijn.

- Niet van toepassing:
 - Er worden geen normen verstrekt omdat het bijvoorbeeld gaat om een **loutere**¹ intra-individuele vergelijking en een dergelijke vergelijking lijkt adequaat te zijn voor het instrument in kwestie (bv. bij ipsatieve instrumenten of instrumenten die vorderingen in de tijd meten);
 - of er wordt gebruik gemaakt van een domein- of criteriumgerichte interpretatie.
- Goed :
 - De normen zijn Vlaams;
 - én de normen zijn representatief voor de beoogde doelgroep, bij voorkeur op basis van de vier basisvariabelen (leeftijd, sekse, etniciteit en regio) want de steekproef werd op een kwaliteitsvolle manier gerekruteerd (bv. aselechte steekproef);
 - én de normen zijn representatief voor andere relevante variabelen voor het instrument in kwestie;
 - én de normen zijn representatief voor hoe het instrument in de praktijk wordt gebruikt.
- Voldoende
- Onvoldoende

¹ Als de testontwikkelaars aangeven dat scores ook vergelijkbaar zijn met scores van andere individuen of groepen, moeten er ook normen aanwezig zijn.

- Een korte beschrijving bij de scoring van vraag 1 is vereist.

Vraag 2. Hoe recent zijn de normen?

- Toelichting bij Vraag 2: *Normen kunnen verouderen door maatschappelijke veranderingen, demografische veranderingen, nieuwe wetenschappelijke kennis, enzovoort. Normeringsonderzoek dient dus na een zekere tijd opnieuw uitgevoerd te worden of de testontwikkelaars dienen aan te tonen dat nieuwe normeringsonderzoek niet nodig is.*
- Scoringssleutel bij vraag 2 : De grenswaardes bij de scoringssleutels werden gebaseerd op de richtlijnen uit het EFPA- en COTAN-beoordelingsmodel. De scoringssleutel hoeft niet strikt gevolgd te worden als de beoordelaar van mening is dat de richtlijnen niet van toepassing zijn op het instrument in kwestie. Afwijkingen van de scoringssleutel dienen onderbouwd te worden bij de beschrijving van deze vraag.
 - o Niet van toepassing
 - Er worden geen normen verstrekt omdat het bijvoorbeeld gaat om een loutere intra-individuele vergelijking en een dergelijke vergelijking lijkt adequaat te zijn voor het instrument in kwestie (bv. bij ipsatieve instrumenten of instrumenten die vorderingen in de tijd meten;
 - of er wordt gebruik gemaakt van een domein- of criteriumgerichte interpretatie.
 - o Goed:
 - De normen zijn niet ouder dan 15 jaar;
 - of de testontwikkelaars hebben op een adequate manier aangetoond dat nieuw normeringsonderzoek onnodig is.
 - o Voldoende: De normen zijn 15 jaar of ouder en de testontwikkelaars hebben niet (op een adequate manier) aangetoond dat nieuw normeringsonderzoek onnodig is.
 - o Onvoldoende:
 - de normen zijn 20 jaar of ouder en de testontwikkelaars hebben niet (op een adequate manier) aangetoond dat niet normeringsonderzoek onnodig is;
 - of er wordt geen of onvoldoende informatie gegeven om vraag 3 te kunnen scoren.
- Een korte beschrijving bij de scoring van vraag 2 is vereist.

Vraag 3. Zijn de steekproefgroottes van het normeringsonderzoek adequaat?

- Toelichting bij vraag 3: In het [EFPA-](#) en [COTAN-beoordelingssysteem](#) worden richtlijnen naar voren geschoven over de gewenste steekproefgroottes voor normeringsonderzoek. Deze richtlijnen worden hier overgenomen.
 - o Voor klassieke normering (bij deze methode zal men per relevante groep, zoals leeftijdsgroepen, apart de normen ontwikkelen) worden onderstaande richtlijnen gehanteerd. Let op: dit gaat om de gewenste steekproefgrootte per normgroep en niet om de totale steekproef.

Tests voor belangrijke* beslissingen op individueel niveau
(bijvoorbeeld personeelsselectie, verwijzing naar speciaal onderwijs, opname/ontslag kliniek, certificering).

$N \geq 400$	Goed
$300 \leq N < 400$	Voldoende
$N < 300$	Onvoldoende

Tests voor relatief minder belangrijke beslissingen op individueel niveau (bijvoorbeeld voortgangscontrole, in het algemeen gebruik, zoals bij beroepskeuzebegeleiding, therapie-indicatie).

$N \geq 300$	Goed
$200 \leq N < 300$	Voldoende
$N < 200$	Onvoldoende

** Met belangrijke beslissingen wordt bedoeld: beslissingen die op basis van de testcores worden genomen, die in principe, of op korte termijn, onomkeerbaar zijn, en die voor een belangrijk deel buiten de geteste persoon om worden genomen.*

- o Voor continue normering (bij deze methode zal men de normen berekenen op basis van de volledige steekproef) worden onderstaande richtlijnen gehanteerd.

Tests voor belangrijke* beslissingen op individueel niveau
(bijvoorbeeld personeelsselectie, verwijzing naar speciaal onderwijs, opname/ontslag kliniek, certificering).

$N \geq 150$	Goed
$100 \leq N < 150$	Voldoende
$N < 100$	Onvoldoende

Tests voor relatief minder belangrijke beslissingen op individueel

niveau (bijvoorbeeld voortgangscontrolle, in het algemeen gebruik, zoals bij beroepskeuzebegeleiding, therapie-indicatie).	$N \geq 100$	Goed
	$70 \leq N < 100$	Voldoende
	$N < 70$	Onvoldoende

** Met belangrijke beslissingen wordt bedoeld: beslissingen die op basis van de testcores worden genomen, die in principe, of op korte termijn, onomkeerbaar zijn, en die voor een belangrijk deel buiten de geteste persoon om worden genomen.*

- Enkele belangrijke kanttekeningen bij deze richtlijnen:
 - Voor continue normering zijn er minder grote steekproeven nodig per groep. Er zijn echter verschillende statistische aannames waaraan voldaan moet worden, namelijk (a) gelijke variantie in de subgroepen, (b) normaalverdeling van de scores in elke subgroep en (c) lineaire regressie van de testscore op leeftijd. Indien er niet voldaan wordt aan deze aannames of de aannames worden niet onderzocht, dan kunnen de resultaten een onderschatting van de benodigde steekproefgroottes opleveren.
 - De richtlijnen gelden voor situaties waarbij er acht subgroepen zijn. Wanneer er meer subgroepen zijn, zullen er minder observaties per subgroep nodig zijn en omgekeerd.
 - Bij continue normering kunnen de aantallen per subgroep verschillen om voor iedere subgroep tot gelijke precisie te komen; er zijn minder observaties nodig in de middelste groepen dan in de extreme groepen. De richtlijnen hoeven dus niet strikt gevolgd te worden.
 - In de richtlijnen wordt een onderscheid gemaakt tussen belangrijke en niet-belangrijke beslissingen. De richtlijnen voor belangrijke beslissingen zouden gevolgd moeten worden als er een belangrijke beslissing wordt genomen op basis van de testcores "die (op korte termijn) onomkeerbaar is en die buiten de geteste persoon om wordt genomen, zoals bij personeelsselectie, doorverwijzing naar een aangepaste vorm van onderwijs, opname/ontslag uit een psychiatrisch ziekenhuis". Dergelijke praktijken (het nemen van belangrijke beslissingen louter op basis van testcores en zonder het betrekken van de cliënt en/of diens context)

liggen echter niet in lijn met de [Algemene Intersectorale Richtlijn voor Diagnostiek](#) (Kwaliteitscentrum voor Diagnostiek, 2021) De beoordelaar dient zich dus af te vragen in welke mate deze opsplitsing tussen beslissingen geldt voor het te beoordelen instrument.

- Scoringssleutel bij vraag 3
 - o Niet van toepassing
 - Er worden geen normen verstrekt omdat het bijvoorbeeld gaat om een loutere intra-individuele vergelijking en een dergelijke vergelijking lijkt adequaat te zijn voor het instrument in kwestie (bv. bij ipsatieve instrumenten of instrumenten die vorderingen in de tijd meten);
 - of er wordt gebruik gemaakt van een domein- of criteriumgerichte interpretatie.
 - o Goed
 - De steekproefgrootte is goed volgens de hierboven beschreven richtlijnen en/of volgens de inschatting van de beoordelaar;
 - én in het geval van continue normering hebben de testontwikkelaars aangetoond dat er is voldaan aan de statistische assumpties of hebben ze aangetoond dat het niet voldoen aan de assumpties geen consequenties heeft voor de berekeningen.
 - o Voldoende
 - De steekproefgrootte is voldoende volgens de hierboven beschreven richtlijnen en/of volgens de inschatting van de beoordelaar
 - én in het geval van continue normering hebben de testontwikkelaars aangetoond dat er is voldaan aan de statistische assumpties of hebben ze aangetoond dat het niet voldoen aan de assumpties geen consequenties heeft voor de berekeningen.
 - o Onvoldoende
 - De steekproefgrootte is onvoldoende volgens de hierboven beschreven richtlijnen en/of volgens de inschatting van de beoordelaar;
 - of in het geval van continue normering hebben de testontwikkelaars niet aangetoond dat er is voldaan aan de statistische assumpties of dat het niet voldoen aan de assumpties geen consequenties heeft voor de berekeningen;
 - of er wordt geen of onvoldoende informatie gegeven om vraag 3 te kunnen scoren.
- Een korte beschrijving bij de scoring van vraag 3 is vereist.

2.2. Criterium gerichte interpretatie

Bij deze benadering worden de ruwe scores geïnterpreteerd door ze te vergelijken met een criteriumscore en deze criteriumscore wordt bepaald op basis van de data over het instrument en over het te voorspellen criterium. Een veelgebruikte methode hierbij is het uitvoeren van ROC-analyses waarbij de criteriumscore wordt bepaald door een afweging te maken tussen de optimale sensitiviteit (proportie van personen met de 'stoornis' en een positief testresultaat ten opzichte van alle personen met de 'stoornis') en/of specificiteit (de proportie van personen zonder de 'stoornis' en een negatief testresultaat ten opzichte van alle personen zonder de 'stoornis') van het instrument.

Merk op dat dit type onderzoek ook informatie geeft over de criteriumvaliditeit van het instrument (zie vraag 2 bij criterium 3 – validiteit). Bij dit criterium wordt de kwaliteit van het onderzoek naar de criteriumscore beoordeeld (d.w.z. transparantie en beredenering bij het bepalen van de criteriumscore, representativiteit van de beoogde doelgroep en gebruikerscontext). Bij het criterium 'validiteit' wordt de kwaliteit van het onderzoek naar de voorspellende waarde van het instrument beoordeeld (d.w.z. de kwaliteit van de referentie, de grootte van de sensitiviteits- en specificiteitswaarden).

Vraag 1. Is de steekproef representatief voor de beoogde doelgroep en gebruikerscontext?

- Toelichting bij vraag 1: De steekproef waarin de grensscore wordt bepaald moet betrekking hebben op de populatie waarvoor de test zal worden gebruikt.
 - o Bijvoorbeeld: Als er een instrument wordt ontwikkeld voor het onderzoeken van een diagnose autisme moeten er personen met een diagnose autisme en personen zonder een diagnose autisme gerekruteerd worden. Bij de groep zonder een diagnose autisme moet er voldoende variabiliteit zijn qua criteria voor een diagnose autisme. Als de steekproef alleen 'extreme' gevallen bevat (alle personen zonder diagnose autisme voldoen aan geen enkel criterium voor een diagnose autisme) dan kan dit leiden tot een overschatting van de sensitiviteit en specificiteit en zal de daaruit afgeleide grensscore dus niet optimaal zijn.
 - o Bijvoorbeeld: Als het instrument bedoeld is voor gebruik binnen een klinische populatie (diagnostisch versus screeningsinstrument) dan moeten de steekproeven ook gerekruteerd worden binnen een klinische populatie.
- Scoringssleutel bij vraag 1: Er is geen scoringssleutel bij vraag 1 omdat de beoordeling van deze vraag zal afhangen van allerlei factoren die eigen zijn aan het instrument en de doelen van het instrument. Er kan een 'onvoldoende' worden gescoord als de testontwikkelaars geen of onvoldoende informatie verschaffen over de steekproef. Indien de beoordelaar van mening is dat er om een andere reden een onvoldoende dient gescoord te worden, dan kan dit mits een toelichting in het open veld. *Let op: bij deze vraag dienen de kwaliteit van de manier waarop de referentie werd gemeten, de grootte*

van de steekproeven, en de grootte van de sensitiviteits- en specificiteitswaarden niet beoordeeld te worden. Deze punten worden beoordeeld bij vraag 2 van het validiteitscriterium.

- Niet van toepassing :
 - Er worden geen criteriumscores verstrekt omdat het bijvoorbeeld gaat om een loutere intra-individuele vergelijking en een dergelijke vergelijking lijkt adequaat te zijn voor het instrument in kwestie (bv. bij ipsatieve instrumenten of instrumenten die vorderingen in de tijd meten;
 - of er wordt gebruik gemaakt van een norm- of domein interpretatie.
- Een korte beschrijving bij de scoring van vraag 1 is vereist.

Vraag 2. Onderbouwen de testontwikkelaars de keuze voor een bepaalde criteriumscore voldoende?

- Toelichting bij vraag 2 : *bij het bepalen van een criteriumscore dienen de testontwikkelaars transparant te zijn en hun keuzes goed te onderbouwen.*
 - *Bij ROC-analyses wordt de criteriumscore bijvoorbeeld bepaald door een afweging te maken tussen optimale sensitiviteit en/of specificiteit van het instrument, er is steeds een wisselwerking. Een hogere sensitiviteit en lagere specificiteit kan bijvoorbeeld wenselijk zijn bij een screeningsinstrument voor borstkanker omdat men zoveel mogelijk gevallen van borstkanker in een vroeg stadium wilt opsporen. De kans op vals-negatieven is dan laag. De specificiteit van het instrument zal in dit geval echter lager zijn, waardoor er meer vals-positieven zijn. Het tijdig opsporen van borstkanker weegt echter zwaarder door dan ongerustheid of onnodig verder onderzoek. Testontwikkelaars dienen dergelijke afwegingen toe te lichten bij het kiezen van een hoge sensitiviteit of specificiteit, en dus het bepalen van de criteriumscore. Deze onderbouwing dient adequaat te zijn.*
 - *Daarnaast dienen de testontwikkelaars te rapporteren welke methode ze gebruikten om de criteriumscore te bepalen. Bij ROC-analyses kan het bijvoorbeeld gaan om positieve en negatieve likelihood ratio's, Youden's J statistic, Euclidian distance, enzoverder.*
- Scoringssleutel bij vraag 2 : Er is geen scoringssleutel bij vraag 2 omdat de beoordeling van deze vraag zal afhangen van allerlei factoren die eigen zijn aan het instrument en de doelen van het instrument. Er kan een 'onvoldoende' worden gescoord als de testontwikkelaars geen of onvoldoende informatie verschaffen over het bepalen van de criteriumscore. Indien de beoordelaar van mening is dat er om een andere reden een

onvoldoende dient gescoord te worden, dan kan dit mits een toelichting in het open veld.

Let op: bij deze vraag dienen de kwaliteit van de manier waarop de referentie werd gemeten, de grootte van de steekproeven, en de grootte van de sensitiviteits- en specificiteitswaarden niet beoordeeld te worden. Deze punten worden beoordeeld bij vraag 2 van het validiteitscriterium.

- Niet van toepassing :
 - Er worden geen criteriumscores verstrekt omdat het bijvoorbeeld gaat om een loutere intra-individuele vergelijking en een dergelijke vergelijking lijkt adequaat te zijn voor het instrument in kwestie (bv. bij ipsatieve instrumenten of instrumenten die vorderingen in de tijd meten;
 - of er wordt gebruik gemaakt van een norm- of domein interpretatie.
- Een korte beschrijving bij de scoring van vraag 2 is vereist.

2.3. Domeingerichte interpretatie

Bij deze benadering worden de ruwe scores geïnterpreteerd door ze te vergelijken met een criteriumscore en deze criteriumscore wordt bepaald op basis van consensus tussen beoordelaars. Beoordelaars zullen voor ieder item of iedere vraag aangeven welk gedrag van een persoon wordt verwacht bij bijvoorbeeld de grensscores voor 'voldoende' en 'onvoldoende'. Dit kan gebeuren aan de hand van verschillende methodes.

Vraag 1. Is er voldoende, aangetoonde overeenstemming tussen de beoordelaars?

- Toelichting bij vraag 1: *Om de beoordelaarsovereenstemming aan te tonen kunnen er verschillende coëfficiënten worden gebruikt. Voor data op nominaal niveau wordt vaak kappa k gebruikt en voor data op ordinaal niveau wordt vaak gewogen kappa k_w gebruikt. Voor data van intervalniveau is de intraklassecorrelatiecoëfficiënt (ICC) een veelgebruikte maat. In het EFPA- en COTAN-beoordelingssysteem worden richtlijnen beschreven in verband met het interpreteren van de groottes van deze effecten.*

Coëfficiënt kappa of intraklassecorrelatiecoëfficiënt	Interpretatie
K of $ICC \geq .80$	Goed
$.60 \leq k$ of $ICC < .80$	Voldoende
k of $ICC < .60$	Onvoldoende

- Scoringssleutel bij vraag 1 : om de scoringssleutel bij Vraag 1 te bepalen, werd gebruik gemaakt van de hierboven vermeldde grenswaardes. Hierbij moet worden opgemerkt dat het om richtlijnen gaat. De beoordelaar mag dus afwijken van de scoringssleutel als die van mening is dat de scoringssleutel niet van toepassing is op het te beoordelen instrument. Afwijkingen van de scoringssleutel dienen toegelicht te worden bij de beschrijving van deze vraag.
 - o Niet van toepassing:
 - Er worden geen criteriumscores verstrekt omdat het bijvoorbeeld gaat om een loutere intra individuele vergelijking en een dergelijke vergelijking lijkt adequaat te zijn voor het instrument in kwestie (bv. bij ipsatieve instrumenten of instrumenten die vorderingen in de tijd meten;
 - of er wordt gebruik gemaakt van een norm- of criteriumgerichte interpretatie.
 - o Goed: de coëfficiënt is groter of gelijk aan .80.
 - o Voldoende: de coëfficiënt is groter of gelijk aan .60.
 - o Onvoldoende:
 - de coëfficiënt is kleiner dan .60;
 - de testontwikkelaars rapporteren geen informatie over de overeenstemming tussen de beoordelaars.
- Een korte beschrijving bij de scoring van vraag 1 is vereist.

Vraag 2. Zijn de beoordelaars goed geselecteerd en getraind?

- Toelichting bij vraag 2 :
 - o *de beoordelaars dienen kennis te hebben van het domein waarop de test betrekking heeft.*
 - o *Het is ook wenselijk dat de beoordelaar training heeft gevolgd over het beoordelen van de getesten.*
 - o *Daarnaast dienen de beoordelaars voldoende begrip te hebben van de te gebruiken standaardbepalingsmethode zodat er geen verschillen in beoordelingen kunnen zijn tussen beoordelaars door het gebruik van verschillende methodes. Een voorbeeld van zo'n methode is de 'examinee-centered-methode'. Hierbij worden de beoordelaars gevraagd om voor ieder item uit een representatieve test aan te geven welk gedrag van de geteste op de grens van voldoende/onvoldoende wordt verwacht.*

- Scoringssleutel bij vraag 2: De scoringssleutel die bij vraag 2 wordt voorzien hoeft niet strikt gevolgd te worden als de beoordelaar van mening is dat de scoring niet van toepassing is op het instrument in kwestie. In dat geval mag de beoordelaar zelf kiezen tussen een score 'goed', 'voldoende' of 'onvoldoende'. Afwijkingen van de scoringssleutel dienen toegelicht te worden bij de beschrijving van deze vraag.
 - o Niet van toepassing :
 - Er worden geen criteriumscores verstrekt omdat het bijvoorbeeld gaat om een loutere intra-individuele vergelijking en een dergelijke vergelijking lijkt adequaat te zijn voor het instrument in kwestie (bv. bij ipsatieve instrumenten of instrumenten die vorderingen in de tijd meten;
 - of er wordt gebruik gemaakt van een norm- of criteriumgerichte interpretatie.
 - o Goed: De testontwikkelaars toonden aan dat de beoordelaars kennis hebben van het domein waar de test betrekking op heeft, de beoordelaars een training hebben gevolgd over het beoordelen van de getesten en dat de beoordelaars voldoende begrip hadden van de gebruikte standaardbepalingsmethode.
 - o Voldoende: De testontwikkelaars verschaffen informatie over twee van de criteria zoals beschreven bij de toelichting van vraag 2 (bv. beoordelaars hebben een training gevolgd en hebben kennis van het domein waar de test betrekking op heeft).
 - o Onvoldoende: De testontwikkelaars verschaffen informatie over (g)een van de criteria zoals beschreven bij de toelichting van vraag 2.
- Een korte beschrijving bij de scoring van vraag 2 is vereist.

Vraag 3. Is er transparantie over de gebruikte methode en werd de methode correct toegepast?

- Toelichting bij vraag 3: De testontwikkelaars dienen de gebruikte methodes goed te onderbouwen, rapporteren en toe te passen. Bij dit laatste kan de beoordelaar nagaan of alle gevolgde stappen in overeenstemming zijn met de voorgeschreven procedures en of de stappen bij alle beoordelaars op dezelfde manier werd gevolgd (bv. dezelfde instructies, materialen).
- Scoringssleutel bij vraag 3: De scoringssleutel bij vraag 3 wordt zo nauw mogelijk gevolgd. Enkel indien de beoordelaar goede redenen heeft, kan worden afgeweken van de scoringssleutel. Dergelijke afwijkingen dienen goed onderbouwd te worden bij de beschrijving van deze vraag.
 - o Niet van toepassing

- Goed : de testontwikkelaars hebben de gebruikte methodes goed onderbouwd, gerapporteerd en toegepast.
- Voldoende : De testontwikkelaars hebben de methode goed toegepast, maar het gebruik ervan niet (voldoende) onderbouwd.
- Onvoldoende :
 - De testontwikkelaars hebben de methode niet goed toegepast (ongeacht een goede onderbouwing);
 - Of de testontwikkelaars hebben geen informatie gerapporteerd over de gebruikte methode.

4.3 Criterium 2 – Betrouwbaarheid

De betrouwbaarheid van een instrument verwijst naar de consistentie en stabiliteit van de testresultaten. Er zijn verschillende vormen van betrouwbaarheid, zoals de interne consistentie en interbeoordelaarsbetrouwbaarheid. Bij dit criterium wordt gevraagd om het instrument te beoordelen op de meest voorkomende types (vraag 1 tot 4) en mogelijke andere vormen (vraag 5) van betrouwbaarheid. Hieronder worden enkele richtlijnen beschreven die de beoordelaar in het achterhoofd moet houden tijdens het scoren van de vragen.

- Om de praktische waarde van de betrouwbaarheidscoëfficiënten te bepalen (is de betrouwbaarheid goed, voldoende of onvoldoende?) worden in de literatuur²

² Charter, R.A. (2003). A breakdown of reliability coefficients by test type and reliability method, and the clinical implications of low reliability. *The Journal of General Psychology*, 130(3), 290–304. doi: 10.1080/00221300309601160.

verschillende grenswaardes gedefinieerd. Onderstaande tabel kan de beoordelaar een indicatie geven van de praktische waarde van de betrouwbaarheidscoëfficiënten maar het is belangrijk om in het achterhoofd te houden dat deze grenswaardes en betekenissen niet generaliseerbaar zijn naar alle instrumenten en situaties. De beoordelaar mag dus andere grenswaarden hanteren om de vragen te scoren indien die dat nodig acht voor het instrument in kwestie. Dergelijke afwijkingen dienen wel gerapporteerd te worden in de daarvoor voorziene ruimte. Let op: deze grenswaardes hebben enkel betrekking op correlatiecoëfficiënten en niet op andere types coëfficiënten.

- In sommige gevallen worden instrumenten bijvoorbeeld gebruikt om belangrijke beslissingen op individueel niveau te maken (bv. personeelselectie), en in die gevallen kan een hogere betrouwbaarheid (bv. $r \geq .90$) wenselijk zijn. Hierbij valt echter op te merken dat het zelden een goed idee is om belangrijke beslissingen over een individu te maken, enkel en alleen op basis van testcores.
- Bij test-hertestbetrouwbaarheid is het extra belangrijk om de grenswaardes niet blind te volgen en kan een grenswaarde die lager dan .70 is bijvoorbeeld al tot een 'voldoende' leiden. Dit type betrouwbaarheid wordt immers bepaald door factoren die niet enkel eigen zijn aan de test, maar bijvoorbeeld ook aan het construct dat gemeten wordt (bv. ontwikkeling).

Betrouwbaarheid	Betrouwbaarheidscoëfficiënt
-----------------	-----------------------------

Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2010). COTAN Beoordelingssysteem voor de kwaliteit van tests. <https://nip.nl/cotan/>

Goed	$r \geq .80$
Voldoende	$.70 \leq r < .80$
Onvoldoende	$r < .70$

- Instrumenten bestaan vaak uit schalen en/of subschalen. In veel gevallen wordt enkel de totaalscore van het instrument gebruikt om interpretaties te maken over het functioneren van een persoon, maar bij sommige instrumenten geven testontwikkelaars aan dat ook de schaalscores en/of subschaalscores gebruikt kunnen worden om het functioneren te interpreteren. Dit is bijvoorbeeld vaak het geval bij intelligentietesten.
 - o Indien de testontwikkelaar aangeeft dat enkel de totaalscore belangrijk is om interpretaties te maken, dan wordt enkel de betrouwbaarheid van de totaalscore beoordeeld.
 - o Indien de testontwikkelaar aangeeft dat de totaalscore het belangrijkste is, maar de scores van schalen en/of subschalen ook geïnterpreteerd kunnen worden, dan wordt bij de beoordeling vooral de betrouwbaarheid van de totaalscore in rekening gebracht. Als de beoordelaar echter van mening is dat de scores van de schalen of subschalen in de praktijk erg vaak geïnterpreteerd worden, dan wordt de betrouwbaarheid van deze scores beoordeeld.
 - o Als de testontwikkelaar geen onderscheid maakt tussen het belang van de totaalscores en de schaal- en/of subschaalscores (even belangrijk), dan wordt bij de beoordeling vooral de betrouwbaarheid van de schaal en/of subschaalscores in rekening gebracht.
- Naast de grootte van de betrouwbaarheidscoëfficiënten dient ook de kwaliteit van het onderzoek naar de coëfficiënten beoordeeld te worden. De beoordelaar kan hierbij nagaan of er meerdere (onafhankelijke) studies zijn uitgevoerd; of de steekproef voldoende beschreven wordt en voldoende groot is (een steekproef bestaande uit >100 proefpersonen wordt doorgaans als adequaat beschouwd, een steekproef bestaande uit

>200 proefpersonen wordt doorgaans als goed beschouwd³); of de standaarddeviaties van de scores gerapporteerd worden; of het aantal beoordelaars en de ervaring van de beoordelaars gerapporteerd wordt bij de interbeoordelaarsbetrouwbaarheid; of er een adequate hoeveelheid tijd tussen de meetmomenten zat bij de test-hertestbetrouwbaarheid (bijvoorbeeld zeer lang of zeer kort); enzoverder.

- Indien het gaat om een vertaling of adaptie van een bestaand instrument kan het origineel betrouwbaarheidsonderzoek gebruikt worden voor de beoordeling als er bewijs is van equivalentie. Als er geen bewezen equivalentie is, dan is generalisering van de originele resultaten niet mogelijk en kunnen deze ook niet meegenomen worden in de beoordeling van het instrument. In dit geval moet de equivalentie eerst aangetoond worden of moet er betrouwbaarheidsonderzoek uitgevoerd worden met de vertaalde of geadapteerde versie van het instrument. Voor een uitgebreidere uiteenzetting over vertaalde en geadapteerde instrumenten verwijzen we naar het [EFPA-testreview model](#).
- Er bestaan verschillende soorten betrouwbaarheidscoëfficiënten per type betrouwbaarheid. Om de interne consistentie van een instrument na te gaan kan bijvoorbeeld Cronbach's alpha, Lambda-2, de "*greatest lowerbound*", enzoverder gerapporteerd worden. Welke betrouwbaarheidscoëfficiënt het meest geschikt is, hangt af van het type test. We bevelen aan dat de beoordelaar op voorhand nagaat of de gerapporteerde betrouwbaarheidscoëfficiënten het meest geschikt zijn voor het instrument in kwestie. Voor adaptieve testen kan Cronbach's alpha bijvoorbeeld niet

³ European Federation of Psychologists' Associations AISBL (EFPA, 2025). Model for the Review, Description and Evaluation of Psychological and Educational Tests (Test Review Model). <https://www.efpa.eu/resources>

zonder meer worden gebruikt⁴. Indien de beoordelaar van mening is dat de betrouwbaarheidscoëfficiënt niet geschikt is voor het instrument maar het toch gerapporteerd werd, kan de coëfficiënt als “onvoldoende” beoordeeld worden. Dergelijke beslissing dient echter goed gestaafd te worden (zie ook de scoringsleutel).

Bij dit criterium wordt gewerkt met scoringsleutels. Per score worden verschillende scenario's beschreven die van toepassing kunnen zijn. Let op: slechts één van de scenario's hoeft van toepassing te zijn om de score toe te wijzen. De richtlijnen die hierboven werden beschreven, helpen de beoordelaar om het best passende scenario te kiezen. De scoringsleutel wordt best zo nauw mogelijk gevolgd, maar er zijn situaties waarbij er van de scoringsleutel kan worden afgeweken.

Vraag 1. Hoe goed is de interne consistentie van het instrument en werd de interne consistentie op een adequate manier onderzocht?

- Toelichting bij vraag 1: *De Interne consistentie heeft betrekking op de mate van samenhang tussen verschillende items, vragen of onderdelen van een instrument. Let op: je beoordeelt bij deze vraag enkel de interne consistentie. Als er bijvoorbeeld subschaalscores zijn waarvoor de interbeoordelaarsbetrouwbaarheid werd berekend, dan hoef je die hier niet mee te nemen in je beoordeling.*
- Scoringsleutel bij vraag 1
 - o **Niet van toepassing:** Er worden geen interne consistentiecoëfficiënten gerapporteerd en dit is volgens de beoordelaar ook niet noodzakelijk (let op: als de beoordelaar dit wel noodzakelijk acht, scoort men een 'onvoldoende'; zie verder). Het onderzoeken en rapporteren van de interne consistentie kan niet of

⁴ Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2010). COTAN Beoordelingssysteem voor de kwaliteit van tests. <https://nip.nl/cotan/>

minder relevant zijn voor het instrument in kwestie (bv. tests met een snelheidskarakter).

- **Goed:**
 - De coëfficiënt voor de totaalscore van het instrument is 'goed' en de coëfficiënt werd adequaat onderzocht;
 - en/of ten minste 80% van de coëfficiënten voor de (sub)schaalscores zijn 'goed' en/of 'voldoende' én de coëfficiënten werden adequaat onderzocht.
- **Voldoende:**
 - De coëfficiënt voor de totaalscore van het instrument is 'voldoende' en de coëfficiënt werd adequaat onderzocht;
 - of de coëfficiënt voor de totaalscore van het instrument is 'goed' maar de coëfficiënt werd niet adequaat onderzocht of er is meer informatie nodig om de kwaliteit van het onderzoek (ten volle) te kunnen beoordelen;
 - en/of ten minste 80% van de coëfficiënten voor de (sub)schaalscores zijn 'goed' en/of 'voldoende', maar de coëfficiënt(en) werd(en) niet adequaat onderzocht, of er is meer informatie nodig om de kwaliteit van het onderzoek (ten volle) te kunnen beoordelen.
- **Onvoldoende**
 - De coëfficiënt voor de totaalscore van het instrument is 'onvoldoende' (ongeacht de kwaliteit van het onderzoek);
 - en/of meer dan 20% van de coëfficiënten voor de (sub)schaalscores zijn 'onvoldoende' (ongeacht de kwaliteit van het onderzoek);
 - of de coëfficiënt voor de totaalscore van het instrument is 'voldoende' maar de coëfficiënt werd niet adequaat onderzocht of er is meer informatie nodig om de kwaliteit van het onderzoek (ten volle) te kunnen beoordelen;
 - of er werden geen interne consistentie coëfficiënt(en) onderzocht of gerapporteerd maar deze zijn wel relevant voor het instrument;
 - of het is belangrijk om de betrouwbaarheidscoëfficiënten van de subschalen te kennen, maar deze werden niet (allemaal) onderzocht en/of gerapporteerd (ongeacht onderzoek naar of rapportage van de betrouwbaarheid voor de totaalscore van het instrument);

- of er worden geen betrouwbaarheidscoëfficiënten per normgroep gerapporteerd, enkel voor de volledige normgroep⁵ (ongeacht de grootte van de coëfficiënt(en) en de kwaliteit van het onderzoek);
- of de beoordelaar is van mening dat de coëfficiënt die gerapporteerd wordt niet geschikt is voor dit type instrument en de informatie kan mogelijk misleidend zijn door bijvoorbeeld een overschatting van de coëfficiënt⁶ (deze beslissing dient goed onderbouwd te worden).

De scoringsleutel wordt zo nauw mogelijk gevolgd, maar er kan van worden afgeweken als bijvoorbeeld geen enkel scenario van toepassing is, als de betrouwbaarheidscoëfficiënten niet gerapporteerd worden voor alle versies van de test en erdoor een afweging moeten worden gemaakt (bv. verschillende leeftijden- of informantversies), of als de beoordelaar van mening is dat de toegewezen score onredelijk is. In die gevallen kan de beoordelaar zelf kiezen tussen een score 'onvoldoende', 'voldoende' of 'goed'. Afwijkingen van de scoringsleutel dienen goed onderbouwd en toegelicht te worden bij de beschrijving van deze vraag.

- Een korte beschrijving bij de scoring van vraag 1 is vereist.

Vraag 2. Hoe goed is de interbeoordelaarsbetrouwbaarheid van het instrument en werd de interbeoordelaarsbetrouwbaarheid op een adequate manier onderzocht?

⁵ De betrouwbaarheidscoëfficiënt voor de normgroep waarmee de verkregen testscore vergeleken wordt, geeft de beste indicatie voor de betrouwbaarheid van de verkregen score. Bovendien zullen betrouwbaarheidscoëfficiënten vaak artificieel hoger zijn voor de volledige normgroep of voor 'extreme' groepen dan voor alle afzonderlijke normgroepen omdat dit soort coëfficiënten sterk afhankelijk zijn van de spreiding van de scores. Het is dus misleidend om niet voor elke normgroep de betrouwbaarheidscoëfficiënten te rapporteren.

- Toelichting bij vraag 2: *De interbeoordelaarsbetrouwbaarheid heeft betrekking op de mate van overeenkomst tussen verschillende beoordelaars bij het invullen van het instrument. Een inschatting van de interbeoordelaarsbetrouwbaarheid is vooral van belang bij instrumenten waarbij de beoordelaar een subjectieve interpretatie moet maken om het instrument te scoren (bv. observatie- en beoordelingsinstrumenten). Let op: je beoordeelt bij deze vraag enkel de interbeoordelaarsbetrouwbaarheid. Als voor de totaalscore van het instrument bijvoorbeeld enkel de interne consistentie werd onderzocht, hoef je die hier niet opnieuw te beoordelen.*
- Scoringsleutel bij vraag 2
 - o **Niet van toepassing:** Er worden geen interbeoordelaarsbetrouwbarheidscoëfficiënten gerapporteerd en dit is volgens de beoordelaar ook niet noodzakelijk (let op: als de beoordelaar dit wel noodzakelijk acht, scoort men een 'onvoldoende'; zie scoringsleutel). Het onderzoeken en rapporteren van de interbeoordelaarsbetrouwbaarheid kan niet of minder relevant zijn voor het instrument in kwestie. Dit is bijvoorbeeld het geval voor instrumenten waarbij de testgebruiker geen subjectieve inschattingen hoeft te maken om het instrument te scoren. Let op: indien er sub(schalen) zijn waarbij de interbeoordelaarsbetrouwbaarheid wel relevant is (bv. observatieschaal) en deze (sub)schaalscores kunnen volgens de testontwikkelaars belangrijke informatie opleveren over het functioneren van de persoon (zie richtlijnen aan het begin van het betrouwbaarheids criterium), dan dient de interbeoordelaarsbetrouwbaarheid wel beoordeeld te worden.
 - o **Goed:**
 - De coëfficiënt voor de totaalscore van het instrument is 'goed' en de coëfficiënt werd adequaat onderzocht;
 - en/of ten minste 80% van de coëfficiënten voor de (sub)schaalscores zijn 'goed' en/of 'voldoende' én de coëfficiënten werden adequaat onderzocht.
 - o **Voldoende:**
 - De coëfficiënt voor de totaalscore van het instrument is 'voldoende' en de coëfficiënt werd adequaat onderzocht;
 - of de coëfficiënt voor de totaalscore van het instrument is 'goed' maar de coëfficiënt werd niet adequaat onderzocht of er is meer informatie nodig om de kwaliteit van het onderzoek (ten volle) te kunnen beoordelen;
 - en/of ten minste 80% van de coëfficiënten voor de (sub)schaalscores zijn 'goed' en/of 'voldoende', maar de coëfficiënt(en) werd(en) niet adequaat

onderzocht, of er is meer informatie nodig om de kwaliteit van het onderzoek (ten volle) te kunnen beoordelen.

○ **Onvoldoende**

- De coëfficiënt voor de totaalscore van het instrument is 'onvoldoende' (ongeacht de kwaliteit van het onderzoek);
- en/of meer dan 20% van de coëfficiënten voor de (sub)schaalscores zijn 'onvoldoende' (ongeacht de kwaliteit van het onderzoek);
- of de coëfficiënt voor de totaalscore van het instrument is 'voldoende' maar de coëfficiënt werd niet adequaat onderzocht of er is meer informatie nodig om de kwaliteit van het onderzoek (ten volle) te kunnen beoordelen;
- of er werden geen interbeoordlaarsbetrouwbaarheidcoëfficiënten onderzocht of gerapporteerd maar deze zijn wel relevant voor het instrument;
- of het is belangrijk om de betrouwbaarheidscoëfficiënten van de subschalen te kennen, maar deze werden niet (allemaal) onderzocht en/of gerapporteerd (ongeacht onderzoek naar of rapportage van de betrouwbaarheid voor de totaalscore van het instrument);
- of er worden geen betrouwbaarheidscoëfficiënten per normgroep gerapporteerd, enkel voor de volledige normgroep (ongeacht de grootte van de coëfficiënt(en) en de kwaliteit van het onderzoek);
- of de beoordelaar is van mening dat de coëfficiënt die gerapporteerd wordt niet geschikt is voor dit type instrument en de informatie kan mogelijk misleidend zijn door bijvoorbeeld een overschatting van de coëfficiënt (deze beslissing dient goed onderbouwd te worden).

De scoringsleutel wordt zo nauw mogelijk gevolgd, maar er kan van worden afgeweken als bijvoorbeeld geen enkel scenario van toepassing is, als de betrouwbaarheidscoëfficiënten niet gerapporteerd worden voor alle versies van de test (bv. verschillende leeftijden- of informantaversies), of als de beoordelaar van mening is dat de toegewezen score onredelijk is. In die gevallen kan de beoordelaar zelf kiezen tussen een score 'onvoldoende', 'voldoende' of 'goed'. Afwijkingen van de scoringsleutel dienen goed onderbouwd en toegelicht te worden bij de beschrijving van deze vraag.

- Een korte beschrijving bij de scoring van vraag 2 is vereist.

Vraag 3. Hoe goed is de test-hertestbetrouwbaarheid van (de subschalen van) het instrument en werd de test-hertestbetrouwbaarheid op een adequate manier onderzocht?

- Toelichting bij vraag 3: *test-hertestbetrouwbaarheid heeft betrekking op de mate van consistentie in testresultaten wanneer dezelfde test verschillende keren wordt afgenomen. Let op: je beoordeelt bij deze vraag enkel de test-hertestbetrouwbaarheid. Als voor de totaalscore van het instrument bijvoorbeeld enkel de interne consistentie werd onderzocht, hoef je die hier niet opnieuw te beoordelen.*
- Scoringsleutel bij vraag 3
 - o **Niet van toepassing:** Er worden geen test-hertestbetrouwbaarheidscoëfficiënten gerapporteerd en dit is volgens de beoordelaar ook niet noodzakelijk (let op: als de beoordelaar dit wel noodzakelijk acht, scoort men een 'onvoldoende'; zie scoringssleutel). Het onderzoeken en rapporteren van de test-hertestbetrouwbaarheid kan niet of minder relevant zijn voor het instrument in kwestie. Dit is bijvoorbeeld het geval voor instrumenten die een tijdelijke toestand of 'state' beogen te meten (bv. angstreactie op een stimulus).
 - o **Goed:**
 - De coëfficiënt voor de totaalscore van het instrument is 'goed' en de coëfficiënt werd adequaat onderzocht;
 - en/of ten minste 80% van de coëfficiënten voor de (sub)schaalscores zijn 'goed' en/of 'voldoende' én de coëfficiënten werden adequaat onderzocht.
 - o **Voldoende:**
 - De coëfficiënt voor de totaalscore van het instrument is 'voldoende' en de coëfficiënt werd adequaat onderzocht;
 - of de coëfficiënt voor de totaalscore van het instrument is 'goed' maar de coëfficiënt werd niet adequaat onderzocht of er is meer informatie nodig om de kwaliteit van het onderzoek (ten volle) te kunnen beoordelen;
 - en/of ten minste 80% van de coëfficiënten voor de (sub)schaalscores zijn 'goed' en/of 'voldoende', maar de coëfficiënt(en) werd(en) niet adequaat onderzocht, of er is meer informatie nodig om de kwaliteit van het onderzoek (ten volle) te kunnen beoordelen.
 - o **Onvoldoende**
 - De coëfficiënt voor de totaalscore van het instrument is 'onvoldoende' (ongeacht de kwaliteit van het onderzoek);
 - en/of meer dan 20% van de coëfficiënten voor de (sub)schaalscores zijn 'onvoldoende' (ongeacht de kwaliteit van het onderzoek);
 - of de coëfficiënt voor de totaalscore van het instrument is 'voldoende' maar de coëfficiënt werd niet adequaat onderzocht of er is meer

informatie nodig om de kwaliteit van het onderzoek (ten volle) te kunnen beoordelen;

- of er werden geen test-hertestbetrouwbaarheidscoëfficiënt(en) onderzocht of gerapporteerd maar deze zijn wel relevant voor het instrument;
- of het is belangrijk om de betrouwbaarheidscoëfficiënten van de subschalen te kennen, maar deze werden niet (allemaal) onderzocht en/of gerapporteerd (ongeacht onderzoek naar of rapportage van de betrouwbaarheid voor de totaalscore van het instrument);
- of er worden geen betrouwbaarheidscoëfficiënten per normgroep gerapporteerd, enkel voor de volledige normgroep (ongeacht de grootte van de coëfficiënt(en) en de kwaliteit van het onderzoek);
- of de beoordelaar is van mening dat de coëfficiënt die gerapporteerd wordt niet geschikt is voor dit type instrument en de informatie kan mogelijk misleidend zijn door bijvoorbeeld een overschatting van de coëfficiënt (deze beslissing dient goed onderbouwd te worden).

De scoringsleutel wordt zo nauw mogelijk gevolgd, maar er kan van worden afgeweken als bijvoorbeeld geen enkel scenario van toepassing is, als de betrouwbaarheidscoëfficiënten niet gerapporteerd worden voor alle versies van de test (bv. verschillende leeftijden- of informantaversies), of als de beoordelaar van mening is dat de toegewezen score onredelijk is. In die gevallen kan de beoordelaar zelf kiezen tussen een score 'onvoldoende', 'voldoende' of 'goed'. Afwijkingen van de scoringsleutel dienen goed onderbouwd en toegelicht te worden bij de beschrijving van deze vraag.

- Een korte beschrijving bij de scoring van vraag 3 is vereist.

Vraag 4. Hoe goed is de paralleltestbetrouwbaarheid van het instrument en werd de paralleltestbetrouwbaarheid op een adequate manier onderzocht?

- Toelichting bij vraag 4: *paralleltestbetrouwbaarheid heeft betrekking op de schatting van de systematische variantie ten opzichte van de totale variantie in de testscore. Paralleltestbetrouwbaarheid wordt vaak nagegaan door de correlatie te berekenen tussen scores op twee parallelle versies van een test (of (sub)schaal) die bij dezelfde groep personen zijn afgenomen. De versies beogen hetzelfde construct te meten maar bevatten andere items van dezelfde moeilijkheidsgraad. Let op: je beoordeelt bij deze vraag enkel de paralleltestbetrouwbaarheid. Als voor de totaalscore van het instrument bijvoorbeeld enkel de interne consistentie werd onderzocht, hoef je die hier niet opnieuw te beoordelen.*

- Scoringsleutel bij vraag 4
 - **Niet van toepassing:** Er worden geen paralleltestbetrouwbaarheidscoëfficiënten gerapporteerd en dit is volgens de beoordelaar ook niet noodzakelijk (let op: als de beoordelaar dit wel noodzakelijk acht, scoort men een 'onvoldoende'; zie scoringsleutel). Het onderzoeken en rapporteren van de paralleltestbetrouwbaarheid kan niet of minder relevant zijn voor het instrument in kwestie.
 - **Goed:**
 - De coëfficiënt voor de totaalscore van het instrument is 'goed' en de coëfficiënt werd adequaat onderzocht;
 - en/of ten minste 80% van de coëfficiënten voor de (sub)schaalscores zijn 'goed' en/of 'voldoende' én de coëfficiënten werden adequaat onderzocht.
 - **Voldoende:**
 - De coëfficiënt voor de totaalscore van het instrument is 'voldoende' en de coëfficiënt werd adequaat onderzocht;
 - of de coëfficiënt voor de totaalscore van het instrument is 'goed' maar de coëfficiënt werd niet adequaat onderzocht of er is meer informatie nodig om de kwaliteit van het onderzoek (ten volle) te kunnen beoordelen;
 - en/of ten minste 80% van de coëfficiënten voor de (sub)schaalscores zijn 'goed' en/of 'voldoende', maar de coëfficiënt(en) werd(en) niet adequaat onderzocht of er is meer informatie nodig om de kwaliteit van het onderzoek (ten volle) te kunnen beoordelen.
 - **Onvoldoende**
 - De coëfficiënt voor de totaalscore van het instrument is 'onvoldoende' (ongeacht de kwaliteit van het onderzoek);
 - en/of meer dan 20% van de coëfficiënten voor de (sub)schaalscores zijn 'onvoldoende' (ongeacht de kwaliteit van het onderzoek);
 - of de coëfficiënt voor de totaalscore van het instrument is 'voldoende' maar de coëfficiënt werd niet adequaat onderzocht of er is meer informatie nodig om de kwaliteit van het onderzoek (ten volle) te kunnen beoordelen;
 - of er werden geen paralleltestbetrouwbaarheidscoëfficiënten onderzocht of gerapporteerd maar deze zijn wel relevant voor het instrument;
 - of het is belangrijk om de betrouwbaarheidscoëfficiënten van de subschalen te kennen, maar deze werden niet (allemaal) onderzocht en/of

gerapporteerd (ongeacht onderzoek naar of rapportage van de betrouwbaarheid voor de totaalscore van het instrument);

- of er worden geen betrouwbaarheidscoëfficiënten per normgroep gerapporteerd, enkel voor de volledige normgroep (ongeacht de grootte van de coëfficiënt(en) en de kwaliteit van het onderzoek);
- of de beoordelaar is van mening dat de coëfficiënt die gerapporteerd wordt niet geschikt is voor dit type instrument en de informatie kan mogelijk misleidend zijn door bijvoorbeeld een overschatting van de coëfficiënt (deze beslissing dient goed onderbouwd te worden).

De scoringsleutel wordt zo nauw mogelijk gevolgd, maar er kan van worden afgeweken als bijvoorbeeld geen enkel scenario van toepassing is, als de betrouwbaarheidscoëfficiënten niet gerapporteerd worden voor alles versies van de test (bv. verschillende leeftijden- of informantversies), of als de beoordelaar van mening is dat de toegewezen score onredelijk is. In die gevallen kan de beoordelaar zelf kiezen tussen een score 'onvoldoende', 'voldoende' of 'goed'. Afwijkingen van de scoringsleutel dienen goed onderbouwd en toegelicht te worden bij de beschrijving van deze vraag.

- Een korte beschrijving bij de scoring van vraag 4 is vereist.

Vraag 5. Hoe goed zijn andere betrouwbaarheidsvormen van het instrument en werden deze andere betrouwbaarheidsvormen op een adequate manier onderzocht?

- Toelichting bij vraag 5: *Bij vraag 1 tot en met 4 werden de meest voorkomende vormen van betrouwbaarheid beoordeeld. Het is mogelijk dat de testontwikkelaars (ook) andere vormen van betrouwbaarheid onderzochten, zoals methoden op basis van item-responstheorie, methoden op basis van generaliseerbaarheidstheorie, structurele vergelijkingsmodellen, enzoverder. Bij deze vraag kan je deze andere vormen van betrouwbaarheid beoordelen. Let op: je beoordeelt bij deze vraag enkel andere vormen van betrouwbaarheid. Als voor de totaalscore van het instrument bijvoorbeeld enkel de interne consistentie werd onderzocht, hoef je die hier niet opnieuw te beoordelen.*
- Scoringsleutel bij vraag 5
 - **Niet van toepassing:** Er worden geen andere vormen van betrouwbaarheid gerapporteerd en dit is volgens de beoordelaar ook niet noodzakelijk (let op: als de beoordelaar dit wel noodzakelijk acht, scoort men een 'onvoldoende'; zie scoringsleutel).
 - **Goed:**
 - De coëfficiënt voor de totaalscore van het instrument is 'goed' en de coëfficiënt werd adequaat onderzocht;

- en/of ten minste 80% van de coëfficiënten voor de (sub)schaalscores zijn 'goed' en/of 'voldoende' én de coëfficiënten werden adequaat onderzocht.
- **Voldoende:**
 - De coëfficiënt voor de totaalscore van het instrument is 'voldoende' en de coëfficiënt werd adequaat onderzocht;
 - of de coëfficiënt voor de totaalscore van het instrument is 'goed' maar de coëfficiënt werd niet adequaat onderzocht of er is meer informatie nodig om de kwaliteit van het onderzoek (ten volle) te kunnen beoordelen;
 - en/of ten minste 80% van de coëfficiënten voor de (sub)schaalscores zijn 'goed' en/of 'voldoende', maar de coëfficiënt(en) werd(en) niet adequaat onderzocht of er is meer informatie nodig om de kwaliteit van het onderzoek (ten volle) te kunnen beoordelen.
- **Onvoldoende**
 - De coëfficiënt voor de totaalscore van het instrument is 'onvoldoende' (ongeacht de kwaliteit van het onderzoek);
 - en/of meer dan 20% van de coëfficiënten voor de (sub)schaalscores zijn 'onvoldoende' (ongeacht de kwaliteit van het onderzoek);
 - of de coëfficiënt voor de totaalscore van het instrument is 'voldoende' maar de coëfficiënt werd niet adequaat onderzocht of er is meer informatie nodig om de kwaliteit van het onderzoek (ten volle) te kunnen beoordelen;
 - of er werden geen andere vormen van betrouwbaarheid onderzocht of gerapporteerd maar deze zijn wel relevant voor het instrument;
 - of het is belangrijk om de betrouwbaarheidscoëfficiënten van de subschalen te kennen, maar deze werden niet (allemaal) onderzocht en/of gerapporteerd (ongeacht onderzoek naar of rapportage van de betrouwbaarheid voor de totaalscore van het instrument);
 - of er worden geen betrouwbaarheidscoëfficiënten per normgroep gerapporteerd, enkel voor de volledige normgroep (ongeacht de grootte van de coëfficiënt(en) en de kwaliteit van het onderzoek);
 - of de beoordelaar is van mening dat de coëfficiënt die gerapporteerd wordt niet geschikt is voor dit type instrument en de informatie kan mogelijk misleidend zijn door bijvoorbeeld een overschatting van de coëfficiënt (deze beslissing dient goed onderbouwd te worden).

De scoringsleutel wordt zo nauw mogelijk gevolgd, maar er kan van worden afgeweken als bijvoorbeeld geen enkel scenario van toepassing is, als de betrouwbaarheidscoëfficiënten niet gerapporteerd worden voor alle versies van de test (bv. verschillende leeftijden- of informantversies), of als de beoordelaar van mening is dat de toegewezen score onredelijk is. In die gevallen kan de beoordelaar zelf kiezen tussen een score 'onvoldoende', 'voldoende' of 'goed'. Afwijkingen van de scoringsleutel dienen goed onderbouwd en toegelicht te worden bij de beschrijving van deze vraag.

- Een korte beschrijving bij de scoring van vraag 5 is vereist.

4.4 Criterium 3 – Validiteit

Validiteit is de mate waarin het instrument meet wat het beoogt te meten. In de literatuur worden verschillende types validiteit onderscheiden en er zijn diverse manieren om deze verschillende types van validiteit op te delen. Zoals bij het EFPA- en COTAN-beoordelingssysteem focussen we in het COI-beoordelingssysteem op de begripsvaliditeit (of constructvaliditeit) en de criteriumvaliditeit als hoofdtypes van validiteit en onderscheiden we daarbij een aantal mogelijke subtypes van validiteit. De verschillende validiteitstypes worden toegelicht bij de beoordelvingsvragen. Enkele algemene opmerkingen:

- De richtlijnen die bij de vragen worden beschreven zijn niet limitatief, maar bieden wel een goede basis om de validiteit te beoordelen.
- Indien het gaat om een vertaalde of geadapteerde versie van een instrument moet er bewijs van equivalentie zijn. Als er geen bewezen equivalentie is, dan is generalisering van de originele onderzoeksresultaten niet mogelijk en kunnen deze ook niet meegenomen worden in de beoordeling van het instrument. In dit geval moet de equivalentie eerst aangetoond worden of er moet er validiteitsonderzoek uitgevoerd worden met de vertaalde of geadapteerde versie van het instrument. Voor een uitgebreidere uiteenzetting over vertaalde en geadapteerde instrumenten en equivalentieonderzoek, verwijzen we naar de validiteitsectie van [het EFPA-beoordelingsmodel](#).

Vraag 1. Hoe goed is de begripsvaliditeit van het instrument en werd de begripsvaliditeit op een adequate manier onderzocht?

- Toelichting bij vraag 1:
 - o *Begripsvaliditeit gaat over de overeenkomst tussen de meetpretentie van het instrument en wat het instrument daadwerkelijk meet: meet de test het bedoelde construct volledig, deels of meet het (ook) iets anders? **Veel voorkomende methodes** om begripsvaliditeit na te gaan zijn: exploratieve of confirmatieve factoranalyse; Item-test correlaties; vergelijking van de gemiddelde scores van*

groepen waarvan verwacht wordt dat er verschillen zullen zijn (bijvoorbeeld: voor een instrument dat de criteria voor een diagnose autisme meet wordt verwacht dat een groep kinderen met een diagnose autisme hoger zal scoren op het instrument dan een groep kinderen zonder een diagnose autisme); convergente validiteit (d.w.z. de mate waarin de resultaten van de test samenhangen met een ander instrument dat hetzelfde zou moeten meten); discriminante validiteit (d.w.z. de mate waarin de resultaten van de test niet samenhangen met de resultaten van een instrument dat iets anders zou moeten meten); Structuur of multi-trait-multi-method onderzoek, enzoverder.

- Bij het beoordelen van vraag 1 dient de beoordelaar niet enkel de uitkomsten, maar ook de **kwaliteit van het validiteitsonderzoek** te beoordelen. Hierbij kan de beoordelaar letten op het type methode en het aantal methodes dat werd gebruikt (de structuur of multi-trait-multi-methode is uitgebreider dan bijvoorbeeld het vergelijken van gemiddeldes tussen twee groepen; dergelijke simpele methodes wordt best aangevuld met resultaten van andere methodes); of de analyse technieken correct worden toegepast; of de steekproeven vergelijkbaar zijn met de groep(en) waarvoor de test is bedoeld (merk op: wanneer de testontwikkelaars aangeven dat het instrument bedoeld is voor gebruik in verschillende situaties en/of verschillende groepen, dan dient er onderzoek uitgevoerd te zijn in meerdere van deze situaties en/of bij meerdere groepen); de grootte van de steekproeven (werden er power-analyses uitgevoerd om de benodigde steekproefgrootte te bereiken?); de kwaliteit van andere instrumenten die werden gebruikt om de validiteit te onderzoeken (andere instrumenten moeten bijvoorbeeld voldoende betrouwbaarheid hebben).
- Wat onvoldoende, voldoende of goede validiteitswaardes zijn, hangt af van de aard van het instrument, de gebruiksdoeleinden van het instrument en het toepassingsgebied. De beoordelaar dient hier zelf kritisch over te reflecteren. Het kan helpend zijn om een vergelijking te maken met validiteitswaarden van gelijkaardige instrumenten waarvan geweten is dat ze een goede begripsvaliditeit hebben.
- Scoringsleutel bij vraag 1
 - Goed : De begripsvaliditeit is goed en gebaseerd op kwaliteitsvol onderzoek
 - Voldoende :
 - De begripsvaliditeit is voldoende en gebaseerd op kwaliteitsvol onderzoek;

- of de begripsvaliditeit is goed, maar het validiteitsonderzoek werd niet helemaal adequaat uitgevoerd en/of er is meer informatie nodig om de kwaliteit van het onderzoek ten volle te kunnen beoordelen.
- Onvoldoende
 - De begripsvaliditeit werd niet onderzocht;
 - of de begripsvaliditeit is onvoldoende (ongeacht de kwaliteit van het onderzoek);
 - of de begripsvaliditeit is voldoende of goed, maar het validiteitsonderzoek werd niet goed uitgevoerd en/of er is te weinig informatie om de kwaliteit van het onderzoek te kunnen beoordelen.

De scoringsleutel wordt zo nauw mogelijk gevolgd, maar er kan van worden afgeweken als bijvoorbeeld geen enkel scenario van toepassing is. In die gevallen kan de beoordelaar zelf kiezen tussen een score 'onvoldoende', 'voldoende' of 'goed'. Afwijkingen van de scoringsleutel dienen goed onderbouwd en toegelicht te worden bij de beschrijving van deze vraag.

- Een korte beschrijving bij de scoring van vraag 1 is vereist.

Vraag 2. Hoe goed is de criteriumvaliditeit van het instrument en werd de criteriumvaliditeit op een adequate manier onderzocht?

- Toelichting bij vraag 2 :
 - *Criteriumvaliditeit verwijst naar de mate waarin de scores van het instrument een relevante uitkomst kunnen voorspellen (bv. schoolprestaties bij een intelligentietest). Criteriumvaliditeit kan onderzocht worden door een criterium in het verleden, in het heden of in de toekomst te voorspellen.*
 - *Er zijn verschillende manieren om criteriumvaliditeit te onderzoeken. In veel gevallen worden correlaties tussen de testscores en het criterium berekend. In bepaalde gevallen (bijvoorbeeld instrumenten voor het onderzoeken of screenen van classificerende diagnoses) is het uitvoeren van ROC-analyses informatiever. ROC-analyses leveren onder andere informatie op over de sensitiviteit (de proportie van personen met de 'stoornis' en een positief testresultaat ten opzichte van alle personen met de 'stoornis') en specificiteit (de proportie van personen zonder de 'stoornis' en een negatief testresultaat ten opzichte van alle personen zonder de 'stoornis') van een instrument. Met de interpretatie van andere resultaten van ROC-analyses zoals 'positive predictive values (PPV)' en 'negative predictive values (NPV)' dient de beoordelaar op te passen omdat deze*

waarden niet intrinsiek zijn aan het instrument, maar sterk beïnvloed worden door de prevalentie van de 'stoornis' in de populatie en onderzochte steekproef.

- In onderstaande tabellen worden een aantal grenswaarden meegegeven die de beoordelaar kunnen helpen om de correlaties, sensitiviteitswaarden en specificiteitswaarden te interpreteren. Het is echter belangrijk om op te merken dat dit geen strikte criteria zijn. Wat als een sterke correlatie wordt beschouwd in de ene context, is dat niet noodzakelijk in de andere context. Hoe sterk een verband moet zijn hangt af van het doel van het instrument en de context waarin het gebruikt wordt. Voor sensitiviteits- en specificiteitswaarden is het van belang om op te merken dat er in de praktijk geen optimale sensitiviteit of specificiteit bestaat (want een hogere sensitiviteit betekent een lagere specificiteit en omgekeerd). De optimale sensitiviteit en/of specificiteit zal altijd afhangen van de populatie waarin je de test gebruikt (bv. klinische populatie) en het doel waarmee je het instrument gebruikt (zoveel mogelijk vals-positieve of vals-negatieve resultaten vermijden). De beoordelaar hoeft deze grenswaarden dus niet toe te passen en dient resultaten met een kritische blik te interpreteren. Het kan nuttig zijn om de waarden te vergelijken met andere gelijkaardige instrumenten die met hetzelfde doel en in dezelfde context worden gebruikt.

	Correlatiecoëfficiënt	Sensitiviteits- en specificiteitswaarden
Onvoldoende	$< .20$	$< 70\%$
Voldoende	$.20 \leq r < .35$	$70\% - 80\%$
Goed	$\geq .35$	$> 80\%$

- Bij het beoordelen van vraag 2 dient de beoordelaar niet enkel de uitkomsten, maar ook de **kwaliteit van het validiteitsonderzoek** te beoordelen. Hierbij kan de beoordelaar letten op het type methode dat werd gebruikt; of de analyse technieken correct worden toegepast; de grootte van de steekproeven (werden er power-analyses uitgevoerd om de benodigde steekproefgrootte te berekenen?); of de criteria op basis waarvan de proefpersonen worden toegewezen adequaat en voor iedereen gelijk zijn (bv. bij een instrument voor het onderzoeken van een diagnose autisme, voldoen aan de criteria voor een diagnose autisme volgens de DSM-5-TR); de kwaliteit van andere instrumenten die werden gebruikt om het criterium te meten (deze andere instrumenten moeten

bijvoorbeeld voldoende betrouwbaarheid hebben); de manier waarop eventuele 'real-life variabelen' (bv. schoolprestaties die objectief versus op basis van zelfrapportage) gemeten werden.

- Scoring bij vraag 2
 - o Niet van toepassing: wanneer in de handleiding wordt aangegeven dat het instrument niet ontwikkeld werd om voorspellingen te doen en dit de beoordelaar ook aannemelijk lijkt (bijvoorbeeld bij instrumenten die gebruikt worden voor voortgangscntrole).
 - o Goed : De criteriumvaliditeit is goed en gebaseerd op kwaliteitsvol onderzoek
 - o Voldoende :
 - De criteriumvaliditeit is voldoende en gebaseerd op kwaliteitsvol onderzoek;
 - of de criteriumvaliditeit is goed, maar het validiteitsonderzoek werd niet helemaal adequaat uitgevoerd en/of er is meer informatie nodig om de kwaliteit van het onderzoek ten volle te kunnen beoordelen.
 - o Onvoldoende
 - De criteriumvaliditeit werd niet onderzocht maar is wel relevant voor het instrument in kwestie;
 - of de criteriumvaliditeit is onvoldoende (ongeacht de kwaliteit van het onderzoek);
 - of de criteriumvaliditeit is voldoende of goed, maar het validiteitsonderzoek werd niet goed uitgevoerd en/of er is te weinig informatie om de kwaliteit van het onderzoek te kunnen beoordelen.

De scoringsleutel wordt zo nauw mogelijk gevolgd, maar er kan van worden afgeweken als bijvoorbeeld geen enkel scenario van toepassing is of als er bijvoorbeeld verschillende studies zijn met heel uiteenlopende sensitiviteits- en/of specificiteitswaarden. In die gevallen kan de beoordelaar zelf kiezen tussen een score 'onvoldoende', 'voldoende' of 'goed'. Afwijkingen van de scoringsleutel dienen goed onderbouwd en toegelicht te worden bij de beschrijving van deze vraag.

- Een korte beschrijving bij de scoring van vraag 2 is vereist.

4.5 Criterium 4 - Kwaliteit van het testmateriaal en de handleiding

Vraag 1. Is het testmateriaal (digitaal of pen-en-papier) adequaat en gebruiksvriendelijk voor de testafnemer⁷?

- Toelichting bij vraag 1 – *Hierbij kan de beoordelaar nagaan of er bijvoorbeeld fouten mogelijk zijn door een gebrek aan kwaliteit van het testmateriaal; of de items duidelijk en helder geformuleerd zijn; of bepaalde constructies makkelijk op te zetten zijn (bv. het klaarleggen van een puzzel); of de tekst overzichtelijk en goed te lezen is. Om deze vraag te beoordelen, kan de beoordelaar eventueel informatie inwinnen bij hulpverleners die het instrument al eerder gebruikten in de praktijk. Dit kan vooral interessant zijn wanneer de beoordelaar geen praktijkervaring met het instrument heeft. Let op: deze vraag gaat niet over de scoring of handleiding.*
- Scoring bij vraag 1 : Er wordt geen scoringsleutel voorzien bij deze vraag omdat bepaalde factoren meer kunnen doorwegen dan anderen bij specifieke instrumenten. Indien de beoordelaar een 'onvoldoende' scoort, dient er een korte toelichting te worden voorzien bij de beschrijving van deze vraag. Indien het gaat om een zelfrapportage instrument dat geen toelichting door een professional behoeft, dan kan de beoordelaar 'niet van toepassing' scoren.
- Eventuele korte beschrijving, aanvullingen en/of opmerkingen bij vraag 1

Vraag 2. Is het testmateriaal (digitaal of pen-en-papier) adequaat en gebruiksvriendelijk voor de geteste⁸?

- Toelichting bij vraag 2 – *Hierbij kan de beoordelaar bijvoorbeeld nagaan of er de instructies voor de geteste voldoende duidelijk zijn; of er (indien relevant) oefenvragen of voorbeelden zijn; of de vragen en eventuele antwoordmogelijkheden voldoende duidelijk zijn. Om deze vraag te beoordelen, kan de beoordelaar eventueel informatie inwinnen bij*

⁷ De testafnemer is de professional die het instrument afneemt bij de cliënt en of diens context.

⁸ De geteste is de persoon die getest wordt.

personen die getest werden. Dit kan vooral interessant zijn wanneer de beoordelaar geen praktijkervaring met het instrument heeft. Let op: deze vraag gaat niet over de scoring of handleiding.

- Scoring bij vraag 2: Er wordt geen scoringsleutel voorzien bij deze vraag omdat bepaalde factoren meer kunnen doorwegen dan anderen bij specifieke instrumenten. Indien de beoordelaar een 'onvoldoende' scoort, dient er een korte toelichting te worden voorzien bij de beschrijving van deze vraag.
- Eventuele korte beschrijving, aanvullingen en/of opmerkingen bij vraag 2

Vraag 3. Is het scoringssysteem (digitaal of pen-en-papier) voldoende duidelijk?

- Toelichting bij vraag 3 – *Met 'scoringssysteem' wordt verwezen naar het toekennen van een score aan de items, het sommeren van de itemscores per (sub)test of (sub)schaal, het omzetten van de somscores naar normscores of criteriumscores, de gebruiksvriendelijkheid van de scoringsformulieren, enzoverder. Hierbij kan de beoordelaar bijvoorbeeld nagaan of de scoringsprocedure duidelijk is omschreven; of er in geval van een subjectief scoringssysteem (d.w.z. scoringssystemen waarbij de beoordelaar een score toekent op basis van eigen inschatting, dit is bijvoorbeeld het geval bij observatieschalen) voldoende richtlijnen en handvaten zijn zodat de scoring zo objectief mogelijk gebeurt; of er voldoende transparantie is over het scoringsalgoritme bij een digitaal scoringssysteem; of de interpretatie van scores door de testontwikkelaars adequaat is (bv. in het geval van het toewijzen van een DSM-categorie aan een score); hoe er moet worden gescoord als er informatie ontbreekt (bv. door afbreekregel).*
- Scoring bij vraag 3: Er wordt geen scoringsleutel voorzien bij deze vraag omdat bepaalde factoren meer kunnen doorwegen dan anderen bij specifieke instrumenten. Indien de beoordelaar een 'onvoldoende' scoort, dient er een korte toelichting te worden voorzien bij de beschrijving van deze vraag.
- Eventuele korte beschrijving, aanvullingen en/of opmerkingen bij vraag 3

Vraag 4. Is de handleiding volledig, helder en gebruiksvriendelijk?

- Toelichting bij vraag 4 – *Hierbij kan de beoordelaar bijvoorbeeld nagaan of de handleiding voldoende standaardisatie toelaat (bv. voldoende instructies over wat de testafnemer wel of niet mag zeggen, over welke mate van ondersteuning er mag worden geboden); of de handleiding alle basisinformatie over het instrument omvat (d.w.z. informatie over de meetpretentie, de beoogde constructen en de bijhorende theorie, de doel/normeringsgroep); of de resultaten van het beschikbare en relevante normerings- en psychometrisch onderzoek voldoende worden gerapporteerd; of er informatie*

beschikbaar is over hoe de test ontwikkeld, vertaald of aangepast werd; of er instructies worden gegeven rond het rapporteren van de testresultaten en het communiceren van de testresultaten naar de geteste; of er voldoende transparantie is over de beperkingen van het instrument; of er – in het geval van digitale testen – voldoende informatie wordt gegeven over de installatie en het gebruik van de software.

- Scoring bij vraag 4: Er wordt geen volledige scoringsleutel voorzien bij deze vraag omdat bepaalde factoren meer kunnen doorwegen dan anderen bij specifieke instrumenten.
 - o Er dient een 'onvoldoende' te worden gescoord als (a) er geen handleiding beschikbaar is, (b) de handleiding niet alle basisinformatie over het instrument bevat en/of (c) er geen resultaten van normerings- en psychometrisch onderzoek worden gerapporteerd in de handleiding. Let op: als de beoordeling van mening is dat er een 'onvoldoende' moet worden gescoord om een andere reden, dan kan dit. Indien dit het geval is, dient er een korte toelichting te worden voorzien bij de beschrijving van deze vraag.
- Eventuele korte beschrijving, aanvullingen en/of opmerkingen bij vraag 4

4.6 Criterium 5 – Fairness

Fairness gaat over de mogelijkheid tot het gebruik van gestandaardiseerde instrumenten binnen groepen met diverse kenmerken. Denk hierbij aan geslacht, genderidentiteit, leeftijd, (motorische) beperkingen, sociaaleconomische status, culturele herkomst, land van herkomst, enzoverder. Testontwikkelaars dienen aandacht te besteden aan fairness tijdens de ontwikkeling van de test. Een test moet bruikbaar zijn binnen de volledig beoogde, diverse doelgroep zonder onbedoelde invloeden van bepaalde kenmerken.

Vraag 1. Wordt er informatie gegeven over de invulling van het beoogde construct in verschillende groepen en de mogelijke gevolgen daarvan voor de interpretatie van de testscores?

- Toelichting bij vraag 1 – *Een voorbeeld hiervan is de invulling van het intelligentieconstruct. In sommige culturen is snelheid een indicatie van intelligentie, terwijl er in andere culturen een grotere focus ligt op nauwkeurigheid. Tijdens de ontwikkelingsperiode kan hierdoor een grotere focus liggen op het aanleren van snelheid, of eerder het aanleren van nauwkeurigheid. Hierdoor kunnen de scores van een Westerse intelligentietest een vertekend beeld opleveren bij iemand met een migratieachtergrond. Let op: het is niet de bedoeling dat de beoordelaar zelf inschattingen maakt over de relevantie van het construct in verschillende culturen. De beoordelaar beoordeelt enkel of dergelijke informatie (voldoende) aanwezig is in de handleiding van het instrument.*
- Scoringsleutel bij vraag 1

- Goed: er wordt dergelijke uitleg rond het construct verstrekt en onderbouwd (bijvoorbeeld door middel van literatuuronderzoek) en de testontwikkelaar is transparant in de mogelijke impact hiervan bij bepaalde groepen.
- Voldoende: dergelijke informatie rond het construct wordt verstrekt, maar het is niet duidelijk waar deze informatie vandaan komt en/of er worden geen mogelijke implicaties beschreven door de testontwikkelaar.
- Onvoldoende: dergelijke informatie rondt het construct is afwezig.

De scoringsleutel bij deze vraag is suggestief en hoeft niet strikt gevolgd te worden. Merk op dat als er geen impact van cultuur op het beoogde construct zou zijn, dit beter ook gerapporteerd en gestaafd wordt door de testontwikkelaars. Afwezigheid van informatie zal dus meestal tot een 'onvoldoende' leiden.

- Eventuele korte beschrijving, aanvullingen en/of opmerkingen bij vraag 1

Vraag 2. Zijn de groepen waarop de normscores, criteriumscores of equivalenten gebaseerd zijn, representatief voor alle personen waarbij het instrument gebruikt kan worden?

- Toelichting bij vraag 2 – *De normeringsgroep dient de diversiteit binnen de beoogde doelgroep te weerspiegelen, met aandacht voor verschillende mogelijke kenmerken (bv. geslacht, genderidentiteit, leeftijd, motorische beperkingen, sociaaleconomische status, culturele herkomst, land van herkomst). Indien het niet mogelijk was om een representatieve groep te rekruteren, dient de mogelijke invloed van bepaalde variabelen op de scores onderzocht te worden door de testontwikkelaars (bv. door na te gaan of er verschillen zijn in scores tussen groepen) en/of dient er in de handleiding gewaarschuwd te worden voor het gebruik van het instrument binnen bepaalde groepen.*
- Scoringsleutel bij vraag 2:
 - Indien de normeringsgroep de diversiteit (op vlak van bv. geslacht, genderidentiteit, leeftijd, (motorische) beperkingen, sociaaleconomische status, culturele achtergrond) binnen de beoogde doelgroep weerspiegelt en deze representatief is, kan deze vraag gescoord worden als 'goed'.
 - Indien het niet mogelijk was om een representatieve groep te rekruteren, maar er wel inspanningen werden geleverd om mogelijke nadelige gevolgen voor de geteste te beperken door bijvoorbeeld de invloed van bepaalde variabelen te onderzoeken en/of de testafnemer te waarschuwen voor het gebruik van het instrument binnen bepaalde doelgroepen, kan deze vraag gescoord worden als 'voldoende'.

- Indien er geen inspanning werd geleverd om een representatieve steekproef te rekruteren of mogelijke nadelige gevolgen voor de geteste op een andere manier te beperken, kan deze vraag gescoord worden als 'onvoldoende'.

De scoringsleutel bij deze vraag is suggestief en hoeft niet strikt gevolgd te worden. Het is bijvoorbeeld mogelijk dat de invloed van bepaalde variabelen op de testscores wel onderzocht werden, maar dat de beoordelaar van mening is dat dit onderzoek niet goed werd uitgevoerd, waardoor de vraag toch als 'onvoldoende' wordt gescoord. Het is ook mogelijk dat er andere methodes werden gebruikt om de mogelijke nadelige gevolgen van de niet representatieve vergelijkingsgroep voor de geteste te minimaliseren.

- Een korte toelichting bij de scoring van vraag 2 is vereist

Vraag 3. Zijn de steekproeven die werden gebruikt bij het onderzoek naar de betrouwbaarheid en validiteit van het instrument representatief voor de volledige groep waarbinnen het instrument kan worden gebruikt?

- Toelichting bij vraag 3 – *de steekproef voor psychometrisch onderzoek dient de diversiteit binnen de beoogde doelgroep te weerspiegelen, met aandacht voor verschillende mogelijke kenmerken (bv. geslacht, leeftijd, motorische beperkingen, sociaaleconomische status, culturele herkomst, land van herkomst). Indien het niet mogelijk was om een representatieve steekproef te rekruteren, maar er aanwijzingen zijn voor mogelijke verschillen tussen subgroepen (bv. verschillende in gemiddelde scores tussen subgroepen of uit ander onderzoek met vergelijkbare instrumenten is gebleken dat de betrouwbaarheid en/of validiteit kan verschillen per subgroep) dan dienen de testbeoordelaars na te gaan of er verschillen zijn op vlak van betrouwbaarheid en/of validiteit en/of dient er in de handleiding gewaarschuwd te worden voor het gebruik van het instrument binnen bepaalde groepen door bijvoorbeeld een lagere betrouwbaarheid van het instrument. Let op: de beoordelaar hoeft niet zelf op zoek te gaan naar informatie over de betrouwbaarheid en validiteit van het instrument bij verschillende subgroepen. De scoring van deze vraag is gerelateerd aan de inspanningen en transparantie van de testontwikkelaar inzake de betrouwbaarheid en validiteit van het instrument bij verschillende subgroepen.*
- Scoringsleutel bij vraag 3.
 - Indien de steekproef voor het psychometrisch onderzoek de diversiteit (op vlak van bv. geslacht, leeftijd, motorische beperkingen, sociaaleconomische status, culturele achtergrond) binnen de beoogde doelgroep weerspiegelt en deze representatief is, dan kan deze vraag gescoord worden als 'goed'.

- Indien het niet mogelijk was om een representatieve steekproef te rekruteren, maar er wel inspanningen werden geleverd om mogelijke nadelige gevolgen voor de geteste te beperken door mogelijke verschillen in betrouwbaarheid en validiteit tussen subgroepen te onderzoeken en/of de testafnemer te waarschuwen voor het gebruik van het instrument binnen bepaalde doelgroepen, dan kan deze vraag gescoord worden als 'voldoende'.
- Indien informatie over de betrouwbaarheid en validiteit van het instrument bij verschillende subgroepen ontbreekt, er geen inspanning werd geleverd om een representatieve steekproef te rekruteren of mogelijke nadelige gevolgen voor de geteste op een andere manier te beperken, kan deze vraag gescoord worden als 'onvoldoende'.

De scoringsleutel bij deze vraag is suggestief en hoeft niet strikt gevolgd te worden. Het is bijvoorbeeld mogelijk dat er geen aanwijzingen zijn voor mogelijke verschillen tussen subgroepen in de betrouwbaarheid en validiteit van het instrument en dat bijkomend onderzoek niet nodig was (dan dient dit wel aangetoond en onderbouwd te worden). Het is ook mogelijk dat er andere methodes werden gebruikt om de mogelijke nadelige gevolgen van de niet representatieve steekproef voor de geteste te minimaliseren.

- Een korte toelichting bij de scoring van vraag 3 is vereist

Vraag 4. Worden er richtlijnen, aanpassingen en/of alternatieve scoringsalgoritmes voorzien voor groepen waarbij de afname van (subschalen van) het instrument minder gemakkelijk is?

- Toelichting bij vraag 4 – *Testontwikkelaars kunnen bijvoorbeeld voorzien in non-verbale index voor personen die het Nederlands minder vaardig zijn of personen met minder verbale mogelijkheden; een extra module voorzien voor specifieke doelgroepen; aanpassingen maken voor personen met beperkte motorische, visuele en/of auditieve mogelijkheden; richtlijnen voorzien voor om het instrument op een faire manier in te zetten; enzoverder.*
- Scoring bij vraag 4: Er is geen scoringsleutel bij deze vraag, de beoordelaar maakt zelf een inschatting van de best passende scoring. Bij de inschatting kan er rekening worden gehouden met de mate van inspanning om het instrument fair te maken. Het voorzien van richtlijnen om het instrument op een faire manier toe te passen, zal vaak minder grote effecten hebben dan het voorzien van mogelijke aanpassingen in het instrument. Bij de afwezigheid van inspanningen om het instrument fair te maken, zou vraag 4 als 'onvoldoende' gescoord kunnen worden, maar het is ook mogelijk dat een instrument werd ontwikkeld voor een specifieke doelgroep, waardoor er geen 'extra' inspanningen

werden geleverd om het instrument fair te maken. In die gevallen zou vraag 4 niet als 'onvoldoende' gescoord kunnen worden.

- Een korte toelichting bij de scoring van vraag 4 is vereist

Vraag 5. Zijn er andere inspanningen die de testontwikkelaar heeft geleverd om de fairness van het instrument te onderzoeken, rapporteren of verbeteren?

- Toelichting bij vraag 5 : hier kunnen alle punten in verband met fairness die niet onder de vier voorgaande vragen passen aangegeven worden.
- Scoring bij vraag 5: dit is een open vraag die niet gescoord wordt.

4.7 Eindbeoordeling van het instrument

De eindbeoordeling bestaat uit drie mogelijke 'scores', namelijk 'goed', 'voldoende' of 'onvoldoende'.

Hierbij wordt een scoringssleutel beschreven. De scoringssleutel dient niet strikt gevolgd te worden en dient vooral als leidraad. De beoordelaar kan om verschillende redenen afwijken van de scoringssleutel. Afwijkingen dienen echter goed onderbouwd en toegelicht te worden bij de sterktes en beperkingen van het instrument (zie 4.1 Beschrijving van het instrument).

Wanneer een 'voldoende' of 'onvoldoende' toegekend wordt aan het instrument, dienen de beoordelaars in het vakje met het eindlabel kort toe te lichten waarom dit eindoordeel werd toegekend.

Om tot deze eindbeoordeling te komen werd rekening gehouden met volgende aandachtspunten:

- De normen/criteriumscores, betrouwbaarheid en validiteit vormen de basisvoorwaarden voor een goed instrument, deze criteria wegen dan ook sterker door bij het bepalen van de eindbeoordeling.
- Het aantal vragen per criterium oefent geen invloed uit op de eindbeoordeling want het aantal vragen zegt niet noodzakelijk iets over het belang van het criterium.
- Het aantal keer dat er 'niet van toepassing' werd gescoord, heeft geen invloed op de eindbeoordeling.
- Het totaal aantal ingevulde vragen (soms kunnen vragen overgeslagen worden) hebben geen invloed hebben op de eindbeoordeling.

Dit leidt tot de volgende scoringssleutel voor het toekennen van eindbeoordeling 'goed', 'voldoende' of 'onvoldoende':

	Criteria
Goed	- Voor de meerderheid (d.w.z. >50%) van het totaal aantal ingevulde vragen over normscores/criteriumscores, betrouwbaarheid en validiteit werd een 'goed' gescoord; - én voor geen enkele van de vragen over normscores/criteriumscores, betrouwbaarheid en validiteit werd een 'onvoldoende' gescoord; - én voor de meerderheid (d.w.z. >50%) van het totaal aantal ingevulde vragen over de kwaliteit van het testmateriaal en de handleiding en fairness werd een 'goed' gescoord.
Onvoldoende	- Bij meer dan één vraag over normen/criteriumscores, betrouwbaarheid of validiteit, werd een 'onvoldoende' gescoord. ➔ Bv. bij een vraag over de normen/criteriumscores en bij een vraag over de betrouwbaarheid werd een 'onvoldoende' werd gescoord. ➔ Bv. Bij twee of meer vragen over de validiteit werd een 'onvoldoende' gescoord. - En/of voor de meerderheid (d.w.z. >50%) van alle ingevulde vragen over de kwaliteit van het testmateriaal en de handleiding en fairness, werd een 'onvoldoende' gescoord (ongeacht de scores voor de kerncriteria). Merk op: in sommige gevallen kan slechts één onvoldoende bij de vragen over normen/criteriumscores, betrouwbaarheid of validiteit tot het eindoordeel 'onvoldoende' leiden. Dergelijke beslissing dient wel goed onderbouw te worden. Merk op: bij dit eindlabel dienen de beoordelaars toe te lichten waarom er een onvoldoende werd toegekend.
Voldoende	Deze score wordt toegewezen in alle andere situaties waarin geen eindbeoordeling 'goed' of 'onvoldoende' wordt toegekend. Merk op: bij dit eindlabel dienen de beoordelaars toe te lichten waarom er een onvoldoende werd toegekend.