**EFPA REVIEW MODEL FOR**

**THE DESCRIPTION AND EVALUATION OF**

**PSYCHOLOGICAL AND EDUCATIONAL TESTS**

***TEST REVIEW FORM AND NOTES FOR REVIEWERS***

**VERSION 4.2.6**

Nederlandse vertaling door het Kwaliteitscentrum voor Diagnostiek vzw

**EFPA-BEOORDELINGSMODEL VOOR**

**DE BESCHRIJVING EN EVALUATIE VAN**

**PSYCHOLOGISCHE EN EDUCATIEVE TESTEN**

***TESTBEOORDELINGSFORMULIER EN AANTEKENINGEN VOOR BEOORDELAARS*VERSIE 4.2.6**

Versie 4.2.6. is een uitgebreide herziening van Versie 3-42   
(2008) door een taskforce van de Evaluatiecommissie van de EFPA bestaande uit:

Arne Evers (voorzitter, Nederland)  
Carmen Hagemeister (Duitsland)  
Andreas Høstmælingen (Noorwegen)  
Patricia Lindley (Verenigd Koninkrijk)  
José Muñiz (Spanje)  
Anders Sjöberg (Zweden)

**Goedgekeurd door de Algemene Vergadering van EFPA, 13-07-2013**

**© EFPA**Gebruikers van dit document en de inhoud ervan zijn door EFPA verplicht deze bron te erkennen met de volgende tekst:

"De EFPA-Testbeoordelingscriteria zijn grotendeels gebaseerd op de vorm en inhoud van de testbeoordelingscriteria van de British Psychological Society (BPS) en de criteria ontwikkeld door de Commissie Testaangelegenheden Nederland (COTAN) en het Nederlands Instituut van Psychologen (NIP). EFPA is de BPS en het NIP erkentelijk voor de toestemming te vertrekken van hun criteria voor de ontwikkeling van het Europese model. Alle intellectuele eigendomsrechten voor de oorspronkelijke BPS en NIP criteria worden erkend en blijven onder deze instanties."

Inhoud

[1 Inleiding 3](#_Toc31638057)

[DEEL 1: BESCHRIJVING VAN HET INSTRUMENT 5](#_Toc31638058)

[2 Algemene beschrijving 6](#_Toc31638059)

[3 Classificering 8](#_Toc31638060)

[4 Meting en scoring 16](#_Toc31638061)

[5 Computergegenereerde rapporten 18](#_Toc31638062)

[6 Voorwaarden en kosten van het aanbod 25](#_Toc31638063)

[DEEL 2: EVALUATIE VAN HET INSTRUMENT 29](#_Toc31638064)

[7 Kwaliteit van de uitleg van de basisprincipes, de presentatie en de beschikbare informatie 32](#_Toc31638065)

[7.1 Kwaliteit van de uitleg van de basisprincipes 32](#_Toc31638066)

[7.2 Toereikendheid van de voor de testleider beschikbare documentatie 33](#_Toc31638067)

[7.3 Kwaliteit van de procedurele instructies voorzien voor de testleider 35](#_Toc31638068)

[8 Kwaliteit van het testmateriaal 38](#_Toc31638069)

[8.1 Kwaliteit van het testmateriaal van pen-en-papier-testen 38](#_Toc31638070)

[8.2 Kwaliteit van de testmaterialen van CBT- en WBT-testen 38](#_Toc31638071)

[9 Normen 40](#_Toc31638072)

[9.1 Normgerichte interpretatie 40](#_Toc31638073)

[9.2 Criteriumgerefereerde interpretatie 46](#_Toc31638074)

[10 Betrouwbaarheid 51](#_Toc31638075)

[11 Validiteit 62](#_Toc31638076)

[11.1 Begripsvaliditeit 64](#_Toc31638077)

[11.2 Criteriumvaliditeit 69](#_Toc31638078)

[11.3 Algemene validiteit 72](#_Toc31638079)

[12 Kwaliteit van computergegenereerde rapporten 74](#_Toc31638080)

[13 Eindevaluatie 79](#_Toc31638081)

[DEEL 3: BIBLIOGRAFIE 81](#_Toc31638082)

[APPENDIX Een geheugensteun met essentiële punten voor evaluatie indien een instrument vertaald en/of aangepast werd van een niet-lokale context. 85](#_Toc31638083)

# Inleiding

De voornaamste doelstelling van het EFPA (European Federation of Psychologists’ Association)-Testbeoordelingsmodel is het beschikbaar stellen van een beschrijving en een gedetailleerde en rigoureuze beoordeling van de psychologische diagnostische testen, schalen en vragenlijsten die gebruikt worden in de domeinen van Arbeid, Onderwijs, Gezondheid en andere contexten. Deze informatie zal beschikbaar worden gesteld aan testleiders[[1]](#footnote-2) en professionals opdat testen en testafnames verbeterd kunnen worden, alsook om hen te helpen de juiste beoordelingskeuzes te maken. Het EFPA-Testbeoordelingsmodel is onderdeel van de informatiestrategie van de EFPA, die ernaar streeft om evaluaties te voorzien van alle noodzakelijke technische informatie over testen om zo het gebruik ervan te verbeteren (Evers et al., 2012; Muñiz & Bartram, 2007). In navolging van de *Standards for Educational and Psychological Testing* wordt het label *test* gebruikt voor elk beoordelingsapparaat of -procedure die een steekproef van het gedrag van de ondervraagde bekomt in een specifiek domein en vervolgens wordt geëvalueerd en gescoord volgens een standaardprocedure. (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 3). Daarom is dit beoordelingsmodel toepasbaar op alle instrumenten die onder deze definitie vallen, ongeacht of dit een schaal, vragenlijst, projectieve techniek of dergelijke wordt genoemd.   
 De originele versie van het EFPA-Testbeoordelingsmodel komt voort uit een aantal bronnen, waaronder: het BPS (British Psychological Society)-Testbeoordelingsformulier (ontwikkeld door *Newland Park Associates Limited*, NPAL, en later overgenomen door het BPS-Stuurgroep voor Teststandaarden); de Spaanse Vragenlijst voor de Evaluatie van Psychometrische Testen (ontwikkeld door de *Spanish Psychological Association*) en het COTAN-Beoordelingssysteem voor de Kwaliteit van Tests (ontwikkeld door de Commissie Testaangelegenheden Nederland van het Nederlandse Instituut van Psychologen). Veel van de inhoud werd aangepast met toestemming van de *Review Proforma*, door *Newland Park Associates Ltd.* initieel ontwikkeld in 1989 voor een beoordeling van testen door opleidingsmedewerkers in het Verenigd Koninkrijk (zie Bartram, Lindley & Foster, 1990). Dit model werd vervolgens gebruikt en verder ontwikkeld voor de beoordeling van een reeks instrumenten van de BPS voor het gebruik in beroepsmatige beoordeling (e.g. Bartram, Lindley, & Foster, 1992; Lindley et al., 2001). Na een initiële EFPA-workshop in maart 2000 en de aansluitende overlegrondes, werd de eerste versie van het EFPA-beoordelingsmodel samengesteld en bewerkt door Dave Bartram (Bartram, 2002a, 2002b). Een uitgebreide update en herwerking werd uitgevoerd door Patricia Lindley, Dave Bartram, en Natalie Kennedy voor het gebruik in het BPS-beoordelingssysteem (Lindley et al., 2004). Dit werd vervolgens overgenomen door EFPA in 2005 (Lindley et al., 2005) met kleine revisies in 2008 (Lindley et al. 2008). De huidige versie van het model werd samengesteld door een werkgroep van de EFPA-Beoordelingsraad, bestaande uit de leden: Arne Evers (voorzitter, Nederland), Carmen Hagemeister (Duitsland), Andreas Høstmælingen (Noorwegen), Patricia Lindley (Verenigd Koninkrijk), José Muñiz (Spanje) en Anders Sjöberg (Zweden). De aantekeningen en checklist voor vertaalde en aangepaste testen, die geproduceerd werden door Pat Lindley en de *Consultant Editors* van de testbeoordelingen in het Verenigd Koninkrijk, werden in deze versie geïntegreerd (Lindley, 2009). De teksten van enkele grote hernieuwde passages zijn gebaseerd op het herwerkte COTAN-Beoordelingssysteem voor de Kwaliteit van Tests (Evers, Lucassen, Meijer, & Sijtsma, 2010; Evers, Sijtsma, Lucassen, & Meijer, 2010).   
 Het EFPA-Testbeoordelingsmodel is onderverdeeld in drie grote delen. In het eerste deel (Beschrijving van het instrument) worden alle kenmerken van de geëvalueerde test in detail beschreven. In het tweede deel (Evaluatie van het instrument) worden de fundamentele eigenschappen van de test geëvalueerd: testmaterialen, normen, betrouwbaarheid, validiteit en computergegenereerde rapporten, inclusief een globale finale evaluatie. In het derde deel (Bibliografie) worden de referenties opgenomen die in de beoordeling zijn gebruikt.   
 Even belangrijk als het model zelf, is de juiste implementatie van het model. De huidige versie is bedoeld voor gebruik door twee onafhankelijke beoordelaars in een *peer review*proces, vergelijkbaar met de gebruikelijke evaluatie van wetenschappelijke artikelen en projecten. Een adviserend redacteur houdt toezicht op de reviews en kan een derde beoordelaar betrekken indien er significante discrepanties tussen de twee reviews worden waargenomen. Enkele variaties in de procedure zijn mogelijk, zolang de competenties en onafhankelijkheid van de beoordelaars en de adviserende redacteur gewaarborgd blijven. EFPA adviseert dat de evaluaties in deze reviews worden gericht aan gekwalificeerde praktiserende testleiders. Niettemin kunnen de reviews ook interessant zijn voor academici, testauteurs en specialisten in psychometrie en psychologische testing.   
 Een ander essentieel punt is de publicatie van de resultaten van een testevaluatie. De resultaten dienen beschikbaar te zijn voor alle professionals en gebruikers (betalend of gratis). Een goede optie is om resultaten beschikbaar te stellen op de website van de *National Psychological Association*, alhoewel deze ook gepubliceerd kunnen worden door derden of in andere media zoals tijdschriften en boeken.   
 De intentie achter het breed beschikbaar maken van dit model is om afstemming binnen de beoordelingsprocedures en -criteria in Europa te stimuleren. Hoewel harmonisering één van de doelen van het model is, is een ander doel van het model om een systeem voor testbeoordeling aan te bieden aan landen die geen eigen beoordelingsprocedure hebben. Wanneer landen effectief het model beginnen te gebruiken, is het mogelijk dat lokale kwesties aanpassingen binnen het EFPA-Testbeoordelingsmodel noodzakelijk maken. Daarom wordt het een ‘model’ genoemd, om te benadrukken dat lokale aanpassingen mogelijk zijn om een betere aansluiting te garanderen met de lokale behoeften.   
 Reacties op het EFPA-Testbeoordelingsmodel zijn welkom vermits de ervaringen van testleiders een rol kunnen spelen in de verbetering en verheldering van de processen.

DEEL 1: BESCHRIJVING VAN HET INSTRUMENT

# Algemene beschrijving

In dit onderdeel van het formulier dient de algemene informatie vermeld te worden die noodzakelijk is voor de identificatie van het instrument en die aangeeft waar het te verkrijgen is. Het hoort de titel van het instrument te vermelden, de uitgever en/of distributeur, de auteur(s), de datum van de originele publicatie en de datum van de versie die wordt beoordeeld.

De vragen 2.1.1. tot 2.7.3. zouden duidelijk moeten zijn. Ze betreffen feitelijke informatie, hoewel enig oordeel nodig zal zijn om de informatie over inhoudsdomeinen in te vullen.

|  |  |  |
| --- | --- | --- |
|  | **Beoordelaar[[2]](#footnote-3)** | Klik of tik om tekst in te voeren. |
|  | **Datum van huidige beoordeling** | Klik of tik om een datum in te voeren. |
|  | **Datum van voorgaande beoordeling** *(indien van toepassing)*[[3]](#footnote-4) | Klik of tik om een datum in te voeren. |
| 2.1.1 | **Naam instrument (lokale versie)** | Klik of tik om tekst in te voeren. |
| 2.1.2 | **Verkorte naam van de test** *(indien van toepassing)* | Klik of tik om tekst in te voeren. |
| 2.2 | **Originele naam van de test** *(indien lokale versie een aanpassing is)* | Klik of tik om tekst in te voeren. |
| 2.3 | **Auteurs van de originele test** | Klik of tik om tekst in te voeren. |
| 2.4 | **Auteurs van de lokale aanpassing** | Klik of tik om tekst in te voeren. |
| 2.5 | **Lokale distributeur/uitgever van de test** | Klik of tik om tekst in te voeren. |
| 2.6 | **Uitgever van de originele versie van de test** *(indien verschillend van huidige distributeur/uitgever)* | Klik of tik om tekst in te voeren. |
| 2.7.1 | **Publicatiedatum van de huidige revisie/editie** | Klik of tik om een datum in te voeren. |
| 2.7.2 | **Publicatiedatum van de aanpassing voor lokaal gebruik** | Klik of tik om een datum in te voeren. |
| 2.7.3 | **Publicatiedatum van de originele test** | Klik of tik om een datum in te voeren. |

|  |
| --- |
| **Algemene beschrijving van het instrument** Korte, op zichzelf staande, niet-evaluatieve beschrijving (200-600 woorden)  Hier dient een bondige, niet-evaluatieve beschrijving van het instrument gegeven te worden in neutrale bewoordingen. De beschrijving moet de lezer een duidelijk beeld geven van wat het instrument stelt te zijn – wat het inhoudt, de schalen die het verklaart te meten, etc. Het moet zo neutraal mogelijk van toon zijn. De beschrijving biedt de lezer een duidelijk beeld van wat het instrument is, wat het bevat, de schalen die het meet, het bedoelde gebruik, de beschikbaarheid en het type van normgroepen, de algemene aandachtspunten of opvallende eigenschappen en relevante historische achtergrond. De beschrijving mag relatief kort zijn (200-300 woorden). Echter, voor sommige, meer complexe instrumenten met meerdere schalen kan een langere beschrijving noodzakelijk zijn (300-600 woorden). Deze omschrijving dient zodanig geschreven te zijn dat het op zichzelf kan staan als een beschrijving van het instrument. Bijgevolg kan meer specifieke informatie die werd verstrekt als antwoord op de secties 2-6 hier herhaald worden. Het dient alle versies weer te geven die beschikbaar zijn en waarnaar wordt verwezen in verdere pagina’s.  Dit onderdeel dient te worden beantwoord op basis van informatie voorzien door de uitgever en te worden gecontroleerd op nauwkeurigheid door de beoordelaar. |
| Klik of tik om tekst in te voeren. |

# Classificering

|  |  |  |
| --- | --- | --- |
| 3.1 | **Inhoudsdomeinen** *(gelieve alles te selecteren wat van toepassing is)*  U dient de inhoudsdomeinen te identificeren die door de uitgever zijn gespecifieerd. Indien deze niet duidelijk zijn, dient dit aangegeven te worden en u dient op basis van de voorziene informatie in de handleiding (standaardisatie, steek-proeven, toepassingen, validering, etc.) te oordelen welke antwoorden het meest geschikt zijn voor 3.1. | Vaardigheid - Algemeen  Vaardigheid – Handmatige vaardigheden/ behendigheid  Vaardigheid – Mechanisch  Vaardigheid – Leren/ geheugen  Vaardigheid – non-verbaal/ abstract/ inductief  Vaardigheid - Numeriek  Vaardigheid – Perceptuele snelheid/ controle  Vaardigheid – Sensomotorisch  Vaardigheid – Spatiaal/ visueel  Vaardigheid - Verbaal  Aandacht/concentratie  Overtuigingen  Cognitieve stijlen  Stoornis en pathologie  Familiaal functioneren  Functioneren in groep  Interesses  Motivatie  Organisatorische functie, geaggregeerde maatregelen, klimaat etc.  Persoonlijkheid – Trek  Persoonlijkheid – Type  Persoonlijkheid – Toestand  Levenskwaliteit  Schoolse prestaties (educatieve test)  Schools of educatief functioneren  Situationeel oordeel  Stress/ burn-out  Therapie uitkomst  Waarden  Welzijn  Andere (gelieve deze te omschrijven):  Klik of tik om tekst in te voeren. |
| 3.2 | **Beoogde of voornaamste toepassings-gebied(en)** *(gelieve deze te selecteren die van toepassing zijn)*  U dient de beoogde toepassingsgebieden te identificeren zoals ze zijn gespecificeerd door de uitgever. Geef aan waar deze niet duidelijk zijn en oordeel welke antwoorden het meest geschikt zijn voor 3.2 op basis van de voorziene informatie in de handleiding (standaardisatie, steekproeven, toe-passingen, validering, etc.) | Klinisch  Advies, begeleiding en carrièrekeuze  Educatief  Forensisch  Algemene gezondheid, leven en welzijn  Neurologisch  Sport en vrije tijd  Werk en beroep  Andere (gelieve deze te omschrijven): Klik of tik om tekst in te voeren. |
| 3.3 | **Beschrijving van de populaties waarvoor de test is bedoeld**  Dit onderdeel dient te worden beantwoord op basis van informatie van de uitgever. Voor sommige testen is dit mogelijks zeer algemeen (bv. volwassenen), voor andere kan dit meer specifiek zijn (bv. handarbeiders, of jongens met een leeftijd tussen 10 en 14 jaar). Enkel de opgegeven populaties dienen hier benoemd te worden. Indien deze ongepast lijken, dient dit te worden becommentarieerd in het Evaluatiedeel van de beoordeling. | Klik of tik om tekst in te voeren. |
| 3.4 | **Aantal schalen en korte beschrijving van de variabele(n) die gemeten worden door het instrument.**  Dit onderdeel dient te worden beantwoord op basis van informatie van de uitgever.  Gelieve het aantal schalen aan te geven (indien meer dan één) en geef een korte beschrijving van elke schaal als de betekenis niet duidelijk is af te leiden uit de naam.  Beoordelingen van het instrument dienen een bespreking te bevatten van andere afgeleide scores wanneer deze scores vaak worden gebruikt bij het instrument en worden beschreven in de standaard-documentatie – bv. scores op de primaire trek en ook Big Five scores op de secundaire trekken voor een multitrekpersoonlijk-heidstest, of subtest-, factor- en totaalscores op een intelligentietest. | Klik of tik om tekst in te voeren. |
| 3.5 | **Wijze(n) waarop antwoorden worden geregistreerd** Dit onderdeel dient te worden beantwoord op basis van informatie van de uitgever.  Indien speciale apparatuur (andere dan de aangegeven opties in de lijst, bv. digitaal opnametoestel) noodzakelijk is, gelieve deze dan hier te beschrijven. Bovendien moeten alle mogelijke speciale test-condities worden beschreven. ‘Standaard testcondities’ worden verondersteld beschikbaar te zijn voor de gesurveilleerde en gecontroleerde beoordeling. Deze omvatten een rustige, goed verlichte en geventileerde kamer met voldoende bureau- en zitruimte voor de testleider(s) en kandida(a)t(en) die noodzakelijk aanwezig zijn voor de beoordeling. | Mondeling interview  Papier en potlood  Manuele (fysieke) handelingen  Directe observatie  Geautomatiseerd/computergestuurd  Andere (gelieve deze te omschrijven): Klik of tik om tekst in te voeren. |
| 3.6 | **Vereisten voor de testdeelnemer**  Dit onderdeel dient te worden beantwoord op basis van informatie van de uitgever.  Welke capaciteiten en vaardigheden zijn noodzakelijk voor de testdeelnemer om met de test aan de slag te gaan en om een eerlijke interpretatie van de testscore mogelijk te maken? Het is normaliter wel duidelijk wanneer het volledig ontbreken van een voorwaarde de mogelijkheid om de test af te ronden verhindert (zoals het geven een van papier-en-pen test wanneer iemand blind is), maar de opgesomde voorwaarden dienen als volgt te worden geclassificeerd:   * ‘Irrelevant/niet essentieel’ be-tekent dat deze vaardigheid in zijn geheel niet nodig is – zoals manuele vaardigheden om mondelinge vragen verbaal te kunnen beantwoorden. * ‘Essentiële informatie is aanwezig’ betekent dat de mogelijke mate van beperking wordt vermeld. * ‘Informatie ontbreekt’ betekent dat er mogelijke beperkingen zijn voor testdeelnemers die een specifieke capaciteit of vaardigheid niet hebben (gekend via theorie of empirische resultaten), maar dat dit niet duidelijk is op basis van de informatie voorzien door de testuitgever bv. als de test een taal gebruikt die verschilt van de deelnemers’ moedertaal. | Manuele capaciteiten **(selecteer één)**  Handigheid **(selecteer één)**  Zicht **(selecteer één)**  Gehoor **(selecteer één)**  Beheersing taal van  de test **(selecteer één)**  Lezen **(selecteer één)**  Schrijven **(selecteer één**) |
| 3.7 | **Vraagmodel** *(selecteer er één)*  Dit onderdeel dient te worden beantwoord op basis van informatie van de uitgever.  Er wordt een onderscheid gemaakt tussen twee soorten meerkeuzemodellen. Het eerste model bevat testen waarin de respondent het juiste antwoord dient te selecteren uit een aantal alternatieven, zoals in vaardigheidstesten (bv. een redeneertest met figuurreeksen). Het tweede model betreft vragenlijsten waarin er geen duidelijk juist antwoord is. Dit model vereist dat testdeelnemers een keuze maken tussen sets van twee of meer items gebaseerd op verschillende schalen (bv. schalen uit een beroepsinteresse- of persoonlijkheidsvragenlijst). Dit model is multidimensionaal, omdat de opties tot verschillende schalen of dimensies behoren. In dit geval is het mogelijk dat de stellingen ofwel gerangschikt ofwel beantwoord moeten worden via de selectie van de opties ‘meest’ of ‘minst van toepassing voor mij.’ Dit type kan mogelijk resulteren in ipsatieve schalen (zie vraag 3.8). Bij Likertschaal beoordelingen moet de testdeelnemer ook kiezen uit een aantal keuzemogelijkheden, maar het essentiële verschil met de meerkeuzemodellen is dat de gebruikte schalen hier eendimensionaal zijn (bv. variërend van ‘nooit’ tot ‘altijd’ of van ‘zeer onwaarschijnlijk’ tot ‘zeer waarschijnlijk’) en dat de testdeelnemer niet hoeft te kiezen tussen alternatieven van verschillende dimensies. Een schaal moet ook als Likertschaal aangeduid worden indien er slechts twee keuzeopties zijn binnen één dimensie (bv. ja/nee of altijd/nooit). | Meerkeuze (vaardigheidstesten, of goed/fout)  Aantal keuzemogelijkheden: Klik of tik om tekst in te voeren.  Meerkeuze (keuzemogelijkheden uit verschillende schalen/dimensies) Aantal keuzemogelijkheden: Klik of tik om tekst in te voeren.  Likertschaal Aantal keuzemogelijkheden: Klik of tik om tekst in te voeren.  Open  Andere (gelieve dit te omschrijven): Klik of tik om tekst in te voeren. |
| 3.8 | **Ipsativiteit**  Zoals vermeld in 3.7 *kan* het multidimensionale meerkeuzemodel waarin keuzemogelijkheden uit verschil-lende schalen/dimensies op-genomen zijn (cfr. gemengde schalen), leiden tot ipsatieve scores. Kenmerkend voor ipsatieve scores is dat de score op elke schaal of dimensie ingeperkt wordt door de scores op de andere schalen of dimensies. In instrumenten die volledig ipsatief zijn, is de som van de schaalscores constant voor elke persoon. Andere scoringsprocedures kunnen resulteren in ipsativiteit (bv. door iemands algeheel gemiddelde af te trekken van elk van hun schaalscores). | Ja, multidimensionaal meerkeuze-model met keuzemogelijkheden uit verschillende schalen/dimensies resulteren in gedeeltelijke of volledige ipsatieve scores  Ja, andere vraagmodellen met scoringsprocedures die resulteren in gedeeltelijke of volledige ipsatieve scores  Nee, multidimensionaal meerkeuze-model met keuzemogelijkheden uit verschillende schalen/dimensies leiden NIET tot ipsatieve scores  Niet relevant |
| 3.9 | **Het totale aantal testitems en het aantal items per schaal of subtest**  Dit onderdeel dient te worden beantwoord op basis van informatie van de uitgever.  Indien het instrument meerdere schalen of subtesten heeft, geef dan het totale aantal items aan en het aantal items per schaal of subtest. Indien één of meerdere items op meer dan één schaal of subtest terug te vinden zijn, dient dit gedocumenteerd te worden. | Klik of tik om tekst in te voeren. |
| 3.10 | **Beoogde gebruikswijze (condities waar-onder het instrument werd ontwikkeld en gevalideerd)**  *(selecteer alles wat van toepassing is)*  Dit onderdeel is belangrijk omdat het bepaalt of het instrument ontworpen is met het doel om het te gebruiken in niet-gesuperviseerde of niet-gecontroleerde afnamecondities. Houd er rekening mee dat gebruikswijzen kunnen variëren tussen versies van een instrument. Dit onderdeel dient te worden beantwoord op basis van informatie van de uitgever en gecontroleerd te worden op nauwkeurigheid. Opmerking: de vier gebruikswijzen worden gedefinieerd in de *International Guidelines on Computer Based and Internet Delivered Testing* (International Test Commission, 2005, pp. 5-6) | *Open afname:* er is geen direct menselijk toezicht op de testafname waardoor er dus geen manier is om de identiteit van de testdeelnemer vast te stellen. Testen via internet zonder enige registratievereiste kunnen als voorbeeld van deze afnamewijze worden beschouwd.  *Gecontroleerde afname*: er is geen direct menselijk toezicht op de testafname, maar de test is enkel beschikbaar gemaakt voor vooraf geïdentificeerde en gekende testdeelnemers. Bij testen via het internet is hier vereist dat testdeelnemers een gebruikersnaam en wachtwoord verkrijgen. Deze zijn vaak voor eenmalig gebruik ontworpen.  *Gesuperviseerde afname*: er is een zekere mate van direct menselijk toezicht aanwezig op de omstandigheden van de testafname. Zo kan de identiteit van de testdeelnemer vastgesteld worden. Bij testen via het internet dient een administrator de kandidaten aan te melden en te bevestigen dat de test naar behoren is afgenomen en afgerond.  *Gecontroleerde afname:* er is een hoge mate van menselijk toezicht en controle over de omgeving van de testafname. Bij computergebaseerde testen wordt dit normaal bereikt door het gebruik van aangewezen testcentra, waar er een hoge mate van controle is op toegang, beveiliging, de kwalificatie van testafnamepersoneel en de kwaliteit en technische specificaties van de testapparatuur. |
| 3.11 | **Wijze(n) van afname** *(selecteer alles wat van toepassing is)*  Dit onderdeel dient te worden beantwoord op basis van informatie van de uitgever.  Indien speciale apparatuur (andere dan de aangegeven opties in de lijst, bv. digitaal opnametoestel) noodzakelijk zijn, dienen deze hier beschreven te worden. Bovendien moeten alle mogelijke speciale testcondities worden beschreven. ‘Standaard testcondities’ worden verondersteld beschikbaar te zijn voor gesurveilleerde en gecontroleerde be-oordeling. Deze omvatten een rustige, goed verlichte en geventileerde kamer met voldoende bureau- en zitruimte voor de vereiste testleider(s) en kandida(a)t(en). | Interactieve individuele afname  Groepsafname onder toezicht  Geautomatiseerde afname onder toezicht: lokaal op computer-geïnstalleerde applicatie  Geautomatiseerde afname onder toezicht: internetapplicatie  Geautomatiseerde afname zonder toezicht/zelfbeoordeling: lokaal op computer-geïnstalleerde applicatie  Geautomatiseerde afname zonder toezicht/zelfbeoordeling: internetapplicatie  Andere (gelieve deze te omschrijven): Klik of tik om tekst in te voeren. |
| 3.12 | **Benodigde tijd voor het afnemen van het instrument** *(gelieve te specificeren voor elke afnamewijze)*  Dit onderdeel dient te worden beantwoord op basis van informatie van de uitgever.  Het antwoord op dit onderdeel kan worden opgedeeld in een aantal componenten. In de meeste gevallen is het enkel mogelijk om algemene schattingen te geven in plaats van exacte cijfers. Het doel is om de potentiële testleider een goed idee te geven van de tijdsinvestering die nodig is voor het gebruiken van dit instrument. Reken de tijd die nodig is om het instrument te leren kennen níét mee, maar ga ervan uit dat de gebruiker ervaren en gekwalificeerd is.   * Voorbereidingstijd: dit omvat de tijd die nodig is voor de testleider om een testafnamesessie voor te bereiden en om alle materialen klaar te zetten; toegangs- en aanmeldingstijd voor een online afname. * Afnameduur per sessie: dit omvat enerzijds de tijd die nodig is om alle items af te ronden en anderzijds een schatting van de benodigde tijd voor het geven van instructies, het doornemen van voorbeelditems en het antwoorden op opmerkingen tijdens de nabespreking aan het einde van de sessie. * Scoring: dit omvat de tijd die nodig is om de ruwe scores te verkrijgen. In vele gevallen is dit mogelijk geautomatiseerd. * Analyse: dit omvat de tijd die nodig is om verder te werken op de ruwe scores om andere metingen af te leiden en om een redelijk uitvoerige interpretatie te geven (ervan uitgaande dat u bekend bent met het instrument). Opnieuw, dit kan mogelijk geautomatiseerd gebeuren. * Feedback: dit omvat de tijd die nodig is om feedback voor te bereiden en te geven aan de testdeelnemer en aan andere belanghebbenden.   De tijd die nodig is voor de laatste twee componenten (namelijk analyse en feedback) kan sterk variëren, afhankelijk van de context waarin het instrument gebruikt wordt. Niettemin zijn enige indicaties of opmerkingen behulpzaam. | Voorbereiding:  Klik of tik om tekst in te voeren.  Afname:  Klik of tik om tekst in te voeren.  Scoring:  Klik of tik om tekst in te voeren.  Analyse:  Klik of tik om tekst in te voeren.  Feedback:  Klik of tik om tekst in te voeren. |
| 3.13 | **Geef aan of verschillende versies van het instrument beschikbaar zijn en welke versie(s) aan deze beoordeling wordt (worden) onderworpen.**  Rapporteer of er wel of geen alternatieve versies (echte of pseudo-parallelle versies, korte versies, digitale versies, etc.) beschikbaar zijn en beschrijf de toepasbaarheid van elke versie voor verschillende groepen mensen. In sommige gevallen zijn verschillende versies van een instrument bedoeld om gelijkwaardig aan elkaar te zijn - bv. alternatieve versies. In andere gevallen bestaan er verschillende versies voor verschillende groepen (bv. een versie voor kinderen en een versie voor volwassenen). Indien er meerdere versies bestaan, geef dan aan of deze gelijkwaardig/alternatief zijn aan elkaar, of ze ontworpen zijn voor andere doeleinden – bv. korte en lange versie; ipsatieve en normatieve versie. Beschrijf ook of een gedeelte van de test gebruikt kan worden in plaats van het volledige instrument. Indien digitale versies bestaan, beschrijf dan kort de software- en hardwarevereisten. Merk op dat op zichzelf staande Computer-gebaseerde Tests (CBT) en onlinepakketten, indien beschikbaar, ook aangegeven dienen te worden. | Klik of tik om tekst in te voeren. |

# Meting en scoring

|  |  |  |
| --- | --- | --- |
| 4.1 | **Scoringsprocedure voor de test** *(selecteer alles wat van toepassing is)*  Dit onderdeel dient te worden beantwoord met verwijzing naar de informatie van de uitgever, de handleidingen en documentatie.  Bureaudiensten zijn diensten die verstrekt worden door de aanbieder of een vertegenwoordiger van de aanbieder voor het scoren en interpreteren. In het algemeen zijn deze diensten optioneel. Indien scoring en/of interpretatie ENKEL door een bureaudienst kan worden uitgevoerd, dan moet dit in de beoordeling kenbaar worden gemaakt en de kosten dienen te worden opgenomen als terugkerende kosten. | Computerscoring met directe invoer van antwoorden door de testdeelnemer  Computerscoring door Optical Mark Reader-invoer van antwoorden op het papieren antwoordenformulier  Computerscoring met handmatige invoer van antwoorden van het papieren antwoordenformulier  Eenvoudige handmatige scoringssleutel – enkel administratieve vaardigheden vereist  Complexe handmatige scoring – vereist training in het scoren van het instrument  Bureaudienst – bv. scoring door het bedrijf dat het instrument verkoopt  Andere (gelieve deze te omschrijven): Klik of tik om tekst in te voeren. |
| 4.2 | **Scores**  Dit onderdeel dient te worden ingevuld met verwijzing naar de informatie van de uitgever, de handleidingen en documentatie.  Korte beschrijving van het scoringssysteem om globale en gedeeltelijke scores te verkrijgen (giscorrectie, kwalitatieve interpre-tatiehulpmiddelen, etc.) | Klik of tik om tekst in te voeren. |
| 4.3 | **Gebruikte schalen** *(selecteer alles wat van toepassing is)*  Dit onderdeel dient te worden ingevuld met verwijzing naar de informatie van de uitgever, de handleidingen en documentatie. | *Scores gebaseerd op percentielen*  Percentielen  5-voudige classificatie: 10:20:40:20:10 honderdste splitsing  Decielen  Andere (gelieve deze te omschrijven): Klik of tik om tekst in te voeren.  *Standaardscores*  Z-scores  *IQ deviation quotients* etc. (bv. gemiddelde: 100, SD: 15 voor Wechsler of 16 voor Stanford-Binet)  Stens  Stanines, C-scores  T-scores  Andere (gelieve deze te omschrijven): Klik of tik om tekst in te voeren.  Grensscores, verwachtingstabellen of andere specifieke keuzegeoriënteerde indexen.  Enkel gebruik van ruwe scores  Andere (gelieve deze te omschrijven):  Klik of tik om tekst in te voeren. |
| 4.4 | **Scoretransformatie voor standaardscores** | Genormaliseerd – standaardscores verkregen door het gebruik van een opzoektabel voor normalisatie  Niet-genormaliseerd – standaardscores verkregen door lineaire transformatie  Niet van toepassing |

# Computergegenereerde rapporten

**Let op dat dit onderdeel puur *beschrijvend* is. Evaluaties van de rapporten dienen in het Evaluatiedeel van de beoordeling gegeven te worden.**

Wanneer bijvoorbeeld meerdere gegenereerde rapporten beschikbaar zijn, gelieve de items 5.2 - 5.13 te vervolledigen voor elk rapport of essentieel onderdeel van een rapport (kopieer pagina’s indien noodzakelijk). Dit classificatiesysteem kan gebruikt worden om twee rapporten te beschrijven die door een systeem werden afgeleverd. Bijvoorbeeld, Rapport 1 kan bedoeld zijn voor de testleider of andere ongetrainde gebruikers, en Rapport 2 voor de getrainde testleider die bekwaam is in het gebruik van het instrument en weet hoe dit te interpreteren.

|  |  |  |
| --- | --- | --- |
| 5.1 | **Zijn er computergegenereerde rapporten beschikbaar bij het instrument?**  Als het antwoord op 5.1 ‘JA’ is, dan dient de volgende classificatie te worden gebruikt om de soorten rapporten te classificeren. Voor vele instrumenten zal er een reeks rapporten beschikbaar zijn. Gelieve een afzonderlijk formulier in te vullen voor elk rapport. | Ja (vul onderstaande onderdelen in)  Nee (ga verder naar 6.1) |
| 5.2 | **Naam of beschrijving van het rapport**  *(cfr. de introductie van dit onderdeel)* | Klik of tik om tekst in te voeren. |
| 5.3 | **Media** *(selecteer alles wat van toepassing is)*  Rapporten kunnen ofwel volledig uit tekst bestaan ofwel naast tekst ook de scores grafisch of in tabelvorm weergeven (bv. stenprofielen). Waar zowel tekst als scores worden gepresenteerd, kunnen deze ofwel op parallelle wijze, ofwel gekoppeld aan elkaar worden weergegeven zodat de relatie tussen tekst en scores expliciet wordt gemaakt. | Enkel tekst  Geen verband tussen tekst en grafische elementen  Tekst en grafische elementen zijn geïntegreerd  Enkel grafische elementen |
| 5.4 | **Complexiteit** *(selecteer er één)*  Sommige rapporten zijn zeer eenvoudig, bijvoorbeeld rapporten waarin elke stenscore vervangen wordt door een tekstuele beschrijving per schaal.  Andere rapporten zijn complexer. In dergelijke rapporten zijn de tekstuele beschrijvingen gerelateerd aan patronen of configuraties van schaalscores en worden interactie-effecten tussen schalen beschouwd. | Eenvoudig (bevat een aantal alinea’s met schaalbeschrijvingen)  Gemiddeld (bevat een verzameling van eenvoudige beschrijvingen en enkele beschrijvingen van configuraties)  Complex (bevat beschrijvingen van patronen en configuraties van schaalscores en interacties tussen schalen) |
| 5.5 | **Structuur van het rapport** *(selecteer er één)*  De structuur is gerelateerd aan de complexiteit van het rapport. | Op schaal gebaseerd: het rapport is rond individuele schalen opgebouwd  Op factor gebaseerd: het rapport is rond hogere-orde factoren opgebouwd, zoals de ‘Big Five’-persoonlijkheidskenmerken  Op construct gebaseerd: het rapport is rond één of meer reeksen van constructen opgebouwd (bv. deze constructen kunnen in een werkcontext teamtypes, leiderschapsstijlen of stresstolerantie zijn; of in een klinische context kunnen dit verschillende soorten psychopathologie zijn; etc.) die gelinkt zijn aan de originele schaalscores.  Op criterium gebaseerd: het rapport richt zich op verbanden met empirische uitkomsten (bv. schoolprestatie, uitkomst van therapie, werkprestatie, absenteïsme, etc.)  Andere (gelieve deze te omschrijven):  Klik of tik om tekst in te voeren. |
| 5.6 | **Gevoeligheid voor context***(selecteer er één)*  Wanneer het rapport wordt geschreven, dan wordt de taal, vorm en inhoud van het rapport aangepast aan de lezer van het rapport en wordt er eveneens rekening gehouden met het doel van de testafname en de context waarin deze plaatsvindt. In een werk- en organisatiecontext zal een rapport dat gemaakt is voor selectiedoeleinden anders zijn dan een rapport dat geschreven is voor begeleiding of ontwikkeling; een rapport voor een manager van middelbare leeftijd zal anders zijn dan een rapport geschreven voor een jong persoon die net begint met een trainingsprogramma, etc. In een onderwijscontext zal een rapport ter evaluatie van de algemene vaardigheid van een student om te leren en te functioneren in een leeromgeving anders zijn dan een rapport om te beoordelen of een student al dan niet een specifieke leerstoornis heeft. Een rapport gericht aan andere professionelen over leerdoelen en interventies zal verschillen van een rapport dat gericht is aan ouders om ze te informeren over de sterktes en zwaktes van hun kind. In een klinische context zal een rapport gemaakt voor diagnostische doeleinden verschillen van een rapport dat potentieel risicovol gedrag van een patiënt evalueert. Een rapport dat als doel heeft om feedback te geven aan patiënten zal verschillend zijn van een rapport dat gemaakt is met het doel om de autoriteiten te informeren of het al dan niet veilig is om een patiënt te ontslaan van een onvrijwillige behandeling. | Eén versie voor alle contexten  Vooraf gedefinieerde context-gerelateerde versies; Aantal contexten: Klik of tik om tekst in te voeren.  Door de gebruiker definieerbare contexten en bewerkbare rapporten. |
| 5.7 | **Klinisch-actuarieel** *(selecteer er één)*  De meeste rapportsystemen zijn gebaseerd op een klinisch oordeel. I.e., de teksteenheden zullen geschreven zijn door één of meerdere personen die ‘expert-testleider’ zijn van het instrument in kwestie. De geschreven tekst in dergelijke rapporten geeft daarom hun particuliere interpretatie van de schalen weer. Sommige rapportsystemen nemen actuariële rapporten op waar de beweringen gebaseerd zijn op empirische validatie-studies die het verband leggen tussen schaalscores en bijvoorbeeld metingen van werkprestatie, klinische classificatie, etc. | Gebaseerd op het klinisch oordeel van één expert  Gebaseerd op het klinisch oordeel van een groep experten  Gebaseerd op empirische/actuariële verbanden |
| 5.8 | **Modificeerbaarheid** *(selecteer er één)*  De output van een rapport is vaak vaststaand. Sommige systemen produceren echter output in de vorm van een bestand dat kan worden verwerkt door de gebruiker. Nog andere systemen bieden online interactieve toegang aan zowel de eindgebruiker als de testafnemer. | Geen mogelijkheid tot aanpassing (vaste geprinte output)  Beperkte mogelijkheid tot aanpassing (enkel binnen bepaalde domeinen, bv. biodata velden)  Onbeperkte mogelijkheid tot aanpassing (bv. door toegang tot een Word-document dat bewerkt kan worden)  Interactief rapport dat de testdeelnemer de mogelijkheid geeft om enerzijds opmerkingen in te voegen of om anderzijds de inhoudelijke nauw-keurigheid te beoordelen (bv. door gedeelde online toegang tot een interactief rapporteringsmechanisme) |
| 5.9 | **Mate van afwerking** *(selecteer er één)*  De mate waarin het systeem ontworpen is om een geïntegreerde tekst te genereren – in de vorm van een gebruiksklaar rapport – of een verzameling ‘aantekeningen’, op-merkingen, hypotheses, etc. | Publicatiekwaliteit  Kwaliteit van een ruwe versie |
| 5.10 | **Transparantie** *(selecteer er één)*  Systemen verschillen in hun openheid of transparantie naar de gebruiker. Bij een open systeem is het verband tussen de schaalscore en de tekst duidelijk en ondubbelzinnig. Dergelijke openheid is alleen mogelijk indien de tekst en de scores weergegeven worden en de verbanden tussen de twee expliciet worden gemaakt. Andere systemen werken als ‘zwarte dozen’ waardoor het voor de testleider moeilijk wordt om de schaalscores te verbinden aan de tekst. | Duidelijk verband tussen constructen, scores en tekst  Verborgen verband tussen constructen, scores en tekst  Mengeling van duidelijk/verborgen ver-banden tussen constructen, scores en tekst |
| 5.11 | **Stijl en toon** (selecteer er één)  Systemen verschillen ook in de mate waarin ze de lezer van het rapport begeleiding of sturing aanbieden. In een werk- en organisatiecontext is een bewering als ‘’Dhr. X is zeer verlegen en zal geen goede verkoper zijn…’’ stipulatief, terwijl andere beweringen ontworpen zijn om hypotheses te suggereren of om vragen op te roepen, zoals “Op basis van zijn scores op schaal Y, lijkt dhr. X in vergelijking met een referentiegroep van verkopers zeer verlegen te zijn. Als dit het geval is, dan zou hij het moeilijk kunnen vinden om in een verkoopomgeving te werken. Dit dient verder met hem verkend te worden.”  In een onderwijscontext kan een stipulatieve bewering zijn: “De resultaten laten zien dat X’s wiskundige vaardigheden twee jaar achterliggen op die van zijn leeftijdsgenoten”, terwijl een bewering ontworpen om hypotheses te suggereren als volgt kan zijn: “De resultaten wijzen erop dat X makkelijk afgeleid wordt door externe stimuli tijdens het uitvoeren van taken. Gedragsobservaties tijdens het testen ondersteunen dit. Dit zou overwogen moeten worden wanneer een optimale leeromgeving voor X wordt ontworpen”.  In een klinische context kan een stipulatieve bewering zijn: “Testscores tonen dat de patiënt een ernstige vorm van visuele verwaarlozing heeft, en daarom niet in staat is om een motorvoertuig veilig te besturen”, terwijl een bewering ontworpen om hypotheses te suggereren als volgt kan zijn: “De testscores van mevr. X laten zien dat ze mogelijks problemen heeft met het opbouwen van stabiele emotionele relaties. Hier is verdere verkenning noodzakelijk vooraleer een conclusie over een diagnose kan worden genomen”. | Directief/ stipulatief  Begeleiding/suggereert hypothesen  Andere (gelieve deze te omschrijven): Klik of tik om tekst in te voeren. |
| 5.12 | **Beoogde ontvangers** *(selecteer alles wat van toepassing is)*  Rapporten worden in het algemeen ontworpen om aan de behoeften van één of meer gebruikerscategorieën te voldoen. Gebruikers kunnen onder-verdeeld worden in vier hoofdgroepen:   1. *Gekwalificeerde testleiders.* Dit zijn mensen die beschikken over voldoende kennis en vaardigheden om hun eigen rapporten te produceren op basis van schaalscores. Ze zouden gebruik moeten kunnen maken van rapporten die psychometrische termino-logie gebruiken en ze moeten expliciete verbanden kunnen leggen tussen schalen en beschrijvingen. Ze zouden ook rapporten moeten kunnen wijzigen en aanpassen. 2. *Gekwalificeerde systeemgebruikers*. Hoewel ze niet competent zijn om hun eigen rapporten te kunnen genereren op basis van een stel schaalscores, zijn de mensen in deze groep wel bekwaam om de gegenereerde output van het systeem te gebruiken. De mate van training die noodzakelijk is om deze competentie te bereiken zal aanzienlijk verschillen, afhankelijk van de aard van de computergegenereerde rapporten (bv. eigenschap-gebaseerd versus competentie-gebaseerd, eenvoudig of complex) en waarvoor de rapporten gebruikt zullen worden (hoge inzet versus lage inzet) 3. *Testdeelnemers*. De persoon die het instrument afneemt, heeft normaal geen voorafgaande kennis van het instrument en het type rapport dat geproduceerd wordt door het systeem. Rapporten gericht aan hen zullen in een taal moeten zijn die geen psychometrische kennis of kennis over instrumenten veronderstelt. 4. *Derde partijen*. Dit betreft mensen, andere dan de kandidaat zelf, voor wie de informatie in het rapport relevant is omwille van hun (directe) betrokkenheid of die een kopie van het rapport kunnen ontvangen. Dit kunnen potentiële werkgevers zijn, iemands manager, leidinggevende of de ouder van een jonge persoon die carrière-advies ontvangt. Het taalniveau dat nodig is voor mensen in deze categorie is gelijkaardig aan dat wat nodig is voor rapporten bedoeld voor testdeelnemers. | Gekwalificeerde testleiders  Gekwalificeerde systeemgebruikers  Testdeelnemers  Derde partijen |
| 5.13 | **Verschaffen de verdelers een dienst om op maat gemaakte geauto-matiseerde rapporten te wijzigen en/of uit te werken?** *(selecteer er één)* | Ja  Nee |

# Voorwaarden en kosten van het aanbod

Dit deel definieert wat de uitgever zal voorzien, aan wie, onder welke voorwaarden en tegen welke kosten. Het definieert de voorwaarden opgelegd door de leverancier over wie de instrumentenmaterialen wel of niet mag verkrijgen. Indien een van de opties niet aansluit bij de aanbodvoorwaarden, geef dan een beschrijving van de relevante voorwaarden.

|  |  |  |
| --- | --- | --- |
| 6.1 | **Documentatie voorzien door de verdeler als onderdeel van het testpakket** *(selecteer alles wat van toepassing is)* | Gebruikershandleiding  Technische (psychometrische) hand-leiding  Aanvullende technische informatie en updates (bv. lokale normen, lokale valideringsstudies, etc.)  Gerelateerde boeken en artikels |
| 6.2 | **Publicatiemethoden** *(selecteer alles wat van toepassing is)*  Bijvoorbeeld, technische handleidingen kunnen up-to-date gehouden en beschikbaar gesteld worden om te downloaden via het internet, terwijl gebruikshandleidingen voorzien worden in papiervorm of op CD/DVD. | Papier  CD of DVD  Download via internet  Andere (gelieve deze te omschrijven): Klik of tik om tekst in te voeren. |
| De onderdelen 6.3 – 6.5 gaan over kosten. Deze informatie zal waarschijnlijk snel achterhaald zijn. Daarom wordt het aanbevolen om de aanbieder of uitgever zo dicht mogelijk bij de publicatiedatum van de beoordeling te contacteren, om actuele informatie over deze onderdelen te voorzien. | | |
| 6.3.1 | **Opstartkosten**  Prijs van een complete materialen set (alle handleidingen en ander materiaal nodig voor tenminste één testafname). Specifieer hoeveel testdeelnemers kunnen worden beoordeeld met de verkregen materialen voor de opstartkosten en of deze kosten materialen voor herhaaldelijke beoordelingen bevatten.  In dit onderdeel wordt getracht de ‘set-up’-kosten te identificeren, namelijk de kosten met betrekking tot het bekomen van een volledige set van materialen, scoringssleutels, etc. Enkel wanneer het instrument ‘gesloten’ is, zal bij de ‘set-up’-kosten ook de trainingskosten worden gerekend. Bij ‘gesloten’ instrumenten is een specifieke trainingskost onvermijdelijk, ongeacht het aanwezige kwalificatieniveau van de testleider. In dergelijke gevallen moet het trainingselement in de kosten expliciet worden gemaakt. De initiële kosten omvatten GEEN apparatuur voor algemeen gebruik (zoals computers, dvd-spelers, etc.). Toch dient de nood hieraan te worden vermeld. Definieer in het algemeen: elke speciale trainingskost; kosten van de gebruikershandleiding; technische handleiding(en); proefexemplaar of referentieset van materialen; initiële software kosten, etc. | Klik of tik om tekst in te voeren. |
| 6.3.2 | **Terugkerende kosten**  Specifieer waar nodig terugkomende kosten met betrekking tot afname en scoring, en kosten voor interpretatie (zie 6.4.1 -6.5)  Dit onderdeel heeft betrekking op de lopende gebruikskosten van het instrument. Het dient de kosten te geven voor instrumentmaterialen (antwoordformulieren, niet-herbruik-bare of herbruikbare vragenboekjes, profielformulieren, authentificatie-codes voor computer-gebruik of ‘dongle’-eenheden, etc.) per persoon per afname. Let op dat voor papier-gebaseerde afname in de meeste gevallen dergelijke materialen niet individueel beschikbaar zijn maar doorgaans voorzien worden in pakken van 10, 25 of 50.  Maak een opdeling tussen iedere jaarlijkse of per capita licentie-vergoeding (inclusief *release*-codes van de software waar relevant), kosten voor aankoop of het leasen van herbruikbare materialen en kandidaat-specifieke kosten voor niet-herbruikbare materialen. | Klik of tik om tekst in te voeren. |
| 6.4.1 | **Prijzen voor rapporten gegenereerd door software geïnstalleerd door de gebruiker (testleider)** | Klik of tik om tekst in te voeren. |
| 6.4.2 | **Prijzen voor rapporten gegenereerd door post/fax-bureaudienst** | Klik of tik om tekst in te voeren. |
| 6.4.3 | **Prijzen voor rapporten via internetdienst** | Klik of tik om tekst in te voeren. |
| 6.5 | **Prijzen voor andere bureaudiensten: verbetering of ontwikkeling van automatische rapporten** | Klik of tik om tekst in te voeren. |
| 6.6 | **Test-gerelateerde kwalificaties vereist door de leverancier van de test** *(selecteer alles wat van toepassing is)*  Dit onderdeel betreft de test-leiderskwalificaties vereist door de leverancier. Als de uitgever informatie over gebruikerskwalificaties heeft voorzien, dient dit in dit onderdeel te worden genoteerd onder de gegeven categorieën. Zijn de kwalificatie-vereisten niet duidelijk, dan dient dit te worden opgenomen bij ‘Andere’ en *niet* bij ‘Geen’. ‘Geen’ verwijst expliciet naar het ontbreken van de kwalificatie-vereisten.  Voor details over de ‘EFPA Niveau 2’-standaard, consulteer de laatste versie hiervan op de website van EFPA. | Geen  Testspecifieke accreditatie  Accreditatie in algemene prestatietesten: metingen van maximaal haalbare prestatie (gelijkwaardig aan EFPA Niveau 2)  Accreditatie in algemene vaardigheids- en geschiktheidstesten: metingen van maximale prestatie in relatie tot potentiële haalbaarheid (gelijkwaardig aan EFPA Niveau 2)  Accreditatie in algemene persoonlijkheid en beoordeling: metingen van typisch gedrag, attitudes en voorkeuren (gelijkwaardig aan EFPA Niveau 2)  Andere (gelieve deze te omschrijven): Klik of tik om tekst in te voeren. |
| 6.7 | **Professionele kwalificaties vereist voor het gebruik van het instrument** *(selecteer alles wat van toepassing is)*  Dit onderdeel betreft de test-leiderskwalificaties vereist door de uitgever. Als de uitgever informatie over gebruikerskwalificaties heeft voorzien, dient dit in dit onderdeel te worden genoteerd onder de gegeven categorieën. Zijn de kwalificatie-vereisten niet duidelijk, dan dient dit te worden opgenomen bij ‘Andere’ en *niet* bij ‘Geen’. ‘Geen’ verwijst expliciet naar het ontbreken van de kwalificatie-vereisten.  Voor details over de EFPA-gebruikersstandaard, consulteer de laatste versie op de EFPA-website. | Geen  Praktiserend psycholoog met kwalificatie in het relevante toepassingsdomein  Praktiserend psycholoog  Onderzoekspsycholoog  Niet-psychologisch academisch onder-zoeker  Beoefenaar in relevante gerelateerde beroepen (therapie, geneeskunde, begeleiding, onderwijs, personeelszaken, etc.) Gelieve deze te omschrijven:  Klik of tik om tekst in te voeren.  *EFPA Test User Qualification Level 1* of nationaal equivalent *EFPA Test User Qualification Level 2* of nationaal equivalent  Gespecialiseerde kwalificatie equivalent aan de *EFPA Test User Standard Level 3*  Andere (gelieve deze te omschrijven):  Klik of tik om tekst in te voeren. |

DEEL 2: EVALUATIE VAN HET INSTRUMENT

**Informatiebronnen**

Er zijn potentieel vier informatiebronnen die geconsulteerd kunnen worden voor de uitvoering van deze evaluatie:

1. De handleiding en/of rapporten die door de uitgever aangeleverd worden voor de testleider: Deze worden altijd aangeleverd door de uitgever/distributeur voordat de beoordelende organisatie een reviewopdracht over het instrument accepteert. De handleiding en/of rapporten vormen de basismaterialen voor de beoordeling.
2. Open informatie die beschikbaar is in de academische of andere literatuur:  
   Open informatie is meestal afkomstig van de beoordelaar en de beoordelaar kan deze gebruiken bij de beoordeling. Het instrument kan geëvalueerd worden met (of zonder) verwijzingen naar deze informatie in de handleiding.
3. Informatie die door de uitgever niet formeel wordt gepubliceerd of verdeeld:   
   De distributeur/uitgever maakt deze informatie beschikbaar bij de aanvang of stuurt het op wanneer de beoordeling teruggestuurd wordt naar de uitgever voor controle op feitelijke nauwkeurigheid. De beoordelaar dient gebruik te maken van deze informatie, maar geeft aan het begin van de opmerkingen over de technische informatie zeer duidelijk aan dat ‘’de classificatie gemarkeerd met een asterisk (\*) in deze beoordeling verwijst naar informatie van de uitgever/distributeur die (normaal) niet worden verleend aan testleiders’’. Als het gaat om waardevolle informatie, dient in de globale evaluatie een aanbeveling opgenomen te worden dat de uitgever deze informatie publiceert en/of beschikbaar stelt voor de kopers van de test.
4. Informatie die op een commercieel vertrouwelijke basis stoelt:   
   In sommige gevallen kunnen uitgevers beschikken over technisch belangrijke informatie die ze niet publiek willen maken omwille van commerciële redenen. De bescherming van de intellectuele eigendom van testontwikkelaars is in de praktijk zeer beperkt (het auteursrecht is zowat de enige wettelijke regeling). Dergelijke informatie kan bijvoorbeeld ook rapporten bevatten over de ontwikkeling van bepaalde scoringsalgoritmes, procedures om testen of items te genereren en technologie om rapporten voort te brengen. Daar waar de inhoud van dergelijke rapporten mogelijks interessant kunnen zijn voor het komen tot een oordeel tijdens een beoordeling, kan de organisatie of vereniging die verantwoordelijk is voor de beoordeling een geheimhoudingsverklaring aanvragen bij de uitgever. Deze overeenkomst zou dan bindend zijn voor de beoordelaars en redacteur. De beoordelaar kan de informatie vervolgens evalueren en opmerkingen plaatsen bij de technische aspecten en de globale evaluatie zodanig dat ‘’de classificatie gemarkeerd met een asterisk (\*) in deze beoordeling verwijst naar informatie van de uitgever/distributeur die onderzocht werd door de beoordelaars op een commercieel vertrouwelijke basis. Deze informatie wordt niet verleend aan eindgebruikers.’’

**Uitleg van de classificaties**

Alle onderdelen worden gescoord op basis van het volgende classificatiesysteem (zie tabel op de volgende pagina). Er zijn gedetailleerde beschrijvingen voorzien die houvast bieden bij elke classificatie/beoordeling.

Waar een [0] of [1] beoordeling wordt gegeven aan een eigenschap die als *kritiek* wordt beschouwd voor een veilig gebruik van het instrument, zal in de beoordeling aanbevolen worden dat het instrument alleen gebruikt mag worden in uitzonderlijke omstandigheden door zeer vakkundige experts of in onderzoek.

De beoordeling van het instrument dient aan te geven welke de kritieke technische eigenschappen zijn, gegeven de aard van het instrument en het bedoelde gebruik ervan. Er wordt voorgesteld om beoordelingen van deze kritische eigenschappen conventioneel in het vetgedrukt weer te geven. In de volgende secties worden de algemene beoordelingen van de sufficiëntie of toereikendheid van informatie met betrekking tot validiteit, betrouwbaarheid en normen, standaard vetgedrukt weergegeven.

***Elk instrument met één of meer beoordelingen van [0] en [1] betreffende eigenschappen die als kritiek gezien worden voor het veilig gebruik van dat instrument, voldoet niet aan de minimumeisen.***

|  |  |
| --- | --- |
| **Beoordeling** | **Uitleg\*** |
| [n.v.t.] | Deze eigenschap is niet van toepassing op het instrument |
| 0 | Niet mogelijk om te beoordelen vermits geen of onvoldoende informatie is voorzien |
| 1 | Onvoldoende |
| 2 | Voldoende |
| 3 | Goed |
| 4 | Uitstekend |

\*Een vijfpuntenschaal is gedefinieerd door EFPA, maar elke gebruiker kan de punten van de schaal combineren (bijvoorbeeld door punten 3 en 4 te combineren in één enkel punt). De enige restrictie is dat er een onderscheid dient te worden gemaakt tussen onvoldoende (of erger) aan de ene kant en voldoende (of beter) aan de andere kant. Beschrijvende termen of symbolen zoals sterren of smileys kunnen gebruikt worden in plaats van nummers. Daar waar de vijfpuntenschaal wordt vervangen of aangepast, dient de gebruiker een sleutel te voorzien die de punten en naamgeving verbindt met de vijfpuntenschaal van EFPA.

# Kwaliteit van de uitleg van de basisprincipes, de presentatie en de beschikbare informatie

In dit onderdeel moeten een aantal classificaties/beoordelingen worden gegeven over diverse aspecten of eigenschappen van de documentatie die bij het instrument (of pakket) geleverd wordt. De term ‘documentatie’ omvat alle materialen die met het instrument meegeleverd worden of gemakkelijk beschikbaar zijn voor de gekwalificeerde gebruiker: bv. de gebruikershandleiding; technische handboeken; normenboekjes; aanvullingen bij de handleiding; updates van uitgevers/leveranciers, etc.

Leveranciers worden verondersteld om voor elke beoordelaar een complete set van dergelijk materiaal te voorzien. Indien u denkt dat er iets is waarmee testleiders voorzien worden dat niet omvat wordt in de informatie die u toegestuurd werd ter beoordeling, gelieve uw redacteur te contacteren.

## Kwaliteit van de uitleg van de basisprincipes

Indien het instrument een *computer-adaptieve test* (CAT)[[4]](#footnote-5) is, dient bijzondere aandacht geschonken te worden aan de items 7.1.1 tot 7.1.6.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Te beoordelen onderdelen: n.v.t. of 0 tot 4** | | **Beoordeling** | | | | | |
| 7.1.1 | Theoretische onderbouwing van de constructen | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 7.1.2 | Procedure van testontwikkeling (en/of vertaling of aanpassing) | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 7.1.3 | Nauwkeurigheid van de analyses van items en van het analysemodel van items | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 7.1.4 | Presentatie van inhoudsvaliditeit | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 7.1.5 | Samenvatting van relevant onderzoek | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 7.1.6 | **Algemene beoordeling van de kwaliteit van de uitleg van de basisprincipes**  Deze algemene beoordeling wordt verkregen door zich te baseren op de classificaties/beoordelingen gegeven op onderdelen 7.1.1 – 7.1.5 | n.v.t. | 0 | 1 | 2 | 3 | 4 |

## Toereikendheid van de voor de testleider beschikbare documentatie

**(gebruikers- en technische handleidingen, aanvullingen bij de normen, etc.)**

Hier ligt de focus op de kwaliteit van de beschikbare documentatie voor gekwalificeerde testleiders: in hoeverre deze sufficiënt en compleet is. Let op, in de sub-sectie 7.2. wordt specifiek ingezoomd op de duidelijkheid en de helderheid van de hoeveelheid aan meegegeven informatie en de toelichting in de beschikbare documentatie voor de testleider (gebruikers- en technische handleidingen, aanvullingen bij de normen, etc.). In de onderdelen 7.1, 7.3, 9, 10 en 11 wordt uitgeweid over de kwaliteit van het instrument zoals gestaafd door de documentatie.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Te beoordelen onderdelen: n.v.t. of 0 tot 4, ‘benchmarks’ of ijkpunten worden voorzien voor een ‘uitstekende’ (4) beoordeling.** | | **Beoordeling** | | | | | |
| 7.2.1 | Basisprincipes (zie beoordeling 7.1.6)  Uitstekend: Logische en duidelijke omschrijving van wat het instrument bedoelt te meten en waarom het is samengesteld zoals het is. | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 7.2.2.1 | Ontwikkeling  Uitstekend: Alle details over itembronnen, ontwikkeling van stimulus materiaal volgens de geaccepteerde richtlijnen (e.g. Haladyna, Downing & Rodriguez, 2002; Moreno, Martinez & Muñiz, 2006), sturing, itemanalyses, vergelijkingsstudies en gemaakte aanpassingen tijdens ontwikkelingsstudies. | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 7.2.2.2 | Ontwikkeling van de test door vertaling/aanpassing  Uitstekend: Informatie in de handleiding die laat zien dat het proces van vertaling/aanpassing gedaan werd volgens de internationale richtlijnen (ITC, 2000) en de volgende onderdelen bevat:   * Bij vertaling naar nieuwe taal: input van *native speakers* van die nieuwe taal * Meervoudige beoordelingen door zowel taal- als inhoudelijke (van de test) deskundigen. * Hervertaling (*back translation*) van nieuwe taal naar oorspronkelijke taal * Aandacht voor culturele en linguïstische verschillen | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 7.2.3 | Standaardisatie  Uitstekend: Duidelijke en gedetailleerde informatie over de omvang (grootte) en de bronnen van de gestandaardiseerde steekproef en procedure. | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 7.2.4 | Normen  Uitstekend: Duidelijke en gedetailleerde informatie over de omvang (grootte) en de bronnen van normgroepen, representativiteit, omstandigheden van de afname, etc. | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 7.2.5 | Betrouwbaarheid  Uitstekend: Uitstekende uitleg van de betrouwbaarheid en standaardmeetfout (SEM – *Standard Error of Measurement*), en een uitgebreide range aan metingen van de interne consistentie, temporele stabiliteit en/of interbeoordelaars-betrouwbaarheid en de resulterende SEM’s voorzien van uitleg over hun relevantie, en de generaliseerbaarheid van het instrument. | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 7.2.6 | Begripsvaliditeit  Uitstekend: Uitstekende uitleg van begripsvaliditeit met een breed scala aan helder, open en degelijk omschreven studies. | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 7.2.7 | Criteriumvaliditeit  Uitstekend: Uitstekende uitleg van criteriumvaliditeit met een breed scala aan helder, open en degelijk omschreven studies. | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 7.2.8 | Computergegenereerde rapporten  Uitstekend: Duidelijke en gedetailleerde informatie over het formaat, de reikwijdte, betrouwbaarheid en validiteit van computergegenereerde rapporten. | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 7.2.9 | **Toereikendheid van de voor de testleider beschikbare documentatie (gebruikers- en technische handleidingen, aanvullingen bij de normen, etc.)**  Deze beoordeling wordt verkregen door een oordeel te vellen op basis van de beoordelingen gegeven op onderdelen 7.2.1 – 7.2.8 | n.v.t. | 0 | 1 | 2 | 3 | 4 |

## Kwaliteit van de procedurele instructies voorzien voor de testleider

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Te beoordelen onderdelen: n.v.t. of 0 tot 4, ‘benchmarks’ of ijkpunten worden voorzien voor een ‘uitstekende’ [4] beoordeling.** | | **Beoordeling** | | | | | |
| 7.3.1 | Voor testafname  Uitstekend: Duidelijke en gedetailleerde uitleg en een systematisch overzicht van procedurele richtlijnen, met een gedegen en gedetailleerd advies over hoe omgegaan kan worden met vragen van kandidaten en problematische situaties. | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 7.3.2 | Voor testscoring  Uitstekend: Duidelijke en gedetailleerde informatie, met inbegrip van een beschrijving van manieren waarop omgegaan kan worden met mogelijke fouten in de scoring.  Indien de scoring met de computer gedaan wordt, is er bewijs dat de scoring correct gedaan wordt? | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 7.3.3 | Voor normering  Uitstekend: Duidelijke en gedetailleerde informatie, met inbegrip van een beschrijving van manieren waarop omgegaan kan worden met mogelijke fouten in normering.  Indien normering gedaan wordt door de computer, is er bewijs dat de omzetting van de score correct is en dat de juiste normgroep wordt toegepast? | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 7.3.4 | Voor interpretatie en rapportering  Uitstekend: Gedetailleerd advies over het interpreteren van verschillende scores, het begrijpen van normatieve metingen en het omgaan met verbanden tussen verschillende schalen, met illustrerende voorbeelden en case studies; daarnaast ook advies over hoe omgegaan kan worden met de mogelijke invloed van inconsistenties in de antwoorden, verschillende antwoordstijlen, antwoorden veinzen, etc. | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 7.3.5 | Voor het verlenen van feedback aan en het nabespreken met testdeelnemers en anderen  Uitstekend: Gedetailleerd advies over hoe feedback verleend kan worden aan de kandidaten, inclusief het gebruik van computergegenereerde rapporten (indien beschikbaar). | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 7.3.6 | Voor het voorzien van ‘good practice’ over kwesties rond openheid, degelijkheid en bias  Uitstekend: Gedetailleerde informatie over studies die handelen over gender en etnische bias, met relevante waarschuwingen over gebruik en generalisatie van validiteit. | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 7.3.7 | Beperkingen op het gebruik  Uitstekend: Duidelijke omschrijving van wie wel en wie niet gemeten zou moeten worden, met een duidelijke rechtvaardiging voor de beperkingen (bv. types van handicap, het vereiste niveau van geletterdheid, etc.) | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 7.3.8 | Software en technische ondersteuning  Uitstekend: In het geval van *Computer Based Testing* (CBT) of *Web Based Testing* (WBT): De informatie met betrekking tot de vereisten voor de browser, de installatie van alle vereiste computersoftware en het gebruik van de software is volledig (het beschrijft ook mogelijke fouten en verschillende systemen). Verder is de beschikbaarheid van technische ondersteuning eveneens duidelijk omschreven. | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 7.3.9 | Referenties en ondersteunend materiaal  Uitstekend: Gedetailleerde referenties naar de relevante ondersteunende academische literatuur en kruisverwijzingen naar andere gerelateerde instrumenten en materialen. | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 7.3.10 | **Kwaliteit van de procedurele instructies die voorzien zijn voor de testleider**    Deze algemene beoordeling wordt verkregen door zich te baseren op de beoordelingen gegeven op onderdelen 7.3.1 – 7.3.9 | n.v.t. | 0 | 1 | 2 | 3 | 4 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | |
| 7.4 | **Algemene toereikendheid**  Deze algemene beoordeling voor onderdeel 7 wordt verkregen door zich te baseren op de beoordelingen gegeven op subonderdelen 7.1, 7.2 en 7.3. | n.v.t. | 0 | 1 | 2 | 3 | 4 |

|  |
| --- |
| **Opmerkingen van de beoordelaar over de documentatie:** (becommentarieer de basisprincipes, presentatie en voorziene informatie) |
| Klik of tik om tekst in te voeren. |

# Kwaliteit van het testmateriaal

## Kwaliteit van het testmateriaal van pen-en-papier-testen

*(dit gedeelte kan overgeslagen worden indien niet van toepassing)*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Te beoordelen onderdelen: n.v.t. of 0 tot 4** | | **Beoordeling** | | | | | |
| 8.1.1 | Algemene kwaliteit van het testmateriaal (testboekjes, antwoordformulieren,  testobjecten, etc.) | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 8.1.2 | Het gemak waarmee de testdeelnemer de taak kan begrijpen | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 8.1.3 | Duidelijkheid en volledigheid van de instructie (inclusief voorbeelden en oefentesten) voor de testdeelnemer | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 8.1.4 | Gemak waarmee reacties of antwoorden gegeven kunnen worden door de testdeelnemer | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 8.1.5 | Kwaliteit van de formulering van de items en duidelijkheid van de grafische inhoud bij non-verbale items | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 8.1.6 | **Kwaliteit van de testmaterialen van pen-en-papier-testen**  Deze algehele beoordeling wordt verkregen door een oordeel te vellen op basis van de beoordelingen gegeven op onderdelen 8.1.1 -8.1.5 | n.v.t. | 0 | 1 | 2 | 3 | 4 |

## Kwaliteit van de testmaterialen van CBT- en WBT-testen

*(dit gedeelte kan overgeslagen worden indien niet van toepassing)*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Te beoordelen onderdelen: n.v.t. of 0 tot 4** | | **Beoordeling** | | | | | |
| 8.2.1 | Kwaliteit van het ontwerp van de software (bv. de robuustheid van de software met betrekking tot de bediening bij een verkeerd gebruik of wanneer onjuiste toetsen worden ingedrukt, bij een zwakke of uitvallende internetverbinding, etc.) | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 8.2.2 | Gemak waarmee de testdeelnemer de taak kan begrijpen | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 8.2.3 | Duidelijkheid en volledigheid van de instructies (inclusief voorbeelditems en oefentesten) voor de testdeelnemer, het gebruik van de software en hoe te antwoorden indien de test wordt afgenomen per computer | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 8.2.4 | Gemak waarmee reacties of antwoorden gegeven kunnen worden door de testdeelnemer | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 8.2.5 | Kwaliteit van het ontwerp van de gebruikersinterface | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 8.2.6 | Beveiliging van de test tegen ongeautoriseerde toegang tot items of antwoorden | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 8.2.7 | Kwaliteitsvolle beschrijving van de items en duidelijke grafische inhoud bij non-verbale items | n.v.t. | 0 | 1 | 2 | 3 | 4 |
| 8.2.8 | **Kwaliteit van de testmaterialen van CBT- en WBT-testen**  Deze algehele beoordeling wordt verkregen door een oordeel te vellen op basis van de beoordelingen gegeven op onderdelen 8.2.1 -8.2.7 | n.v.t. | 0 | 1 | 2 | 3 | 4 |

|  |
| --- |
| **Opmerkingen van de beoordelaar over de kwaliteit van de testmaterialen** |
| Klik of tik om tekst in te voeren. |

# Normen

**Algemene richtlijnen voor het toekennen van beoordelingen voor deze sectie**

Het is moeilijk om duidelijke criteria op te stellen voor de beoordeling van de technische kwaliteiten van een instrument.

Deze informatie voorziet een richtlijn met betrekking tot de waarden voor een inadequate, adequate, goede en uitstekende beoordeling. Deze zijn echter alleen bedoeld als richtlijnen. De aard van het instrument, het toepassingsgebied, de kwaliteit van de data waarop normen gebaseerd zijn en het soort beslissingen waarvoor het zal worden gebruikt, zouden allemaal van invloed moeten zijn op de manier waarop de beoordelingen worden toegekend.

Er kunnen twee manieren worden onderscheiden om een ruwe testscore in te schalen of te categoriseren, om er zo betekenis aan te geven (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). In de eerste plaats kan een set van ingeschaalde scores of normen afgeleid worden uit de scoreverdeling van een referentiegroep. Dit wordt normgerichte interpretatie genoemd (zie sectie 9.1). In de tweede plaats kunnen normen afgeleid worden uit een vaardigheidsdomein of een onderwerp dat moet worden beheerst (domeingerichte interpretatie) of grensscores kunnen afgeleid worden van de resultaten van empirisch validiteitsonderzoek (criteriumgerichte interpretatie, zie subsectie 9.2). Bij de laatste twee mogelijkheden worden ruwe scores onderverdeeld in twee (bijvoorbeeld ‘geslaagd’ of ‘gefaald’) of meer verschillende scorecategorieën, bv. het toewijzen van patiënten uit verschillende   
score-categorieën aan verschillende behandelingsprogramma’s, het toewijzen van scholieren die lager scoren dan een vastgelegde minimumscore aan remediërend leren, of het aannemen of weigeren van sollicitanten bij personeelselectie.

## Normgerichte interpretatie

*(dit gedeelte kan overgeslagen worden indien niet van toepassing)*

**Informatie over internationale normen**

Er moet goed nagedacht worden over de geschiktheid van internationale (zelfde-taal-)normen. Daar waar deze zorgvuldig zijn vastgelegd op basis van steekproeven uit een groep van landen, dienen deze beoordeeld te worden op dezelfde basis als nationaal gebaseerde   
(enkele-taal-)normgroepen. Als er een niet-lokale norm wordt voorzien, dient er een sterk bewijs aangeleverd te worden voor de gelijkwaardigheid van zowel testversies als steekproeven om het gebruik te rechtvaardigen. Over het algemeen zijn er voor een dergelijk bewijs studies nodig die de schaal-equivalentie demonstreren tussen de oorspronkelijke versie en de vertaalde versies. *Waar dit niet werd aangegeven, dient dit te worden toegelicht in de ‘Opmerkingen van de Beoordelaar’ aan het einde van sectie 9.*   
Een internationale norm is wellicht het meest geschikt voor internationaal gebruik (bv. vergelijken van mensen die de test hebben afgelegd in verschillende talen) maar de hieronder opgenomen punten dienen te worden overwogen bij het bepalen van de geschiktheid ervan. In het algemeen moet bij het gebruik van een internationale norm ten minste de equivalentie van metingen worden aangetoond tussen de oorspronkelijke versie en de vertaalde versies van de test.

De aard van de steekproef:

* De balans van de samenstelling van de steekproef (bv. een steekproef die voor 95% Duits is en voor slechts 2% Italiaans en 3% Brits, is niet echt een internationale steekproef). Om de diverse samenstelling beter te reflecteren kan een steekproef gewogen zijn.
* De equivalentie of gelijkwaardigheid van de achtergrond (beroep, opleiding, test-omstandigheden, etc.) van de verschillende elementen van de steekproef. Normsteekproeven die niet toelaten dit te evalueren zijn ontoereikend.

Het testtype:

* Waar er tests zijn die weinig of geen verbale inhoud hebben zal de impact kleiner zijn op de vertaling. Dit is van toepassing op prestatietesten en in zekere mate op abstracte en schematische redeneringstesten waar er een kleinere impact op de scores zou moeten zijn.

De gelijkwaardigheid van de versie van de test die gebruikt wordt in de steekproeven met verschillende talen:

* Er dient bewijs te zijn dat alle taalversies goed vertaald/aangepast zijn
* Is er enig bewijs dat een van de groepen de test heeft afgelegd in een niet-primaire taal?

Overeenkomsten van scores in verschillende steekproeven:

* Bewijs moet worden aangeleverd over de relatieve scorepatronen van de steekproefsecties van verschillende landen. Grote verschillen dienen te worden verklaard en de implicaties voor het gebruik besproken. Bv. indien een Spaanse steekproef hoger scoort op een schaal dan een Nederlandse steekproef, is er dan een verklaring over wat het betekent om leden van beide groepen of een derde groep te vergelijken met het gemiddelde? Is er een interpretatie van het verschil?

*Het ontbreken van deze bewijsbronnen dient te worden toegelicht in de ‘Opmerkingen van de Beoordelaar’ aan het eind van deze sectie.*

In de handleiding van het instrument dient specifieke informatie meegegeven te worden over de generaliseerbaarheid van de normen voorbij die groepen die opgenomen zijn in de internationale normen.

* Bv. indien een norm bestaat uit 20% Duits, 20% Frans, 20% Italiaans, 20% Brits en 20% Nederlands, is deze mogelijk geschikt als een vergelijkingsgroep voor Zwitserse of Belgische kandidaten, maar deze is mogelijk niet geschikt als een vergelijking voor een groep van Chinese aanmelders.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 9.1 | Norm-gerefereerde interpretatie  Wanneer een instrument ontworpen is voor gebruik zonder steun van normen of referentiegroepen (bv. ipsatieve testen enkel ontworpen voor intra-individuele vergelijkingen), dient de ‘niet van toepassing’ categorie gebruikt te worden in plaats van de ‘geen informatie beschikbaar’. Niettemin moet de beoordelaar wel nagaan of de redenering voor het niet voorzien van normen gerechtvaardigd is, anders dient de categorie ‘geen informatie beschikbaar’ te worden gebruikt. | | | |
| 9.1.1 | Geschiktheid voor lokaal gebruik, bij lokale dan wel internationale normen  Houd er rekening mee dat voor adaptieve testen enkel lokale (op nationaal niveau) of echte internationale normen in aanmerking komen voor het beoordelen van 2, 3 of 4 ook al is er construct-equivalentie gevonden over culturen heen. Waar problemen met meetinvariantie ontstaan, moeten er aparte normen worden opgesteld voor (sub)groepen en moeten eventuele problemen worden verklaard. | | | |
| Niet van toepassing | | | n.v.t. |
| Geen informatie beschikbaar | | | 0 |
| Lokaal niet relevant (bv. ongeschikte buitenlandse steekproeven) | | | 1 |
| Lokale steekproef/steekproeven die niet goed past/passen bij het relevante toepassingsgebied maar met enige voorzichtigheid gebruikt kan/kunnen worden | | | 2 |
| Lokale (nationale) steekproeven of relevante internationale steekproeven met goede relevantie naar de beoogde toepassing | | | 3 |
| Lokale (nationale) steekproeven of relevante internationale steekproeven gebaseerd op goed gedefinieerde populaties van het relevante toepassingsgebied | | | 4 |
| 9.1.2 | Geschiktheid voor beoogde toepassingen | | | |
| Niet van toepassing | | | n.v.t. |
| Geen informatie beschikbaar | | | 0 |
| Norm of normen niet voldoende voor beoogde toepassingen | | | 1 |
| Voldoende algemene populatienormen en/of een brede range aan normtabellen, of voldoende normen voor sommige, maar niet voor alle beoogde toepassingen | | | 2 |
| Goede en brede range aan normtabellen | | | 3 |
| Uitstekend, breed scala aan steekproef relevante, leeftijd-gerelateerd en geslacht-gerelateerde normen met informatie over andere verschillen binnen groepen (bv. etnische groep samenstelling) | | | 4 |
| 9.1.3 | Steekproefgrootte (klassieke normering)  Steekproeven van minder dan 200 testdeelnemers zullen voor de meeste doeleinden te klein zijn, omdat de representatie van deelnemers in de staarten van de verdeling erg beperkt zal zijn. De *SE*gemiddelde voor een *z*-score met *N*= 200 is 0,071 van de SD – of gewoon beter dan één *T-*scorepunt. Hoewel deze mate van onnauwkeurigheid mogelijk slechts minimale gevolgen heeft voor het centrum van de verdeling, kan de impact aan de uiteinden van de verdeling toch behoorlijk groot zijn (en dit kan net de scorespreiding zijn die het meest relevant is voor beslissingen die moeten genomen worden op basis van de testscores). Als er internationale normen zijn, moeten deze in het algemeen vanwege hun heterogeniteit groter zijn dan de typische vereisten van lokale steekproeven.  Verschillende richtlijnen worden gegeven voor gebruik bij besluitvorming met lage versus hoge inzet. Over het algemeen is het gebruik met hoge inzet van toepassing waar een belangrijke beslissing voor een deel gebaseerd is op de testscore(s). | | | |
|  | gebruik met lage inzet | gebruik met hoge inzet |  |
| Niet van toepassing | | | n.v.t. |
| Geen informatie beschikbaar | | | 0 |
| Onvoldoende steekproefgrootte | bv. < 200 | bv. 200-299 | 1 |
| Voldoende steekproefgrootte | bv. 200-299 | bv. 300-399 | 2 |
| Goede steekproefgrootte | bv. 300-999 | bv. 400-999 | 3 |
| Uitstekende steekproefgrootte | bv. ≥ 1000 | bv. ≥ 1000 | 4 |
| 9.1.4 | Steekproefgrootte bij continue normering  Continue normeringsprocedures winnen aan populariteit. Ze worden voornamelijk gebruikt in testen die bedoeld zijn voor gebruik in scholen (bv. van het eerste tot het zesde leerjaar in het basisonderwijs) of voor een specifieke leeftijdsgroep (bv. een intelligentietest voor 6-16 jarigen). Continue normering is efficiënter aangezien er minder respondenten nodig zijn om dezelfde nauwkeurigheid van de normen te verkrijgen. Bechger, Hemker, en Maris (2009) hebben een aantal waarden berekend voor de groottes van continue normgroepen die een gelijkwaardige nauwkeurigheid zouden opleveren als klassieke normering. Bij gebruik van 8 subgroepen, geeft N= 70 (8x70) dezelfde nauwkeurigheid als *N*s van 200 (8x200) in een klassieke aanpak; *N*=100 (x8) vergeleken met 300 (x8) en *N*=150 (x8) met 400 (x8). In deze gevallen is de nauwkeurigheid op basis van de continue normeringsaanpak zelfs beter in de middelste groepen, maar iets slechter in de buitenste groepen. Een ander voordeel, naast een grotere efficiëntie, is dat er waarden voor tussenliggende normgroepen berekend kunnen worden op basis van een regressielijn. Deze aanpak is echter gebaseerd op vrij strikte statistische aannames. De auteur van de test moet aantonen dat aan deze aannames worden voldaan, of dat afwijkingen van deze aannames geen serieuze gevolgen hebben voor de nauwkeurigheid van de normen.  Houd er rekening mee dat wanneer het aantal groepen groter is, het aantal respondenten in elke groep kleiner kan zijn en omgekeerd. Voor belangrijke beslissingen, zoals schooltoelating, wordt het vereiste aantal met één stap verhoogd. | | | |
| Niet van toepassing | | | n.v.t. |
| Geen informatie beschikbaar | | | 0 |
| Onvoldoende steekproefgrootte (bv. minder dan 8 subgroepen met elk een maximum van 69 respondenten) | | | 1 |
| Voldoende steekproefgrootte (bv. 8 subgroepen met elk 70-99 respondenten) | | | 2 |
| Goede steekproefgrootte (bv. 8 subgroepen met elk 100-149 respondenten) | | | 3 |
| Uitstekende steekproefgrootte (bv. 8 subgroepen met elk ten minste 150 respondenten) | | | 4 |
| 9.1.5 | Gebruikte procedures in steekproefselectie *(selecteer er één)*  Een normgroep moet representatief zijn voor de bedoelde groep. Een steekproef is representatief voor de bedoelde populatie indien de samenstelling van de steekproef voor een aantal variabelen (bv. leeftijd, geslacht, scholing) vergelijkbaar is met die van de populatie, *en* wanneer de steekproef verzameld is met behulp van een stochastisch steekproefmodel. In een dergelijk model is de kans om opgenomen te worden in de steekproef even groot voor elk element in de populatie. In zowel stochastische als niet-stochastische steekproeven kunnen verschillende methoden gebruikt worden.  Wanneer bij een stochastische steekproef een individueel persoon de eenheid van selectie is, kan er een onderscheid gemaakt worden tussen drie verschillende methodes: *volledig stochastisch;* *systematisch* (bv. elk tiende element van de populatie) en *gestratificeerd* (voor enkele belangrijke variabelen, bv. geslacht, worden beoogde aantallen vastgesteld om representativiteit te garanderen met betrekking tot deze variabelen). Desondanks (bv. in het belang van efficiëntie) kunnen ook groepen personen worden geselecteerd (bv. schoolklassen), of een combinatie van groepen en individuen.  Bij niet-stochastische steekproeven kan er een onderscheid gemaakt worden tussen vier verschillende methoden: *gelegenheidssteekproeftrekking* (elke geteste persoon wordt simpelweg toegevoegd aan de normgroep, zoals gebruikelijk is bij steekproeven voor personeelselectie; post-hoc gegevens kunnen geclassificeerd worden in betekenisvolle subgroepen gebaseerd op biografische en situationele informatie); *quota steekproeftrekking(*zoals bij de gelegenheidssteekproef, maar het wordt op voorhand gespecifieerd hoeveel respondenten er nodig zijn in elke subgroep, zoals gebruikelijk is bij vragenlijstonderzoek*)*, *sneeuwbalsteekproeftrekking* (hier worden vrienden en vrienden van vrienden, etc. gemobiliseerd om te participeren aan de steekproef) en *doelgericht* (bv. het doelbewust selecteren van extreme groepen). | | | |
| Geen informatie beschikbaar | | |  |
| Stochastische steekproef – volledig stochastisch | | |  |
| Stochastische steekproef – systematisch | | |  |
| Stochastische steekproef – gestratificeerd | | |  |
| Stochastische steekproef – cluster | | |  |
| Stochastische steekproef – getrapt (bv. eerst cluster en vervolgens stochastisch binnen clusters) | | |  |
| Niet-stochastische steeproef - gemak | | |  |
| Niet-stochastische steeproef - quota | | |  |
| Niet-stochastische steeproef – ‘sneeuwbal’ | | |  |
| Niet-stochastische steeproef - doelgericht | | |  |
| Andere (gelieve deze te omschrijven): Klik of tik om tekst in te voeren. | | |  |
| 9.1.6 | Representativiteit van de norm steekproef/steekproeven | | | |
| Niet van toepassing | | | n.v.t. |
| Geen informatie beschikbaar | | | 0 |
| Onvoldoende representatief voor het beoogde toepassingsgebied of de representativiteit kan onvoldoende worden vastgesteld op basis van de voorziene informatie | | | 1 |
| Voldoende | | | 2 |
| Goed | | | 3 |
| Uitstekend: gegevens worden verzameld door middel van een stochastische steekproefmodel; een grondige beschrijving is voorzien van de samenstelling van de steekproef/steekproeven en de populatie(s) rekening houdend met relevante achtergrondvariabelen (zoals geslacht, leeftijd, scholing, culturele achtergrond, beroep); er is een goede representativiteit voor deze variabelen vastgesteld | | | 4 |
| 9.1.7 | Kwaliteit van de beschikbare informatie over verschillen tussen minderheidsgroepen en beschermde groepen, effecten van leeftijd, geslacht etc. | | | |
| Niet van toepassing | | | n.v.t. |
| Geen informatie beschikbaar | | | 0 |
| Onvoldoende informatie | | | 1 |
| Voldoende algemene informatie, met minimale analyse | | | 2 |
| Goede omschrijving en analyses van groepen en verschillen | | | 3 |
| Uitstekend en uitgebreid scala aan analyses en besprekingen van relevante kwesties met betrekking tot gebruik en interpretatie | | | 4 |
| 9.1.8 | Hoe oud zijn de normatieve studies? | | | |
| Niet van toepassing | | | n.v.t. |
| Geen informatie beschikbaar | | | 0 |
| Onvoldoende, 20 jaar of ouder | | | 1 |
| Voldoende, normen tussen 15 en 19 jaar oud | | | 2 |
| Goed, normen tussen 10 en 14 jaar oud | | | 3 |
| Uitstekend, normen minder dan 10 jaar oud | | | 4 |
| 9.1.9 | Oefeneffecten (enkel relevant voor prestatietesten) | | | |
| Niet van toepassing | | | n.v.t |
| Geen informatie beschikbaar, alhoewel oefeneffecten verwacht kunnen worden | | |  |
| Algemene informatie beschikbaar | | |  |
| Normen voor tweede meting na een typisch test-hertestinterval | | |  |

## Criteriumgerefereerde interpretatie

*(dit gedeelte kan overgeslagen worden indien niet van toepassing)*

Om de grensscore(s) te bepalen kan een onderscheid gemaakt worden tussen procedures die gebruik maken van het oordeel van experts (dit is de domeingerichte normering, zie subcategorie 9.2.1) en procedures die gebruik maken van feitelijke gegevens met betrekking tot de relatie tussen de testscore en een extern criterium (dit is de criteriumgerichte normering in enge zin, zie subcategorie 9.2.2).

|  |  |  |
| --- | --- | --- |
| **9.2.1** | **Domein-gerefereerde normering** | |
| 9.2.1.1 | Indien het oordeel van experts wordt gebruikt om de kritieke score te bepalen, zijn de experts dan op gepaste wijze geselecteerd en getraind?  Van de experts wordt verwacht dat ze over de nodige kennis beschikken wat betreft het inhoudsdomein van de test, dat ze een gepaste training hebben genoten in enerzijds het beoordelen van (het werk van) testafnemers en anderzijds in het gebruik van de standaardbepalingsmethode die wordt toegepast. De selectieprocedure van experts en de aangeboden training moeten worden omschreven. | |
| Niet van toepassing | n.v.t. |
| Geen informatie beschikbaar | 0 |
| Onvoldoende | 1 |
| Voldoende | 2 |
| Goed | 3 |
| Uitstekend | 4 |
| 9.2.1.2 | Indien het oordeel van experts wordt gebruikt om de kritieke score te bepalen, is het aantal ingeschakelde experts dan voldoende?  Het vereiste aantal experts hangt af van de taken en de contexten. De hieronder voorgestelde aantallen dienen te worden gezien als een absoluut minimum. | |
| Niet van toepassing | n.v.t. |
| Geen informatie beschikbaar | 0 |
| Onvoldoende (minder dan twee) | 1 |
| Voldoende (twee experts) | 2 |
| Goed (drie experts) | 3 |
| Uitstekend (vier experts of meer) | 4 |
| 9.2.1.3 | Indien het oordeel van experts gebruikt wordt om de kritieke score te bepalen, welke standaardbepalingsmethode wordt er dan aangegeven? *(selecteer er één)* | |
| Nedelsky |  |
| Angoff |  |
| Ebel |  |
| Zieky and Livingstone (beperkte groep) |  |
| Berk (contrast groep) |  |
| Beuk |  |
| Hofstee |  |
| Andere (gelieve deze te omschrijven):Klik of tik om tekst in te voeren. |  |
| 9.2.1.4 | Indien het oordeel van experts gebruikt wordt om de kritieke score te bepalen, welke methode om interbeoordelaarsovereenstemming te berekenen wordt er aangegeven *(selecteer er één)* | |
| Coëfficiënt p₀ |  |
| Coëfficiënt Kappa |  |
| Coëfficiënt Livingstone |  |
| Coëfficiënt Brennan and Kane |  |
| Intraclass correlatiecoëfficiënt (ICC) |  |
| Andere (gelieve deze te omschrijven): Klik of tik om tekst in te voeren. |  |
| 9.2.1.5 | Indien het oordeel van experts gebruikt wordt, wat is de grootte van de interbeoordelaars overeenstemmingscoëfficiënten (bv. Kappa of ICC)  In de wetenschappelijke literatuur zijn er geen eenduidige normen voor de interpretatie van dit soort coëfficiënten, alhoewel over het algemeen waarden onder de .60 gezien worden als onvoldoende. Hieronder wordt de classificatie van Shrout (1998) gevolgd. Het gebruik van de classificatie vereist wel enige voorzichtigheid, omdat de prevalentie of basisratio een invloed kan hebben op de waarde van Kappa. | |
| Niet van toepassing | n.v.t. |
| Geen informatie beschikbaar | 0 |
| Onvoldoende (bv.  *r* < 0.60) | 1 |
| Voldoende (bv. 0.60≤ *r <* 0.70) | 2 |
| Goed (bv. 0.70≤ *r <* 0.80) | 3 |
| Uitstekend (bv. 0.70≤ *r <* 0.80) | 4 |
| 9.2.1.6 | Hoe oud zijn de normatieve studies? | |
| Niet van toepassing | n.v.t. |
| Geen informatie beschikbaar | 0 |
| Onvoldoende, 20 jaar of ouder | 1 |
| Voldoende, normen tussen 15 en 19 jaar oud | 2 |
| Goed, normen tussen 10 en 14 jaar oud | 3 |
| Uitstekend, normen minder dan 10 jaar oud | 4 |
| 9.2.1.7. | Oefeneffecten (enkel relevant voor prestatietesten) | |
| Geen informatie beschikbaar, alhoewel oefeneffecten verwacht kunnen worden |  |
| Algemene informatie beschikbaar |  |
| Normen voor tweede meting na een typisch test-hertestinterval |  |
| **9.2.2** | **Criteriumgerichte normering** | |
| 9.2.2.1 | Indien de grensscore gebaseerd is op empirisch onderzoek, wat zijn dan de resultaten en de kwaliteit van dit onderzoek?  Om deze vraag te beantwoorden, kunnen er geen expliciete richtlijnen worden gegeven over het aanvaardbare niveau van de relatie, niet enkel omdat wat gezien kan worden als ‘hoog’ en ‘laag’ potentieel kan verschillen voor elk criterium dat wordt voorspeld, maar ook omdat voorspelde resultaten beïnvloed zullen worden door andere variabelen zoals de prevalentie of basisratio. Daarom moet de beoordelaar vertrouwen op zijn/haar expertise voor zijn/haar oordeel. Eveneens moet de samenstelling van de steekproef die gebruikt werd voor dit onderzoek (is de steekproef vergelijkbaar met de groep waarvoor de test is bedoeld, heterogener, of homogener?) en de grootte van deze groep in beschouwing worden genomen. | |
| Niet van toepassing | n.v.t. |
| Geen informatie beschikbaar | 0 |
| Onvoldoende | 1 |
| Voldoende | 2 |
| Goed | 3 |
| Uitstekend | 4 |
| 9.2.2.2 | Hoe oud zijn de normatieve studies? | |
| Niet van toepassing | n.v.t. |
| Geen informatie beschikbaar | 0 |
| Onvoldoende, 20 jaar of ouder | 1 |
| Voldoende, normen tussen 15 en 19 jaar oud | 2 |
| Goed, normen tussen 10 en 14 jaar oud | 3 |
| Uitstekend, normen minder dan 10 jaar oud | 4 |
| 9.2.2.3 | Oefeneffecten (enkel relevant voor prestatietesten) | |
| Geen informatie beschikbaar, alhoewel oefeneffecten verwacht kunnen worden |  |
| Algemene informatie beschikbaar |  |
| Normen voor tweede meting na een typisch test-hertestinterval |  |
| 9.3 | **Globale toereikendheid**  Deze globale beoordeling wordt verkregen door een oordeel te vellen op basis van de beoordelingen gegeven voor items 9.1 – 9.2.2.3  De globale beoordeling voor *normgerichte interpretatie* kan nooit hoger zijn dan de beoordeling voor het item over steekproefgrootte, maar het kan wel lager zijn afhankelijk van de andere beschikbare informatie. Hier is vooral de informatie over de representativiteit en het verjaren van de normen relevant. Indien niet-stochastische normgroepen gebruikt worden kan de kwaliteit hoogstens als ‘voldoende’ worden gekwalificeerd, maar alleen wanneer de beschrijving van de normgroep laat zien dat de verdeling van de relevante variabelen vergelijkbaar is met de doelgroep. De globale beoordeling dient de karakteristieken van de grootste en meest betekenisvolle normen weer te geven, veeleer dan een gemiddelde van alle gepubliceerde normen.  De globale beoordeling voor *criteriumgerichte interpretatie* kan, in het geval dat er een beroep wordt gedaan op experts voor het vaststellen van de kritieke scores, nooit hoger zijn dan de beoordeling voor de grootte van de interbeoordelaarsovereenkomst, maar het kan wel lager zijn afhankelijk van de andere beschikbare informatie. Hierbij zijn volgende zaken vooral van belang: de correcte toepassing van de gebruikte methode, de kwaliteit, de training en het aantal experts.   Indien de kritieke score gebaseerd is op empirisch onderzoek, kan de beoordeling nooit hoger zijn dan de beoordeling voor item 9.2.2.1, maar het kan wel lager zijn als de studies te oud zijn. | |
| Niet van toepassing | n.v.t. |
| Geen informatie beschikbaar | 0 |
| Onvoldoende | 1 |
| Voldoende | 2 |
| Goed | 3 |
| Uitstekend | 4 |

|  |
| --- |
| **Opmerkingen van de beoordelaar over de normen:** kort verslag over de normen en hun geschiedenis, inclusief informatie over de specifieke maatregelingen die de uitgever of auteur voorziet voor het regelmatig updaten van normen. Opmerkingen met betrekking tot niet-lokale normen dienen hier gemaakt te worden. |
| Klik of tik om tekst in te voeren. |

# Betrouwbaarheid

**Algemene leidraad voor het toekennen van beoordelingen voor deze sectie**

Betrouwbaarheid verwijst naar de mate waarin scores vrij zijn van meetfoutvariantie (i.e. de spreiding van verwachte meetfouten). Voor betrouwbaarheid zijn de richtlijnen gebaseerd op de vereiste om een zo klein mogelijke standaardfout te hebben bij schattingen van betrouwbaarheid. Richtlijnen worden voorgesteld met betrekking tot twee verschillende contexten: het gebruik van instrumenten om beslissingen te nemen over groepen mensen (bv. organisatiediagnose) en voor individuele beoordelingen. De betrouwbaarheidsvoorwaarden zijn hoger voor individuele dan voor groepsbeoordelingen. Andere factoren kunnen eveneens een invloed uitoefenen op de betrouwbaarheidsvoorwaarden, zoals het soort beslissingen dat wordt genomen en of schalen al dan niet individueel geïnterpreteerd worden of met andere schalen samengevoegd worden in een samengestelde schaal. In het laatste geval dient de betrouwbaarheid van de samengestelde schaal de focus te zijn van de beoordeling, niet de betrouwbaarheden van de componenten.

Wanneer een instrument vertaald en/of aangepast is van een niet-lokale context, kan het bewijs voor de betrouwbaarheid van de originele versie gebruikt worden als ondersteuning van de kwaliteit van de vertaalde/aangepaste versie. In dit geval dient de equivalentie van de test in een nieuwe taal met het origineel te worden aangetoond. Zonder dit bewijs is het niet mogelijk om resultaten te generaliseren van één land/taal-versie naar een andere versie. Echter, voor interne consistentie geniet een betrouwbaarheidsbewijs op basis van lokale groepen de voorkeur, omdat dit bewijs nauwkeuriger is en vaak makkelijker te verkrijgen. Voor enkele richtlijnen aangaande de vaststelling van equivalentie, verwijzen we naar de introductie in de sectie over validiteit. In de bijlage is een lijst met de belangrijkste aandachtspunten als geheugensteuntje opgenomen voor een vertaling of aanpassing van een instrument vanuit een niet-lokale context.

Het is moeilijk om duidelijke criteria te geven voor het beoordelen van de technische kwaliteiten van een instrument. Deze informatie biedt ondersteuning en sturing bij het toekennen van waarden aan de volgende beoordelingen: ‘onvoldoende’, ‘voldoende’, ‘goede’ en ‘uitstekende’. Deze zijn echter alleen bedoeld als richtlijnen. De aard van het instrument, het toepassingsgebied, de kwaliteit van de gegevens waarop betrouwbaarheidsschattingen zijn gebaseerd, en de beslissingstypes waar het voor wordt gebruikt, zouden allemaal invloed moeten hebben op de manier waarop de beoordelingen worden toegekend. In sommige gevallen is een betrouwbaarheid van 0.70 prima, in andere gevallen zal dit niet voldoende zijn. Omwille van deze redenen, dienen de samengevatte beoordelingen gebaseerd te zijn op het oordeel en de expertise van de beoordelaar in kwestie en niet door simpelweg het gemiddelde te nemen van alle andere beoordelingen.

Om een beeld te geven van de range en de verdeling van waarden die geassocieerd worden met de verschillende schalen waar een instrument uit bestaat, wordt in elke sectie het *aantal schalen* ingevuld. Bijvoorbeeld, indien een instrument dat gebruikt wordt voor *beslissingen op groepsniveau* 15 schalen heeft, waarvan er vijf een test-hertestbetrouwbaarheid hebben die lager ligt dan 0.60, zes tussen de 0.60 en 0.70 en de andere vier tussen de 0.70 en 0.80, dan kan de mediane stabiliteit van de mediaan als ‘voldoende’ beoordeeld worden (aangezien dit de categorie is waarbinnen de mediaan van de 15 waarden valt). Indien het gaat over meer dan één studie, dan dient eerst de mediaan per schaal te worden berekend met inachtneming van de steekproefgroottes; in sommige gevallen zijn wellicht resultaten van een meta-analyse beschikbaar, deze kunnen beoordeeld worden op dezelfde manier. Dit wordt dan ingegeven als volgt:

|  |  |  |
| --- | --- | --- |
| **Stabiliteit** | Aantal schalen *(indien van toepassing)* | M\* |
| Geen informatie beschikbaar | [ - ] | 0 |
| Onvoldoende (bv.  *r* < 0.60) | [ 5 ] | 1 |
| Voldoende (bv. 0.60≤ *r <* 0.70) | [ 6 ] | 2 |
| Goed (bv. 0.70≤ *r <* 0.80) | [ 4 ] | 3 |
| Uitstekend (bv. *r ≥* 0.80) | [ 0 ] | 4 |

\* M = stabiliteit van mediaan

Voor elk van de mogelijke beoordelingen worden de voorbeeldwaarden *enkel als richtlijn* gegeven, vooral bij het onderscheid tussen ‘voldoende’, ‘goed’ en ‘uitstekend’. Voor *beslissingen met een hoge inzet*, zoals een personeelsselectie, zullen deze voorbeelden .10 hoger zijn. Echter, het dient te worden opgemerkt dat beslissingen vaak gebaseerd zijn op geaggregeerde schaalscores. Geaggregeerde schalen kunnen een veel hogere betrouwbaarheid hebben dan hun primaire schalen van het onderdeel. Bijvoorbeeld, primaire schalen in een multi-schaal instrument kunnen betrouwbaarheden hebben van rond de 0.70 terwijl de Big Five secundaire samengevoegde schalen die hierop gebaseerd zijn een betrouwbaarheid kunnen hebben van in de 0.90. Goede testhandleidingen zullen de betrouwbaarheden rapporteren van zowel secundaire als primaire schalen.

Het is duidelijk dat het in vele gevallen wellicht onmogelijk is om de feitelijke mediaan te berekenen. Daarom is het noodzakelijk om op basis van de beschikbare informatie in de documentatie de beste schatting te maken. Er is ruimte voorzien om opmerkingen toe te voegen. U kan hier uw bedenkingen over de nauwkeurigheid van uw schattingen noteren. Bijvoorbeeld, in sommige gevallen kan bij een zeer hoog niveau van interne consistentie opgemerkt worden dat het gaat om een ‘bloated specific’.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **10** | **Betrouwbaarheid** | | | | | |
| 10.1 | Beschikbare gegevens over betrouwbaarheid *(selecteer er twee indien van toepassing)* | | | | | |
| Geen informatie beschikbaar | | | | |  |
| Slechts één betrouwbaarheidscoëfficiënt (voor elke schaal of subschaal) | | | | |  |
| Slechts één schatting of standaard metingsfout (voor elke schaal of subschaal) | | | | |  |
| Betrouwbaarheidscoëfficiënten voor een aantal verschillende groepen (voor elke schaal of subschaal) | | | | |  |
| Standaard metingsfout voor een aantal verschillende groepen (voor elke schaal of sub-schaal) | | | | |  |
| **10.2** | **Interne consistentie**  Het gebruik van interne consistentie coëfficiënten is weinig zinvol voor het beoordelen van de betrouwbaarheid van speed-tests heterogene schalen (ook wel empirische of criterium-versleutelde schalen; Cronbach, 1970), effectindicatoren (Nunnally & Bernstein, 1994) en spontaan optredende eigenschappen (Schneider & Hough, 1995). In deze gevallen dienen alle items betreffende de interne consistentie gemarkeerd te worden als *‘niet van toepassing’*. Interne consistentie is ook niet geschikt als methode voor het schatten van de betrouwbaarheid van ipsatieve schalen. Alternatieve methodes of test-hertestmetingen zijn beter geschikt voor dit type schalen.  Interne consistentie coëfficiënten geven een betere schatting van betrouwbaarheid dan *split-half-*coëfficiënten die gecorrigeerd zijn met de Spearman-Brown formule. Daarom is het gebruik van *split-half*, ongeacht de achterliggende reden, enkel gerechtvaardigd als er geen informatie beschikbaar is over de antwoorden op de individuele items. *Split-half-*coëfficiënten kunnen worden opgenomen in item 10.7 (‘Andere methoden’). | | | | | |
| 10.2.1 | Steekproefgrootte | | | | | |
| Niet van toepassing | | | | | n.v.t. |
| Geen informatie beschikbaar | | | | | 0 |
| Één onvoldoende grote studie (bv. steekproefgrootte kleiner dan 100) | | | | | 1 |
| Één voldoende grote studie (bv. steekproefgrootte van 100-200) | | | | | 2 |
| Één grote (bv. steekproefgrootte groter dan 200) of meer dan één voldoende grote studie | | | | | 3 |
| Breed scala aan voldoende grote tot grote studies | | | | | 4 |
| 10.2.2 | Soorten gerapporteerde coëfficiënten *(selecteer zoveel* *als mogelijk)* | | | | | |
| Niet van toepassing | | | | | n.v.t. |
| Coëfficiënt alpha of KR-20 | | | | |  |
| Lambda-2 | | | | |  |
| *Greatest lower bound* (De grootste ondergrens) | | | | |  |
| Omega (factor analyse) | | | | |  |
| Theta (factor analyse) | | | | |  |
| Andere (gelieve deze te omschrijven): Klik of tik om tekst in te voeren. | | | | |  |
| 10.2.3 | Grootte van coëfficiënten | | | Aantal schalen  *(indien van toepassing)* | | M\* |
| Niet van toepassing | | | | | n.v.t. |
| Geen informatie beschikbaar | | | [Klik of tik om tekst in te voeren.] | | 0 |
| Onvoldoende (bv.  *r* < 0.70) | | | [Klik of tik om tekst in te voeren.] | | 1 |
| Voldoende (bv. 0.70≤ *r <* 0.80) | | | [Klik of tik om tekst in te voeren.] | | 2 |
| Goed (bv. 0.80≤ *r <* 0.90) | | | [Klik of tik om tekst in te voeren.] | | 3 |
| Uitstekend (bv. *r ≥* 0.90) | | | [Klik of tik om tekst in te voeren.] | | 4 |
| 10.2.4 | Betrouwbaarheidscoëfficiënten worden vermeld op basis van steekproeven die ... *(selecteer er één)* | | | | | |
| …niet overeenkomen met de beoogde testdeelnemers, wat leidt tot meer gunstige coëfficiënten (bv. inflatie door kunstmatige heterogeniteit) | | | | |  |
| …niet overeenkomen met de beoogde testdeelnemers, maar het effect op de grootte van de coëfficiënten is onduidelijk | | | | |  |
| …niet overeenkomen met de beoogde testafnemers, wat leidt tot minder gunstige coëfficiënten (bv. reductie door beperking van de *range*) | | | | |  |
| …overeenkomen met de beoogde testdeelnemers | | | | |  |
| Niet van toepassing | | | | | n.v.t. |
| **10.3** | **Test-hertestbetrouwbaarheid – temporele stabiliteit**  Test-hertest verwijst naar een relatief kort tijdsinterval, terwijl temporele stabiliteit eerder refereert naar een langer tijdsinterval waarin meer verandering acceptabel is. In het bijzonder voor testen die gebruikt worden om voorspellingen op langere termijn te doen, zijn beide aspecten relevant. Om de temporele stabiliteit te kunnen beoordelen kan meer dan één hertest noodzakelijk zijn.  Het gebruik van een test-hertestdesign is weinig zinvol voor het beoordelen van de betrouwbaarheid van een momentopname (eigenlijk zou een hoge test-hertestcoëfficiënt het momentkenmerk van een test ongeldig maken). In dit geval dienen alle onderdelen met betrekking tot test-hertestbetrouwbaarheid aangeduid worden als ‘niet van toepassing’. | | | | | |
| 10.3.1 | Steekproefgrootte | | | | | |
| Niet van toepassing | | | | | n.v.t. |
| Geen informatie beschikbaar | | | | | 0 |
| Één onvoldoende grote studie (bv. steekproefgrootte kleiner dan 100) | | | | | 1 |
| Één voldoende grote studie (bv. steekproefgrootte van 100-200) | | | | | 2 |
| Één grote (bv. steekproefgrootte groter dan 200) of meer dan één voldoende grote studie | | | | | 3 |
| Breed scala aan voldoende grote tot grote studies | | | | | 4 |
| 10.3.2 | Grootte van coëfficiënten | | | | Aantal schalen  *(indien van toepassing)* | M\* |
| Niet van toepassing | | | | | n.v.t. |
| Geen informatie beschikbaar | | | | [Klik of tik om tekst in te voeren.] | 0 |
| Onvoldoende (bv.  *r* < 0.60) | | | | [Klik of tik om tekst in te voeren.] | 1 |
| Voldoende (bv. 0.60≤ *r <* 0.70) | | | | [Klik of tik om tekst in te voeren.] | 2 |
| Goed (bv. 0.70≤ *r <* 0.80) | | | | [Klik of tik om tekst in te voeren.] | 3 |
| Uitstekend (bv. *r* ≥ 0.80) | | | | [Klik of tik om tekst in te voeren.] | 4 |
| 10.3.3 | Gegevens voorzien over de test-hertestinterval *(selecteer of vul test hertestinterval in)* | | | | | |
| Niet van toepassing | | | | | n.v.t. |
| Geen informatie beschikbaar | | | | |  |
| De interval bedraagt: … | | | | |  |
| 10.3.4 | Betrouwbaarheidscoëfficienten worden vermeld op basis van steekproeven die ... *(selecteer er één)* | | | | | |
| …niet overeenkomen met de beoogde testdeelnemers, wat leidt tot meer gunstige coëfficiënten (bv. inflatie door kunstmatige heterogeniteit) | | | | |  |
| …niet overeenkomen met de beoogde testdeelnemers, maar het effect op de grootte van de coëfficiënten is onduidelijk | | | | |  |
| …niet overeenkomen met de beoogde testdeelnemers, wat leidt tot minder gunstige coëfficiënten (reductie door beperking van de *range*) | | | | |  |
| …overeenkomen met de beoogde testdeelnemers | | | | |  |
| Niet van toepassing | | | | | n.v.t. |
| **10.4** | **Paralleltestbetrouwbaarheid (parallel of alternatieve vormen)** | | | | | |
| 10.4.1 | Steekproefgrootte | | | | | |
| Niet van toepassing | | | | | n.v.t. |
| Geen informatie beschikbaar | | | | | 0 |
| Één onvoldoende grote studie (bv. steekproefgrootte kleiner dan 100) | | | | | 1 |
| Één voldoende grote studie (bv. steekproefgrootte van 100-200) | | | | | 2 |
| Één grote (bv. steekproefgrootte groter dan 200) of meer dan één voldoende grote studie | | | | | 3 |
| Breed scala aan voldoende grote tot grote studies | | | | | 4 |
| 10.4.2 | Wordt er voldaan aan de aannames voor parallelisme\* voor de verschillende versies van de test waarvoor de paralleltestbetrouwbaarheid wordt onderzocht?  \*Houd er rekening mee dat testen als parallel kunnen worden beschouwd indien in dezelfde groep, de gemiddelde scores, variantie en correlaties met andere testen hetzelfde zijn. | | | | | |
| Niet van toepassing | | | | | n.v.t. |
| Geen informatie beschikbaar | | | | | 0 |
| Onvoldoende | | | | | 1 |
| Voldoende | | | | | 2 |
| Goed | | | | | 3 |
| Uitstekend | | | | | 4 |
| 10.4.3 | Grootte van coëfficiënten | Aantal schalen  *(indien van toepassing)* | | | | M\* |
| Niet van toepassing | | | | | n.v.t. |
| Geen informatie beschikbaar | [Klik of tik om tekst in te voeren.] | | | | 0 |
| Onvoldoende (bv.  *r* < 0.70) | [Klik of tik om tekst in te voeren.] | | | | 1 |
| Voldoende (bv. 0.70≤ *r <* 0.80) | [Klik of tik om tekst in te voeren. ] | | | | 2 |
| Goed (bv. 0.80≤ *r <* 0.90) | [ Klik of tik om tekst in te voeren.] | | | | 3 |
| Uitstekend (bv. *r ≥* 0.90) | [Klik of tik om tekst in te voeren.] | | | | 4 |
| 10.4.4 | Betrouwbaarheidscoëfficiënten worden vermeld op basis van steekproeven die ... *(selecteer er één)* | | | | | |
| …niet overeenkomen met de beoogde testdeelnemers, wat leidt tot meer gunstige coëfficiënten (bv. inflatie door kunstmatige heterogeniteit) | | | | |  |
| …niet overeenkomen met de beoogde testdeelnemers, maar het effect op de grootte van de coëfficiënten is onduidelijk | | | | |  |
| …niet overeenkomen met de beoogde testdeelnemers, wat leidt tot minder gunstige coëfficiënten (reductie door beperking in de range) | | | | |  |
| …overeenkomen met de beoogde testdeelnemers | | | | |  |
| Niet van toepassing | | | | | n.v.t. |
| **10.5** | **IRT (item-responstheorie)-gebaseerde methode** | | | | | |
| 10.5.1 | Steekproefgrootte  Bij het gebruik van IRT-methodes voor de schatting van de betrouwbaarheid is het moeilijk om uniforme richtlijnen te geven over de geschiktheid van steekproefgroottes, omdat de vereiste verschillen afhankelijk van het antwoordopzet en het antwoordmodel van het item. Afhankelijk van het item response model zijn minimale waarden voor ‘voldoende’ steekproefgroottes: 200 voor studies met 1 parameter, 400 voor studies met 2 parameters en 700 voor studies met 3 parameters (gebaseerd op Parshall, Davey, Spray & Kalohn, 2001). Deze waarden zijn van toepassing op dichotome modellen, maar kunnen ook hulp bieden aan de beoordelaar wanneer polytome modellen gebruikt worden waarvoor de steekproefgroottes kleiner kunnen zijn. | | | | | |
| Niet van toepassing | | | | | n.v.t. |
| Geen informatie beschikbaar | | | | | 0 |
| Één onvoldoende grote studie | | | | | 1 |
| Één voldoende grote studie | | | | | 2 |
| Één grote of meer dan één voldoende grote studie | | | | | 3 |
| Breed scala aan voldoende grote tot grote studies | | | | | 4 |
| 10.5.2 | Soorten gerapporteerde coëfficiënten *(selecteer alles wat van toepassing is)*  De eerste methode geeft de betrouwbaarheid van de geschatte latente trek die in IRT de geschatte feitelijke score (i.e. testscore) vervangt, (zie Embretson & Reise, 2000). De tweede methode is gebaseerd op informatie over de individuele items en geeft een schatting van de betrouwbaarheid wanneer aan de vereisten die kenmerkend zijn voor IRT voldaan is (Mokken, 1971). De derde methode geeft een schatting van de nauwkeurigheid van de meting met betrekking tot de positie op de latente trek. | | | | | |
| Betrouwbaarheid van de geschatte latente trek | | | | |  |
| Rho | | | | |  |
| Testinformatie functie | | | | |  |
| Andere (gelieve deze te omschrijven): Klik of tik om tekst in te voeren. | | | | |  |
| Niet van toepassing | | | | | n.v.t. |
| 10.5.3 | Grootte van coëfficiënten (gebaseerd op de finale testlengte)  Er worden richtlijnen gegeven voor zowel betrouwbaarheidscoëfficiënten (inclusief rho) als voor de testinformatiefunctie. De richtlijnen voor de testinformatiefunctie zijn gebaseerd op die van betrouwbaarheidscoëfficiënten aangezien Informatie = 1/SE2, en gegeven enkele vaak gehanteerde aannames, *r* = 1 - SE2. Houd er rekening mee dat SE en informatiewaarden afhankelijk zijn van de latente trek-waarde en dat iedere test een *range* heeft waarbinnen de informatiewaarde optimaal is. De beoordeling dient niet a priori gebaseerd te zijn op deze optimale waarde, maar op de informatiewaarde van de score of de *range* aan scores die van specifiek belang zijn (bv. kritieke scores). Voor deze scores kan de informatiewaarde optimaal zijn, maar dat is niet noodzakelijkerwijs zo. Als er geen dergelijke scores zijn, dient de beoordeling gebaseerd te zijn op de gemiddelde informatiewaarde (zie ook Reise & Havilund, 2005). Aangezien deze vuistregels nog niet veel gebruikt zijn, adviseren we beoordelaars om deze met de nodige voorzichtigheid te gebruiken. | | Aantal schalen  *(indien van toepassing)* | | | M\* |
| Niet van toepassing | | | | | n.v.t. |
| Geen informatie beschikbaar | | [ Klik of tik om tekst in te voeren. ] | | | 0 |
| Onvoldoende (bv.  *r* < 0.70; informatie < 3.33) | | [ Klik of tik om tekst in te voeren. ] | | | 1 |
| Voldoende (bv. 0.70≤ *r <* 0.80; 3.33 ≤ informatie < 5.00) | | [Klik of tik om tekst in te voeren.] | | | 2 |
| Goed (bv. 0.80≤ *r <* 0.90; 5.00 ≤ informatie < 10.00) | | [ Klik of tik om tekst in te voeren. ] | | | 3 |
| Uitstekend (bv. *r ≥* 0.90; informatie ≥ 10.00) | | [Klik of tik om tekst in te voeren.] | | | 4 |
| **10.6** | **Interbeoordelaarsbetrouwbaarheid**  Als de scoring van de test geen beoordelend proces inhoudt (bv. het simpelweg optellen van de antwoorden op meerkeuzevragen) dan is dit type betrouwbaarheid niet noodzakelijk en dienen alle items aangaande interbeoordelaarsbetrouwbaarheid gemarkeerd te worden met *‘niet van toepassing’*. Houd er rekening mee dat, hoewel het kan zijn dat de interbeoordelaars-betrouwbaarheid niet van toepassing is op de test als geheel, het wel van toepassing kan zijn op één of meerdere subtesten (bv. enkele subtesten van een intelligentietest). | | | | | |
| 10.6.1 | Steekproefgrootte | | | | | |
| Niet van toepassing | | | | | n.v.t. |
| Geen informatie beschikbaar | | | | | 0 |
| Één onvoldoende grote studie (bv. steekproefgrootte kleiner dan 100) | | | | | 1 |
| Één voldoende grote studie (bv. steekproefgrootte van 100-200) | | | | | 2 |
| Één grote (bv. steekproefgrootte groter dan 200) of meer dan één voldoende grote studie | | | | | 3 |
| Breed scala aan voldoende grote tot grote studies | | | | | 4 |
| 10.6.2 | Soorten gerapporteerde coëfficiënten *(selecteer zoveel mogelijk)* | | | | | |
| Niet van toepassing | | | | | n.v.t. |
| Percentage overeenkomst | | | | |  |
| Coëfficiënt Kappa | | | | |  |
| Intraclass correlatiecoëfficient | | | | |  |
| Coëfficiënt Iota | | | | |  |
| Andere (gelieve deze te omschrijven): Klik of tik om tekst in te voeren. | | | | |  |
|  | Grootte van coëfficiënten  Voor sommige vermelde methodes in 10.6.2 zijn de voorbeeldwaarden niet van toepassing omdat er geen *r’s* worden berekend. | Aantal schalen  *(indien van toepassing)* | | | | M\* |
| Niet van toepassing | | | | | n.v.t. |
| Geen informatie beschikbaar | [Klik of tik om tekst in te voeren. ] | | | | 0 |
| Onvoldoende (bv.  *r* < 0.60) | [Klik of tik om tekst in te voeren.] | | | | 1 |
| Voldoende (bv. 0.60≤ *r <* 0.70) | [ Klik of tik om tekst in te voeren.] | | | | 2 |
| Goed (bv. 0.70≤ *r <* 0.80) | [Klik of tik om tekst in te voeren.] | | | | 3 |
| Uitstekend (bv. *r* ≥ 0.80) | [Klik of tik om tekst in te voeren.] | | | | 4 |
| **10.7** | **Andere methoden voor betrouwbaarheidsschatting** | | | | | |
| 10.7.1 | Steekproefgrootte | | | | | |
| Niet van toepassing | | | | | n.v.t. |
| Geen informatie beschikbaar | | | | | 0 |
| Één onvoldoende grote studie (bv. steekproefgrootte kleiner dan 100) | | | | | 1 |
| Één voldoende grote studie (bv. steekproefgrootte van 100-200) | | | | | 2 |
| Één grote (bv. steekproefgrootte groter dan 200) of meer dan één voldoende grote studie | | | | | 3 |
| Breed scala aan voldoende grote tot grote studies | | | | | 4 |
| 10.7.2 | Omschrijf methode:  Klik of tik om tekst in te voeren. | | | | | |
| 10.7.3 | Resultaten | Aantal schalen  *(indien van toepassing)* | | | | M\* |
| Niet van toepassing | | | | | n.v.t. |
| Geen informatie beschikbaar | [ Klik of tik om tekst in te voeren. ] | | | | 0 |
| Onvoldoende | [ Klik of tik om tekst in te voeren. ] | | | | 1 |
| Voldoende | [ Klik of tik om tekst in te voeren. ] | | | | 2 |
| Goed | [ Klik of tik om tekst in te voeren. ] | | | | 3 |
| Uitstekend | [ Klik of tik om tekst in te voeren. ] | | | | 4 |
| 10.8 | **Globale toereikendheid**  Deze globale beoordeling wordt verkregen door een oordeel te vellen op basis van de beoordelingen gegeven voor items 10.1 – 10.7.3. *Bereken niet simpelweg de gemiddelde scores om de globale beoordeling te verkrijgen.*  Voor sommige instrumenten kan interne consistentie ongeschikt zijn (brede kenmerken of geaggregeerde schalen), in deze moet gevallen meer nadruk worden gelegd op de test-hertestgegevens. In andere gevallen (momentopnamen), zouden test-hertestbetrouwbaarheden ongeschikt zijn, dus dient de nadruk te worden gelegd op interne consistenties. Houd voor het formuleren van een eindoordeel rekening met volgende punten:   * of de test gebruikt wordt voor individuele- of groepsbeoordelingen * de aard van de keuze/ beslissing (beslissingen met hoge versus lage inzet) * of er één of meerdere (types van) betrouwbaarheidsstudies gerapporteerd worden * of er ook standaardmeetfouten (Standard errors of measurement) worden gerapporteerd/weergegeven * procedurele kwesties, bv. groepsgrootte, het aantal betrouwbaarheidsstudies, heterogeniteit van de groep(en) waarop de coëfficiënten zijn berekend, aantal beoordelaars indien interbeoordelaarsovereenkomst berekend wordt, lengte van het test-hertestinterval, etc. * volledigheid van de rapportage over de betrouwbaarheidsstudies | | | | | |
| Geen informatie beschikbaar | | | | | 0 |
| Onvoldoende | | | | | 1 |
| Voldoende | | | | | 2 |
| Goed | | | | | 3 |
| Uitstekend | | | | | 4 |

|  |
| --- |
| **Opmerkingen van de beoordelaar over betrouwbaarheid:** Benadruk de sterke en zwakke aspecten van het beschikbare bewijs voor betrouwbaarheid. Opmerkingen met betrekking tot betrouwbaarheidsgeneralisatie dienen hier ook gemaakt te worden (indien van toepassing). |
| Klik of tik om tekst in te voeren. |

# Validiteit

**Algemene richtlijnen bij het toewijzen van beoordelingen bij dit onderdeel**

Validiteit is de mate waarin een test aan zijn doel beantwoordt: kunnen de beoogde conclusies afgeleid worden uit de testscores? In de literatuur worden vele types validiteit onderscheiden, bv. Drenth en Sijtsma (2006, p. 334-340) vermelden acht verschillende types. De verschillen kunnen betrekking hebben op ofwel het doel ofwel het proces van de validering bij specifieke technieken van gegevensanalyse. In de laatste decennia van vorige eeuw was er een groeiende consensus om validiteit als één concept te beschouwen en dat variaties in types validiteit enkel dienen gepercipieerd te worden als verschillende manieren om bewijs te vergaren (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Borsboom, Mellenbergh, en Van Heerden (2004) stellen dat een test valide is voor het meten van een attribuut wanneer variatie in het attribuut causaal gerelateerd is aan variatie in de gemeten uitkomsten. Hoewel dit een andere benadering is, zijn deze auteurs eveneens de mening toegedaan dat een differentiatie tussen types validiteit niet relevant is.

Welke validiteitsbenadering ook gekozen wordt, het is niettemin noodzakelijk om bij een gestandaardiseerde beoordeling het concept validiteit enigszins te structureren. Daarom wordt een onderscheid gemaakt tussen begrips- en criteriumvaliditeit in aparte subsecties. Afhankelijk van het doel van de test kan één van deze aspecten van validiteit meer relevant zijn dan de andere. Hoe dan ook is begripsvaliditeit het fundamenteler concept en kan het vinden van bewijs voor criteriumvaliditeit bijdragen tot het oordeel over de begripsvaliditeit van een test. Verschillende soorten validiteit kunnen van toepassing zijn op een test, afhankelijk van het soort beslissingen die op basis van de test genomen worden, de gebruikte steekproeventypes, etc. Hoe dan ook vormt het maken van een kwaliteitsbeoordeling over *de* (begrips- of criterium)validiteit van een test een inherent onderdeel van een testbeoordelingssysteem. Deze beoordeling moet een beschouwing zijn van de kwaliteit van het bewijs ter onderbouwing van de stelling dat de test gebruikt kan worden voor de in de handleiding beschreven interpretaties. Hoe breder de bedoelde toepassingen, hoe meer bewijs aangaande de validiteit de auteur/uitgever zou moeten aanleveren. Houd er rekening mee dat de eindbeoordeling van begrips- en criteriumvaliditeit een soort gemiddelde van dit bewijs is en er situaties of groepen kunnen zijn waarvoor de test een hogere of lagere validiteit heeft (of waarvoor de validiteit helemaal niet bestudeerd kon worden).

Wanneer een instrument vertaald en/of aangepast werd vanuit een niet-lokale context, dan moet bewijs worden geleverd aangaande de equivalentie van de meting in een nieuwe taal ten opzichte van het origineel. Zonder dit bewijs is het niet mogelijk om bevindingen te veralgemenen van één land/taal versie naar de andere. Voorbeelden van bewijs voor equivalentie zijn:

* Invariantie in constructstructuur – bv. via factorstructuur of correlatie met standaardmetingen.
* Vergelijkbare criteriumgerelateerde validiteit – bv. vergelijkbaar profiel van correlaties van een multi-schaalinstrument met een onafhankelijk extern criterium – zoals een maat voor jobcompetenties.
* Items tonen gelijkaardige patronen van schaalladingen – bv. items correleren volgens hetzelfde patroon met andere schalen; sterkste/zwakste itemladingen zijn gelijk in de originele en nieuwe taal.
* Tweetalige kandidaten hebben gelijke profielen in de twee talen (cf. alternatieve vorm betrouwbaarheid).

**Er is sterker bewijs nodig voor de generaliseerbaarheid van de validiteit wanneer testen vertaald worden over linguïstische families heen** (bv. van een Indo-Europese naar een Semitische taal). In een dergelijke situatie is er een grotere bedreiging voor de equivalentie door de verschillen in taalstructuur en culturele verschillen. Validiteitsgeneralisatie kan echter afgeleid worden **wanneer invariantie werd aangetoond in de validiteit bij voorgaande vertalingen van de test in meerdere talen**. Dit is bijvoorbeeld het geval als een Zweedse test reeds vertaald is geweest in het Frans, Duits en Italiaans en er equivalentie werd aangetoond voor de versies in deze talen.

Aangaande de kwestie van equivalentie, kan het bruikbaar zijn om de classificatie van Van de Vijver en Poortinga (2005) te volgen:

* Structurele/functionele equivalentie
  + Er is bewijs dat de versies van de oorspronkelijke taal en de doeltaal dezelfde psychologische constructen meten over groepen heen. Dit wordt over het algemeen aangetoond door aan te tonen dat patronen van correlaties tussen variabelen dezelfde zijn over groepen heen.
* Equivalentie van meeteenheden
  + Er is bewijs dat de meeteenheden hetzelfde zijn, maar er zijn verschillende oorsprongen over de groepen heen (i.e. individuele verschillen gevonden in groep A kunnen vergeleken worden met verschillen gevonden in groep B, maar de absolute ruwe scores van A en B zijn niet direct vergelijkbaar zonder enige vorm van herschaling).
* Equivalentie van schalen/volledige scores
  + Dezelfde meeteenheid en dezelfde oorsprong (i.e. ruwe scores hebben dezelfde betekenis en kunnen vergeleken worden over de groepen heen).

De criteria en opmerkingen in de subsecties 11.1 en 11.2 bieden enige begeleiding bij welke waarden verbonden moeten worden met de beoordelingen onvoldoende, voldoende, goed en uitstekend. Deze zijn echter enkel bedoeld als richtlijnen. De aard van het instrument, het toepassingsgebied, de kwaliteit van de gegevens waarop de validiteitschattingen worden gebaseerd, en de soorten beslissingen waarvoor het instrument zal gebruikt worden, moeten allen worden meegenomen in manier waarop de beoordelingen worden gegeven. Voor validiteit zijn de richtlijnen over de steekproefgroottes gebaseerd op een poweranalyse van de steekproefgroottes die nodig zijn voor het al dan niet vinden van een middelmatige validiteit, indien deze bestaat.

## Begripsvaliditeit

Het doel van begripsvaliditeit is het vinden van een antwoord op de vraag of een test daadwerkelijk het bedoelde begrip meet of, gedeeltelijk of voornamelijk, iets anders. Veelgebruikte methoden om begripsvaliditeit te onderzoeken zijn exploratieve of confirmatorische factoranalyse, item-test correlaties, het vergelijken van de gemiddelde scores van groepen waarvan men mag verwachten dat ze verschillen zullen vertonen, het testen van invariantie van factorstructuur en *item-bias* *(Differential Item Functioning (DIF))* voor verschillende groepen, het berekenen van correlaties met tests die hetzelfde begrip(convergente validiteit) of verschillende begrippen zouden moeten meten (discriminante validiteit), *Multi-Trait-Multi-Method*-onderzoek (MTMM), Item-Respons Theorie-methodologie en (quasi-)experimentele designs.

|  |  |  |
| --- | --- | --- |
| **11.1** | **Begripsvaliditeit** | |
| 11.1.1 | Gebruikte designs *(selecteer alles wat van toepassing is)* | |
| Geen informatie beschikbaar |  |
| Exploratieve factoranalyse |  |
| Confirmatorische factoranalyse |  |
| (Gecorrigeerde) item-test correlaties |  |
| Testen van structuurinvariantie en *differential item functioning* over groepen heen |  |
| Verschillen tussen groepen |  |
| Correlaties met andere instrumenten en prestatiecriteria |  |
| MTMM-correlaties |  |
| IRT-methodologie |  |
| (Quasi-)Experimentele Design |  |
| Andere, gelieve deze te omschrijven: Klik of tik om tekst in te voeren. |  |
| 11.1.2 | Ondersteunen de resultaten van de (exploratieve of confirmatorische) factoranalyse de structuur van de test? | |
| Geen informatie beschikbaar | 0 |
| Onvoldoende | 1 |
| Voldoende | 2 |
| Goed | 3 |
| Uitstekend | 4 |
| 11.1.3 | Correleren de items voldoende goed met de (sub)testscore?  Houd er rekening mee dat zeer hoge correlaties kunnen betekenen dat items min of meer synoniem zijn en dat het gemeten concept zeer nauw kan zijn (een zo genoemde ‘*bloated specific*’). | |
| Geen informatie beschikbaar | 0 |
| Onvoldoende | 1 |
| Voldoende | 2 |
| Goed | 3 |
| Uitstekend | 4 |
| 11.1.4 | Is de factorstructuur invariant over groepen heen en/of is de test vrij van itembias (DIF)?  Dit soort onderzoek kan plaatsvinden op basis van modellen die passen binnen de klassieke testtheorie of de IRT-. Als item-bias gevonden wordt, dan moet het effect hiervan op de totale score geschat worden (kleine effecten zijn acceptabel). | |
| Geen informatie beschikbaar | 0 |
| Onvoldoende | 1 |
| Voldoende | 2 |
| Goed | 3 |
| Uitstekend | 4 |
| 11.1.5 | Zijn de verschillen in gemiddelde scores tussen relevante groepen zoals verwacht?  Bv. van leerlingen uit het 6e leerjaar wordt verwacht dat ze hoger scoren op een test voor numerieke vaardigheden dan leerlingen uit het 4e leerjaar; van kinderen met de diagnose ADHD wordt verwacht dat ze hoger scoren op een test voor hyperactiviteit dan kinderen die niet gediagnosticeerd zijn met ADHD; van verkooppersoneel wordt verwacht dat ze hoger scoren op een test die commerciële kennis nagaat dan de gemiddelde werkende populatie. Ook al zijn de resultaten in de verwachte richting, dit biedt meestal geen uitsluitsel over de begripsvaliditeit van de test. De waarde van dit onderzoek is echter dat wanneer de verwachte verschillen niet gevonden worden, dit sterke twijfels doet rijzen over de begripsvaliditeit van de test. | |
| Geen informatie beschikbaar | 0 |
| Onvoldoende | 1 |
| Voldoende | 2 |
| Goed | 3 |
| Uitstekend | 4 |
| 11.1.6 | De mediaan en spreiding van de correlaties tussen de test en testen die gelijkaardige constructen meten  Een essentieel element van het proces van begripsvaliditeit is het correleren van de testscore(s) met schalen van gelijkaardige instrumenten, de zogenoemde congruente of convergente validiteit. De richtlijnen voor congruente validiteitscoëfficiënten moeten flexibel geïnterpreteerd worden. Waar twee zeer gelijkaardige instrumenten zijn gecorreleerd (met gelijktijdig verzamelde gegevens), zouden we correlaties van 0.60 of meer verwachten te vinden voor een ‘voldoende’. Wanneer de instrumenten minder gelijkaardig zijn aan elkaar of de afnamesessies van elkaar gescheiden zijn door een bepaald tijdsinterval, dan kunnen lagere waarden ‘voldoende’ zijn. Bij evaluatie van de congruente validiteit is voorzichtigheid aangewezen bij het interpreteren van zeer hoge correlaties. Als de correlaties hoger zijn dan 0.90, bestaat de kans dat de schalen in kwestie exact hetzelfde construct meten. Dit is geen probleem indien de schalen in kwestie een nieuwe schaal en een reeds gevestigd ijkpunt representeren. Het zou wel een probleem zijn als de desbetreffende scha(a)l(en) bedoeld was (waren) om bruikbare variantie toe te voegen aan wat andere schalen al meten. De gegeven richtlijnen betreffen de correlaties die niet aangepast zijn aan variantie te wijten aan een gelijkaardige methode. Daarom moeten ook de betrouwbaarheden van beide instrumenten in rekening gebracht worden bij het beoordelen van de congruente validiteitscoëfficiënten. Bv., wanneer beide instrumenten een betrouwbaarheid hebben van .75, dan is de maximum correlatie tussen de instrumenten .56. Indien de betrouwbaarheidscoëfficiënten hoger liggen, dan worden hogere correlaties verwacht. | |
| Geen informatie beschikbaar | 0 |
| Onvoldoende (*r* < 0.55) | 1 |
| Voldoende (0.55 ≤ *r* < 0.65) | 2 |
| Goed (0.65 ≤ *r* < 0.75) | 3 |
| Uitstekend (*r* ≥ 0.75) | 4 |
| 11.1.7 | Tonen de correlaties met andere instrumenten goede discriminante validiteit aan met betrekking tot constructen die de test niet verondersteld wordt te meten? | |
| Geen informatie beschikbaar | 0 |
| Onvoldoende | 1 |
| Voldoende | 2 |
| Goed | 3 |
| Uitstekend | 4 |
| 11.1.8 | Wanneer een *Multi-Trait-Multi-Method*-design wordt gebruikt, ondersteunen de resultaten dan de begripsvaliditeit van de test (meet de test echt wat het verondersteld wordt te meten en niet iets anders)?  Houd er rekening mee dat indien een MTMM-design gebruikt wordt, onderzoek zoals vermeld in 11.1.6 en 11.1.7 mogelijk niet meer nodig is. | |
| Geen informatie beschikbaar | 0 |
| Onvoldoende | 1 |
| Voldoende | 2 |
| Goed | 3 |
| Uitstekend | 4 |
| 11.1.9 | Andere, bv. IRT-methodologie, (quasi-)experimentele designs (beschrijf): | |
| Geen informatie beschikbaar | 0 |
| Onvoldoende | 1 |
| Voldoende | 2 |
| Goed | 3 |
| Uitstekend | 4 |
| 11.1.10 | Steekproefgroottes  De richtlijnen hieronder hebben betrekking op studies binnen het klassieke testtheoriekader. Voor de schatting van item-parameters binnen de IRT-methodologie worden steekproefgroottes in de volgende gevallen als ‘voldoende groot’ beschouwd: meer dan 200 voor 1-parameter studies, meer dan 400 voor 2-parameter studies en meer dan 700 voor 3-parameter studies (gebaseerd op Parshall, Davey, Spray, & Kalohn, 2001). | |
| Geen informatie beschikbaar | 0 |
| Eén onvoldoende grote studie (bv. steekproefgrootte minder dan 100) | 1 |
| Eén voldoende grote studie (bv. steekproefgrootte van 100-200) | 2 |
| Eén grote (bv. steekproefgrootte meer dan 200) of meer dan één voldoende grote studie | 3 |
| Breed scala aan voldoende grote tot grote studies | 4 |
| 11.1.11 | Kwaliteit van de instrumenten als criteria of ijkpunten | |
|  | Geen informatie beschikbaar | 0 |
|  | Onvoldoende kwaliteit | 1 |
|  | Voldoende kwaliteit | 2 |
|  | Goede kwaliteit | 3 |
|  | Uitstekende kwaliteit met een uitgebreid scala aan relevante ijkpunten voor convergente en divergente validiteit | 4 |
| 11.1.12 | Hoe oud zijn de validiteitsstudies?  Het is moeilijk om een algemene regel te formuleren die de recentheid van het onderzoek in rekening brengen. Voor testen die bedoeld zijn om constructen te meten in een onderzoeksgebied waar belangrijke theoretische ontwikkelingen hebben plaats gevonden, kan onderzoek dat 15 jaar oud is bijna zinloos zijn, terwijl voor andere testen onderzoek dat 20 jaar oud is (of zelfs ouder) nog steeds relevant kan zijn. | |
|  | Aantal jaren | Klik of tik om tekst in te voeren. |
| 11.1.13 | **Begripsvaliditeit – Algemene toereikendheid**  Deze algemene beoordeling wordt bekomen door zich te baseren op de beoordelingen van de items 11.1.1 – 11.1.12. *Bereken niet simpelweg de gemiddelde scores om de globale beoordeling te verkrijgen.*  Naast de uitkomsten van het onderzoek aangaande begripsvaliditeit, dient bij de eindbeoordeling ook in rekening gebracht te worden of analysetechnieken op een correcte wijze zijn gebruikt (bv., werd het significantieniveau gecorrigeerd voor het correleren van het instrument met andere instrumenten zonder duidelijke hypotheses, het zogenoemde ‘fishing’), en of de onderzoeksteekproeven gelijkaardig zijn aan de groep(en) voor wie de test bedoeld is (zijn) (bv., meer heterogeniteit zal de correlaties opdrijven, steekproeven van studenten kunnen mogelijk resultaten geven die niet generaliseerbaar zijn). Ook moet men bij de eindbeoordeling de grootte van de onderzoeksteekproe(f)(ven), de kwaliteit van de andere gebruikte instrumenten (bv., in convergent en discriminant validiteitsonderzoek) en de actuele staat van de studies in rekening brengen. | |
| Geen informatie beschikbaar | 0 |
| Onvoldoende | 1 |
| Voldoende | 2 |
| Goed | 3 |
| Uitstekend | 4 |

## Criteriumvaliditeit

Criteriumvaliditeit (concurrente en predictieve validiteit) verwijst naar studies die de samenhang nagaan tussen de eigen meting en een andere ‘*real-world*’ criterium meting. Bewijs voor criteriumvaliditeit wordt gevonden in studies waarin metingen uit het dagelijkse leven die het criterium vormen (i.e., geen scores van andere instrumenten) gecorreleerd worden met schalen. Voorspellende studies hebben over het algemeen betrekking op situaties waarin een testafname werd uitgevoerd op een ‘kwalitatief’ verschillend tijdsmoment dan de criteriummeting – bv. voor een werkgerelateerde selectiemeting bedoeld om job succes te voorspellen, zou het instrument op het ogenblik van selectie moeten worden afgenomen - veeleer dan dat het een kwestie is van hoe lang het tijdsinterval was tussen de instrument- en criteriummeting.

Studies kunnen ook retrospectief/postdictief een voorspellende waarde hebben, dit kan bijvoorbeeld het geval zijn wanneer scores op een potentiële selectietest gecorreleerd zijn met de vroegere prestatiebeoordelingen van personeelsleden door de werkgever. Eigenlijk is bewijs voor criteriumvaliditeit vereist voor allerlei soorten testen. Wanneer echter de handleiding expliciet vermeldt dat de test geen voorspellende doeleinden heeft (zoals testen voor het meten van schoolvorderingen), dan kan criteriumvaliditeit als ‘niet van toepassing’ beschouwd worden.

|  |  |  |
| --- | --- | --- |
| **11.2** | **Criteriumvaliditeit** | |
| 11.2.1 | Type van criteriumstudie of -studies *(selecteer alles wat van toepassing is)* | |
| Predictief |  |
| Gelijktijdig |  |
| Retrospectief/postdictief |  |
| 11.2.2 | Steekproefgroottes | |
| Geen informatie beschikbaar | 0 |
| Eén onvoldoende grote studie (bv. steekproefgrootte minder dan 100) | 1 |
| Eén voldoende grote studie (bv. steekproefgrootte van 100-200) | 2 |
| Eén grote (bv. steekproefgrootte meer dan 200) of meer dan één voldoende grote studie | 3 |
| Breed scala aan voldoende grote tot grote studies | 4 |
| 11.2.3 | Kwaliteit van de criteriummetingen | |
| Geen informatie beschikbaar | 0 |
| Onvoldoende kwaliteit | 1 |
| Voldoende kwaliteit | 2 |
| Goede kwaliteit | 3 |
| Uitstekende kwaliteit met betrekking tot betrouwbaarheid en representatie van het criterium construct | 4 |
| 11.2.4 | Sterkte van de relatie tussen de test en de criteria  Het is moeilijk om duidelijke criteria op te stellen voor het beoordelen van de grootte van de criteriumvaliditeitscoëfficiënten van een instrument. Een criterium-validiteitscoëfficiënt van 0.20 kan behoorlijk bruikbaar zijn in sommige situaties, terwijl een coëfficiënt van 0.40 van weinig belang kan zijn voor andere. Een coëfficiënt van 0.30 kan in het geval van een personeelsselectie als relatief bruikbaar beschouwd worden, terwijl in onderwijssituaties hogere coëfficiënten gebruikelijk zijn.  Daarom moeten beoordelingen gebaseerd zijn op het specifieke oordeel en de expertise van de beoordelaar en niet door simpelweg de gemiddelde scores te berekenen van reeksen van correlatiecoëfficiënten. Deze richtlijnen zijn gebaseerd op Hemphill (2003; zie ook Meyer et al., 2001) en betreffen correlaties die niet gecorrigeerd zijn voor verzwakking in noch de predictor, noch het criterium. Nochtans, coëfficiënten kunnen gecorrigeerd zijn voor beperkingen in de waarden.  De hieronder gegeven waarden hebben betrekking op validiteitscoëfficiënten, omdat *correlaties* tussen testen en criteria de meest gebruikte manier is om criteriumvaliditeit voor te stellen. Gegevens over de sensitiviteit en specificiteit van een test kunnen echter meer bruikbare informatie geven over de relatie tussen een test en een criterium, voornamelijk voor gebruik in klinische situaties. Een veelgebruikte manier om de sensitiviteit en specificiteit te kwantificeren zijn ROC-curves. Swets (1988) presenteert een overzicht van waarden van ROC-curves in verschillende onderzoeksgebieden. Waarden vallen tussen .81 en .97 voor bepaalde types van een medische diagnose, tussen .70 en .95 voor leugendetectie en tussen .71 en .94 voor schoolse prestaties (slagen/falen). Deze waarden mogen als richtlijnen gebruikt worden, maar het wordt aan de expertise van de beoordelaar overgelaten om te beslissen in welke mate de test een waardevolle bijdrage kan leveren aan de beslissing in kwestie. Dit geldt eveneens indien er nog andere indicatoren gerapporteerd worden, zoals de positieve en negatieve voorspellende waarde van een test, de waarschijnlijkheidsratio, etc. | |
| Geen informatie beschikbaar | 0 |
| Onvoldoende (*r* < 0.20) | 1 |
| Voldoende (0.20 ≤ *r* < 0.35) | 2 |
| Goed (0.35 ≤ *r* < 0.50) | 3 |
| Uitstekend (*r* ≥ 0.50) | 4 |
| 11.2.5 | Hoe oud zijn de validiteitsstudies?  Het formuleren van een algemene regel om de actuele staat van het onderzoek in rekening te brengen is moeilijk. Voor testen die bedoeld zijn om gedrag te voorspellen in snel wijzigende omgevingen kan onderzoek dat 15 jaar oud is mogelijk gedateerd en dus bijna zinloos zijn, terwijl voor andere testen onderzoek dat 20 jaar oud is (of zelfs ouder) nog steeds relevant kan zijn. | |
| Aantal jaren | Klik of tik om tekst in te voeren. |
| 11.2.6 | **Criteriumvaliditeit – algemene toereikendheid**  Deze algemene beoordeling wordt verkregen door zich te baseren op de beoordelingen van de items 11.2.1 – 11.2.5. *Bereken niet simpelweg de gemiddelde scores om de globale beoordeling te verkrijgen.*  Afgezonderd van de resultaten van het onderzoek naar criteriumvaliditeit, dient bij de eindbeoordeling ook in rekening gebracht te worden of procedures en analysetechnieken op een correcte wijze zijn gebruikt (bv. controle op de aanwezigheid van criteriumcontaminatie, correctie voor verzwakking, cross-validatie) en of de onderzoeksteekproeven gelijkaardig zijn aan de groep(en) waarvoor de test bedoeld is (bv. correctie voor beperking in de range). Ook moet men bij de eindbeoordeling de grootte van de onderzoekssteekproe(f)(ven), de kwaliteit van de gebruikte criterium instrumenten (bv. is er tekort aan criteria) en de actuele staat van de studies in rekening brengen. | |
| Geen informatie beschikbaar | 0 |
| Onvoldoende | 1 |
| Voldoende | 2 |
| Goed | 3 |
| Uitstekend | 4 |

## Algemene validiteit

Bij de beoordeling van de algemene validiteit is het belangrijk om stil te staan bij het belang van begripsvaliditeit als de beste indicator om na te gaan of een test daadwerkelijk meet wat deze beweert te meten. In sommige gevallen zou het grootste bewijs hiervoor kunnen liggen in criteriumonderzoek. Een dergelijke test kan mogelijk een ‘voldoende’ of betere beoordeling hebben voor criteriumvaliditeit en een ‘onvoldoende’ voor begripsvaliditeit. Over het algemeen zal de beoordeling van de Algemene Validiteit gelijk zijn aan ofwel de Begripsvaliditeit ofwel de Criteriumvaliditeit, afhankelijk van welke de grootste is. Echter kan, afhankelijk van de specifieke doelstelling van de test, het ene type als meer relevant dan het andere type beschouwd worden. De beoordeling van de Algemene Validiteit mag niet beschouwd worden als een gemiddelde of als de laagste gemene deler.

|  |  |  |
| --- | --- | --- |
| 11.3 | **Validiteit – algemene toereikendheid**  Deze algemene beoordeling wordt verkregen door zich te baseren op de beoordelingen van de items 11.1.1 – 11.2.6. *Bereken niet simpelweg de gemiddelde scores om de globale beoordeling te verkrijgen.* | |
| Geen informatie beschikbaar | 0 |
| Onvoldoende | 1 |
| Voldoende | 2 |
| Goed | 3 |
| Uitstekend | 4 |

|  |
| --- |
| **Opmerkingen van de beoordelaar over validiteit** (inclusief al het bewijs voor validiteit). Opmerkingen over equivalentie/generalisatie van validiteit dienen hier evenwel gemaakt te worden (indien van toepassing). |
| Klik of tik om tekst in te voeren. |

# Kwaliteit van computergegenereerde rapporten

Het beoordelen van computergegenereerde rapporten wordt bemoeilijkt door het feit dat vele leveranciers begrijpelijkerwijs hun intellectuele eigendomsrechten op algoritmes en scoringsregels wensen te beschermen. In de praktijk dient de technische handleiding voor beoordelingsdoeleinden voldoende informatie te bevatten waarin de ontwikkeling van het rapporteringsproces en de basisprincipes nauwkeurig worden omschreven en onderbouwd via een voorbeeld van testcasussen van scoreconfiguraties. Idealiter beschrijft de documentatie ook de procedures die werden gebruikt om het genereren van specifieke rapporten te testen op accuraatheid, consistentie en relevantie. Voor de beoordelingsdoeleinden dienen minimum drie rapporten gebaseerd op verschillende scoreprofielen, inclusief ruwe scores, voorzien te zijn, ook al zijn de algoritmes voor het genereren van de rapporten confidentieel.

Voor elk van de volgende eigenschappen werden een aantal vragen gesteld die zouden kunnen helpen bij het maken van een oordeel. Verder is een definitie van een ‘uitstekende’ (4) beoordeling gegeven.

|  |  |  |
| --- | --- | --- |
| **Items die beoordeeld moeten worden met n.v.t. of 0 tot 4, criteria zijn voorzien voor een ‘uitstekende’ (4) beoordeling.** | | |
| 12.1 | **Reikwijdte of specificiteit (dekking)**  Rapporten kunnen variëren in zowel hun reikwijdte als hun specificiteit. Daarnaast kunnen rapporten mogelijk ook variëren in de range van mensen voor wie ze geschikt zijn. In sommige gevallen is het mogelijk dat afzonderlijke op maat gemaakte rapporten voorzien worden voor verschillende groepen van ontvangers.   * *Dekt het rapport de uitgebreide range aan kenmerken gemeten door het instrument?* * *Is de mate van specificiteit verantwoord op basis van het detailleringsniveau dat in de scores van het instrument voorhanden is?* * *Kan de ‘granulariteit’ van het rapport (i.e. het aantal afzonderlijke scoreverdelingen op een schaal die gebruikt worden om de verschillende teksteenheden uit het rapport in kaart te brengen) verantwoord worden in termen van de meetfouten van de schalen?* * *Is het rapport ontworpen voor dezelfde populaties waarvoor het instrument was ontwikkeld? (bv. groepen waarvoor de normgroepen relevant zijn, of waarvoor er relevante criteriumgegevens zijn, etc.).* | |
| Geen informatie beschikbaar | 0 |
| Onvoldoende | 1 |
| Voldoende | 2 |
| Goed | 3 |
| Uitstekend: uitstekende afstemming tussen de reikwijdte van het instrument en die van het rapport, met een koppeling van de specificiteitsgraad van het rapport aan het detailleringsniveau gemeten door de schalen. Er is goed gebruik gemaakt van alle scores van het instrument. | 4 |
| 12.2 | **Betrouwbaarheid**   * *Hoe consistent zijn de rapporten in hun interpretatie van gelijkaardige reeksen van scoregegevens?* * *Indien de rapportinhoud gevarieerd is (bv. door willekeurige selectie van gelijkaardige teksteenheden), wordt dit dan op een gepaste manier gedaan?* * *Is de interpretatie van de scores en de verschillen tussen de scores gerechtvaardigd in termen van meetfouten van de schalen?* | |
| Geen informatie beschikbaar | 0 |
| Onvoldoende | 1 |
| Voldoende | 2 |
| Goed | 3 |
| Uitstekend: Uitstekende consistentie in de interpretatie en er werden gepaste waarschuwingen voorzien inzake de onderliggende meetfouten van stellingen, interpretatie en aanbevelingen. | 4 |
| 12.3 | **Relevantie of validiteit**  Het verband tussen het instrument en de inhoud van het rapport kan ofwel in het rapport zelf uitgelegd worden, ofwel apart gedocumenteerd worden. Bij rapporten die gebaseerd zijn op een klinisch oordeel zou het proces waarin de expert(en) gekomen zijn tot het creëren van de inhoud en de regels die de scores aan de inhoud relateren, moeten worden gedocumenteerd.   * *Hoe sterk is de relatie tussen de inhoud van het rapport en de scores op het instrument? In welke mate gaat het rapport verder dan of wijkt het af van de informatie die teruggevonden kan worden in de scores van het instrument?* * *Relateert de inhoud van het rapport duidelijk aan de kenmerken gemeten door het instrument?* * *Biedt het redelijke gevolgtrekkingen over criteria waarvan we kunnen verwachten dat ze gerelateerd zijn aan dergelijke kenmerken?* * *Welk empirisch bewijs bestaat er om aan te tonen dat deze relaties echt bestaan?*   Het is relevant om zowel de begripsvaliditeit van een rapport (i.e. de mate waarin voorzien wordt in een interpretatie die in lijn ligt met de definitie van de onderliggende constructen) als de criteriumvaliditeit (i.e. de mate waarin de gegeven verklaringen teruggebracht kunnen worden naar de empirische gegevens) in beschouwing te nemen. | |
| Geen informatie beschikbaar | 0 |
| Onvoldoende | 1 |
| Voldoende | 2 |
| Goed | 3 |
| Uitstekend: Een uitstekende verantwoording is voorzien voor het verband tussen de schalen en de inhoud van het rapport. | 4 |
| 12.4 | **Vrijheid van kwetsende inhoud, systematische afwijking of vertekening**   * *Is het mogelijk dat de inhoud van en de gebruikte taal in het rapport de indruk nalaten misplaatst te zijn voor bepaalde groepen?* * *Geeft het rapport duidelijk aan waar mogelijke afwijkingen of vertekeningen in de resultaten van het instrument kunnen voorkomen?* * *Zijn er alternatieve taalformulieren beschikbaar gesteld? Indien ja, zijn er voldoende stappen gezet om hun equivalentie te waarborgen?* | |
| Geen informatie beschikbaar | 0 |
| Onvoldoende | 1 |
| Voldoende | 2 |
| Goed | 3 |
| Uitstekend: Duidelijke waarschuwingen en toelichtingen over de mogelijke afwijkingen of vertekeningen, beschikbaar in alle relevante gebruikerstalen. | 4 |
| 12.5 | **Aanvaardbaarheid**  Dit zal in belangrijke mate afhangen van de complexiteit van de gebruikte taal in het rapport, de complexiteit van de beschreven constructen en het doeleinde waarvoor het is bedoeld.   * *Is het waarschijnlijk dat de vormgeving en inhoud van het rapport aanvaardbaar is voor de bedoelde ontvangers?* * *Is het rapport geschreven in een taal die aangepast is aan de mogelijke niveaus van rekenvaardigheid en geletterdheid van de bedoelde lezer?* | |
| Geen informatie beschikbaar | 0 |
| Onvoldoende | 1 |
| Voldoende | 2 |
| Goed | 3 |
| Uitstekend: zeer hoge aanvaardbaarheid, het is uitstekend vormgegeven en uitermate geschikt voor het doelpubliek. | 4 |
| 12.6 | **Lengte**  De eigenschap ‘*Lengte*’ hoort ook onder ‘*Praktische Informatie*’ en onder dit luik zou hierover in de beoordeling gereflecteerd moeten worden. Maar omdat te uitgebreide en te lange rapporten mogelijk ook een indicatie kunnen zijn van overinterpretatie van scores, wordt de lengte van de rapporten ook apart beoordeeld. In het algemeen wordt aangenomen dat rapporten die gemiddeld langer zijn dan één pagina per schaal (exclusief titel pagina’s, copyright mededelingen, etc.) te lang en dus overgeïnterpreteerd zijn. | |
| Geen informatie beschikbaar | 0 |
| Onvoldoende | 1 |
| Voldoende | 2 |
| Goed | 3 |
| Uitstekend | 4 |
| 12.7 | **Algemene toereikendheid van computergegenereerde rapporten**  Deze algemene beoordeling wordt verkregen door zich te baseren op de beoordelingen van de items 12.1 – 12.6. *Bereken niet simpelweg de gemiddelde scores om de globale beoordeling te verkrijgen.* | |
| Geen informatie beschikbaar | 0 |
| Onvoldoende | 1 |
| Voldoende | 2 |
| Goed | 3 |
| Uitstekend | 4 |

|  |
| --- |
| **Opmerkingen van de beoordelaar over computergegenereerde rapporten**  In de evaluatie kan eveneens het licht geworpen worden op aanvullende zaken zoals de vraag of rapporten al dan niet beschikken over ingebouwde controlesystemen om de consistentie in antwoordpatronen na te gaan. Verder kan het ook gaan over bv. het opnemen van metingen van antwoordafwijkingen of -vertekeningen (bv. metingen van de centrale neiging in beoordelingen) en andere indicatoren van het vertrouwen waarmee de scores van de persoon kunnen worden geïnterpreteerd.  Bovendien kunnen opmerkingen over de complexiteit van de algoritmes ook toegevoegd worden, bv. wordt er over verschillende schalen tegelijk nagedacht, hoe wordt er omgegaan met schaalprofielen, etc. Dergelijke complexiteit dient uiteraard ondersteund te worden door helder geformuleerde basisprincipes in de handleiding. |
| Klik of tik om tekst in te voeren. |

# Eindevaluatie

|  |  |  |
| --- | --- | --- |
| **Evaluatierapport van de test**  Het evaluatierapport moet een beknopt en duidelijk beargumenteerd oordeel over de test bevatten. Verder dient het ook de pro’s en contra’s te beschrijven, algemene aanbevelingen te doen over hoe en wanneer het gebruikt kan worden en (waar nodig) waarschuwingen te geven over wanneer het niet gebruikt dient te worden.  Een samenvatting van alle positieve en negatieve punten van de aangepaste en vertaalde testen dient hier opgenomen te worden. Een checklist van belangrijke overwegingen omtrent dergelijke instrumenten is toegevoegd in de Appendix als een geheugensteuntje van de opmerkingen in de relevante onderdelen. Becommentarieer deze enkel indien dit aangewezen is.  De evaluatie zou onderwerpen moeten omvatten zoals de geschiktheid van het instrument voor verschillende beoordelingsfuncties of toepassingsgebieden; alle speciale trainingsbehoeften of vaardigheden die noodzakelijk zijn; of de trainingsvereisten op het juiste niveau afgesteld werden; gebruiksgemak; de kwaliteit en kwantiteit van de informatie die voorzien werd door de leverancier; of er belangrijke informatie is die niet aan de gebruikers aangeleverd wordt en of er zaken tevoorschijn komen wanneer het instrument vertaald of aangepast is (zie Appendix).  Voeg opmerkingen toe over relevante onderzoeken die lopende zijn, over de plannen van de leveranciers voor toekomstige ontwikkelingen en verfijningen, etc. | | |
| Klik of tik om tekst in te voeren. | | |
| Conclusies | | |
| Klik of tik om tekst in te voeren. | | |
| **Aanbevelingen** (*selecteer er één)*  De toepasselijke aanbeveling zou, op basis van de lijst hiernaast, moeten worden aangeduid. Dit vereist gewoonlijk enig commentaar, verantwoording of beschrijving. Verder dient een korte uiteenzetting te worden toegevoegd over de situaties waarvoor en wijze waarop het instrument gebruikt kan worden, alsook waarschuwingen over mogelijke gebieden van verkeerd gebruik .  **Alle kenmerken hieronder zouden beoordelingen moeten hebben van ofwel n.v.t., 2, 3 of 4 indien een instrument ‘aanbevolen’ wordt voor algemeen gebruik (aanbeveling 4 of 5).**  9 Normering  10 Betrouwbaarheid-algemeen  11 Validiteit-algemeen  12 Computergegenereerde rapporten  Indien één van deze beoordelingen 0 of 1 is, dan zal het instrument gewoonlijk worden geclassificeerd onder de Aanbeveling 1, 2 of 3 of het zal worden geclassificeerd onder ‘Andere’ samen met een gepaste verklaring. | 1 Vereist verdere ontwikkeling. Enkel geschikt voor gebruik in onderzoek, niet voor gebruik in de praktijk. |  |
| 2 Enkel geschikt voor gebruik door een deskundige testleider (hoger dan EFPA User Qualification Level 2) onder nauwkeurig gecontroleerde condities of in zeer gelimiteerde toepassings-gebieden. |  |
| 3 Geschikt voor gesuperviseerd gebruik in toepassingsgebieden(en) gedefinieerd door de verdeler, door elke testleider met een algemene competentie in testgebruik en testafname (hoger dan EFPA User Qualification Level 2). |  |
| 4 Geschikt voor gebruik in toepassingsgebieden(en) gedefinieerd door de verdeler, door testleiders die voldoen aan de specifieke kwalificatievereisten van de verdeler (ten minste EFPA User Qualification Level 2) |  |
| 5 Geschikt voor niet-gesuperviseerde zelfbeoordeling in toepassingsgebieden(en) gedefinieerd door de verdeler |  |
| 6 Andere Klik of tik om tekst in te voeren. |  |

DEEL 3: BIBLIOGRAFIE

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

Bartram, D. (1996). Test qualifications and test use in the UK: The competence approach. *European Journal of Psychological Assessment*, *12*, 62–71.

Bartram, D. (2002a). *EFPA Review Model for the description and evaluation of psychological instruments: Version 3.2. Evaluation Form.* Brussels: EFPA Standing Committee on Tests and Testing (September, 2002).

Bartram, D. (2002b). *EFPA Review Model for the description and evaluation of psychological instruments: Version 3.2. Notes for Reviewers*. Brussels: EFPA Standing Committee on Tests and Testing (September, 2002).

Bartram, D., & Hambleton, R. K. (Eds.) (2006). *Computer-based testing and the Internet*. Chichester, UK: Wiley and Sons.

Bartram, D., Lindley, P. A., & Foster, J. M. (1990). *A review of psychometric tests for assessment in vocational training*. Sheffield, UK: The Training Agency.

Bartram, D., Lindley, P. A., & Foster, J. M. (1992). *Review of psychometric tests for assessment in vocational training*. BPS Books: Leicester.

Bechger, T., Hemker, B., & Maris, G. (2009). *Over het gebruik van continue normering* [On the use of continuous norming]. Arnhem, The Netherlands: Cito.

Bennett, R. E. (2006). Inexorable and inevitable: The continuing story of technology and assessment. In D. Bartram & R. K. Hambleton (Eds.), *Computer-based testing and the Internet* (pp. 201-217). Chichester, UK: Wiley and Sons.

Brennan, R. L. (Ed.) (2006). *Educational measurement*. Westport, CT: ACE/Praeger.

Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). New York: Harper & Row.

Downing, S. M., & Haladyna, T. M. (Eds.) (2006). *Handbook of test development.* Hillsdale, NJ: Erlbaum.

Drasgow, F., Luecht, R. M., & Bennett, R. E. (2006). Technology and testing. In R. L. Brennan (Ed.), *Educational measurement* (pp. 471-515).Westport, CT: ACE/Praeger.

Drenth, P. J. D., & Sijtsma, K. (2006). *Testtheorie. Inleiding in de theorie van de psychologische test en zijn toepassingen* (4e herziene druk) [Test theory. Introduction in the theory and application of psychological tests (4th revised ed.)]. Houten, The Netherlands: Bohn Stafleu van Loghum.

Embretson, S. E. (Ed.) (2010*). Measuring psychological constructs. Advances in model-based approaches.* Washington, D. C.: American Psychological Association.

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

Evers, A. (2001a). Improving test quality in the Netherlands: Results of 18 years of test ratings. *International Journal of Testing*, *1*, 137–153.

Evers, A. (2001b). The revised Dutch rating system for test quality. *International Journal of Testing*, *1*, 155–182.

Evers, A., Braak, M., Frima, R., & van Vliet-Mulder, J. C. (2009-2012). *Documentatie van Tests en Testresearch in Nederland* [Documentation of Tests and Testresearch in The Netherlands]. Amsterdam: Boom test uitgevers.

Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2010). *COTAN Beoordelingssysteem voor de Kwaliteit van Tests (geheel herziene versie; gewijzigde herdruk)* [COTAN Rating system for test quality (completely revised edition; revised reprint)]. Amsterdam: NIP.

Evers, A., Muñiz, J., Bartram, D., Boben, D., Egeland, J., Fernández-Hermida, J. R., et al. (2012). Testing practices in the 21st Century: Developments and European psychologists’ opinions. *European Psychologist*, in press.

Evers, A., Sijtsma, K., Lucassen, W., & Meijer, R. R. (2010). The Dutch review process for evaluating the quality of psychological tests: History, procedure and results. *International Journal of Testing, 10,* 295-317*.*

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*, 309-334.

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

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

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist, 58*, 78-80.

International Test Commission. (2005). *International Guidelines on Computer-Based and Internet Delivered Testing*. Bruxelles, Belgium: Author.

Kersting, M. (2008). DIN Screen, Version 2. Leitfaden zur Kontrolle und Optimierung der Qualität von Ver-fahren und deren Einsatz bei beruflichen Eignungsbeurteilungen [DIN Screen, Version 2. Guide line for monitoring and optimizing the quality of instruments and their application in proficiency assessment procedures.]. In M. Kersting. *Qualitätssicherung in der Diagnostik und Personalauswahl - der DIN Ansatz* (S. 141-210) [Guaranteeing quality in diagnostics and personnel selection (p. 141-210)]. Göttingen: Hogrefe.

Lindley, P. A. (2009). *Reviewing translated and adapted tests: Notes and checklist for reviewers:5 May 2009*. Leicester, UK: British Psychological Society. Retrieved from <http://www.efpa.eu/professional-development/tests-and-testing>.

Lindley, P.A. (2009, July). Using EFPA Criteria as a common standard to review tests and instruments in different countries. In D.Bartram (Chair), *National approaches to test quality assurance*. Symposium conducted at The 11th European Congress of Psychology, Oslo, Norway.

Lindley, P., Bartram, D., & Kennedy, N. (2004). *EFPA Review Model for the description and evaluation of psychological tests: test review form and notes for reviewers: Version 3.3.* Leicester, UK: British Psychological Society (November, 2004).

Lindley, P., Bartram, D., & Kennedy, N. (2005). *EFPA Review Model for the description and evaluation of psychological tests: test review form and notes for reviewers: Version 3.41.* Brussels: EFPA Standing Committee on Tests and Testing (August, 2005).

Lindley, P., Bartram, D., & Kennedy, N. (2008). *EFPA Review Model for the description and evaluation of psychological tests: test review form and notes for reviewers: Version 3.42.* Brussels: EFPA Standing Committee on Tests and Testing (September, 2008).

Lindley, P. A. (Senior Editor), Cooper, J., Robertson, I., Smith, M., & Waters, S. (Consulting Editors). (2001). *Review of personality assessment instruments (Level B) for use in occupational settings. 2nd Edition.* Leicester, UK: BPS Books.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., et al. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56*, 128-165.

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

Moosbrugger, H., Kelava, A., Hagemeister, C., Kersting, M., Lang, F., Reimann, G., et al. (2009, July).

The German Test Review System (TBS-TK) and first experiences. In D. Bartram (Chair), *National approaches to test quality assurance*. Symposium conducted at The 11th European Congress of Psychology, Oslo, Norway.

Moreno, R., Martínez, R. J., & Muñiz, J. (2006). New guidelines for developing multiple-choice ítems. *Methodology, 2*, 65-72.

Muñiz, J., & Bartram, D. (2007). Improving international tests and testing. *European Psychologist*, *12*, 206-219.

Nielsen, S. L. (2009, July). Test certification through DNV in Norway. In D. Bartram (Chair), *National approaches to test quality assurance*. Symposium conducted at The 11th European Congress of Psychology, Oslo, Norway.

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

Parshall, C. G., Spray, J. A., Davey, T., & Kalohn, J. (2001). *Practical Considerations in Computer-based Testing.* New York: Springer Verlag.

Prieto, G., & Muñiz, J. (2000). Un modelo para evaluar la calidad de los tests utilizados en España [A model for the evaluation of test quality in Spain]. *Papeles del Psicólogo*, *77*, 65–71.

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

Tideman, E. (2007). Psychological tests and testing in Sweden. *Testing International*, *17*(June), 5–7.

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

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

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

Testkuratorium. (2009). TBS-TK. Testbeurteilungssystem des Testkuratoriums der Föderation Deutscher Psychologenvereinigungen. Revidierte Fassung vom 09. September 2009 [TBS-TK. Test review system of the board of testing of the Federation of German psychologists’ associations]. *Report Psy-chologie, 34,* 470-478.

Van de Vijver, F. J. R., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Erlbaum.

Van der Linden, W. J., & Glas, C. A. W. (Eds.) (2010). *Elements of adaptive testing*. London: Springer.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.

Ziegler, M., MacCann, C., & Roberts, R. (Eds.) (2011). *New perspectives on faking in personality assessment*. Oxford, UK: Oxford University Press.

APPENDIX   
  
Een geheugensteun met essentiële punten voor evaluatie indien een instrument vertaald en/of aangepast werd van een niet-lokale context.

|  |  |
| --- | --- |
| **Ontwikkeling** | |
| Bewijs of discussie van | *Input van ‘native speakers’ van de nieuwe taal* |
| *Verscheidene beoordelingen door zowel taal- als (test)inhoudsexperten* |
| *Terugvertaling van de nieuwe taal in de originele taal* |
| Basis psychometrische kenmerken | *Item prestaties* |
| *Betrouwbaarheid* |
| **Normen** | |
|  | *Een lokale norm werd voorzien* |
| Niet-lokale normen | *Sterk bewijs voor equivalentie voor zowel de testversies als de steekproeven* |
| Internationale normen | *Groter dan de normale vereisten voor lokale steekproeven* |
| De aard van de steekproef | *Evenwicht in de bronnen van de steekproef* |
| *Equivalentie van de achtergrond van de verschillende onderdelen van de steekproef* |
| Het type meting | *Weinig of geen verbale inhoud* |
| De equivalentie van de testversie | *Alle taalversies zijn goed vertaald/aangepast* |
| *Sommige groepen hebben de test afgerond in een niet-primaire taal* |
| Gelijkenissen van scores in verschillende steekproeven | *Wanneer er grote verschillen zijn, moeten deze, alsook de implicaties ervan voor het gebruik, bediscussieerd worden* |
| Richtlijnen aangaande de veralgemening van de normen |  |
| **Equivalentie/Betrouwbaarheid/Validiteit** | |
| Invariantie in constructstructuur | *Aan de hand van factorstructuur, equivalentie van correlatiematrices of gelijkenis in correlatiepatronen met standaardmaten* |
| Gelijkaardig criteriumvaliditeit | *Sterkste correlatie met gelijkaardige competenties* |
| Gelijkaardige patronen van schaalladingen | *Items correleren in hetzelfde patroon met andere schalen* |
| *Sterkste/zwakste itemladingen zijn gelijk in de originele en nieuwe talen* |
| Alternerende vormbetrouwbaarheid | *Tweetalige kandidaten hebben gelijkaardige profielen in de twee talen* |
| **Generalisatie van de validiteit** | |
| Validiteitsgeneralisatie vereist sterk bewijs | *Bij het vertalen van testen over linguïstische families (bv. van een Indo-Europese naar een Semitische taal)* |
| Validiteitsgeneralisatie kan afgeleid worden | *Wanneer een test werd vertaald in verschillende talen, dan kan enige validiteitsveralgemening afgeleid worden uit bewijs van validiteitsinvariantie in vorige vertalingen: een Zweedse test is reeds vertaald geweest in het Frans, Duits en Italiaans en er werd equivalentie van de test in deze talen aangetoond* |

1. Binnen deze tekst wordt de term ‘testleider’ gebruikt om de persoon aan te duiden die de test afneemt. [↑](#footnote-ref-2)
2. Elk land kan kiezen om ofwel de namen van de beoordelaars te publiceren of om te opteren voor anonieme beoordeling. [↑](#footnote-ref-3)
3. Deze informatie dient ingevuld te worden door de redacteur of de administratie. [↑](#footnote-ref-4)
4. Een *computer adaptieve test* (CAT) is een computergebaseerde test waarbij de moeilijkheidsgraad van de vragen aangepast wordt aan de capaciteiten van de gebruiker van de test (n.v.d.v). [↑](#footnote-ref-5)