Principes van de psychometrie

Hoe beoordeel je de kwaliteit van psychologische testen?

Tim Vantilborgh

acco learn

Inhoud

1

Psychometrie

LEERDOELEN

Na het bestuderen van dit hoofdstuk:

- Begrijp je het belang van psychologische testen voor psychologen en in het dagelijkse leven;
- Ken je de definitie van psychologische testen;
- Begrijp je de gevolgen van de definitie van psychologische testen;
- Heb je inzicht in de geschiedenis van psychologische testen.

We zijn ons er niet altijd van bewust, maar psychologische testen spelen een grote rol in ons dagelijkse leven. Of je nu een student, leraar, psycholoog, arts, patiënt, advocaat, of manager bent, de kans is groot dat je in aanraking komt met psychologische testen. In de loop van je hele leven leg je op verschillende momenten psychologische testen af. En als psycholoog zul je tijdens je carrière regelmatig psychologische testen van anderen afnemen, bijvoorbeeld om diagnoses te stellen. Psychologische testen spelen ook een grote rol in wetenschappelijk onderzoek. Ze worden gebruikt om attituden, gedrag, en gevoelens van participanten te meten in studies. Via statistische analyses worden die data geanalyseerd om hypotheses te toetsen. Zonder psychologische testen zou heel wat onderzoek dan ook niet mogelijk zijn.

Van bij de geboorte worden mensen in allerhande situaties getest. De eerste test die wellicht iedereen ondergaat, is de Apgartest (ontwikkeld door Virginia Apgar in 1952). Daarbij worden pasgeboren baby's geëvalueerd op vijf kenmerken (kwaliteit van de ademhaling, aanwezigheid van de reflexen, spiertonus, pols, en kleur). Elk kenmerk krijgt vervolgens een score (2 = goed, 0 = slecht). Aan de hand van de testscore (= de som van de scores op de vijf kenmerken) wordt nagegaan of extra medische aandacht nodig is. Een score van 7 of meer is een indicatie van goede gezondheid van het kind. Op latere leeftijd komt een peuter die vroeger een lage Apgarscore had mogelijk in aanmerking voor een onderzoek naar een ontwikkelingsstoornis. Een kleuter moet misschien een schoolrijpheidstest afleggen. Eens op school wachten er vele testen – denk maar aan examens – die je moet ondergaan

Virginia Apgar

vooraleer je afstudeert. En met een diploma op zak wachten weer andere testen (bv. in sollicitatieprocedures of bij medisch onderzoek). In scholen, universiteiten, de zakenwereld, de industrie, de welzijnszorg enzovoort worden veel beslissingen gebaseerd op de kenmerken van mensen, meer bepaald de kenmerken waarin mensen onderling van elkaar verschillen. Die kenmerken worden veelal gemeten aan de hand van psychologische testen. Zoals we in dit hoofdstuk zullen zien is de rol van psychologische testen voor het nemen van beslissingen in de loop van de geschiedenis alsmaar belangrijker geworden.

	KENMERK	0 PUNTEN	1 PUNT	2 PUNTEN
A	Activiteit (spiertonus)	Afwezig	Armen en benen gebogen	Actieve beweging
P	Pols	Afwezig	< 100 slagen / minuut	> 100 slagen / minuut
G	Grimas (reflex responsiviteit)	Geen reactie	Grimas	Niezen; proesten; wegtrekken
A	Aanblik (huidkleur)	Blauwgrijs; helemaal bleek	Normaal, behalve aan de extremiteiten	Normaal over hele lichaam
R	Respiratie (ademhaling)	Afwezig	Traag; onregelmatig	Goed; huilen

Figuur 1.1 Virginia Apgar ontwikkelde de Apgartest (bron: By March of Dimes Public Domain, https://commons.wikimedia.org/w/index.php?curid=43770603).

Gelet op het belang van psychologische testen, is het belangrijk dat men goede van slechte testen kan onderscheiden. De studie van de kwaliteit van psychologische testen vormt het onderwerp van de psychometrie. In dit hoofdstuk zullen we enkele basisconcepten binnen de psychometrie toelichten. We lichten kort toe wat het belang is van psychologische testen in de huidige maatschappij en voor psychologen, en bespreken daarna wat psychologische testen juist zijn. Vervolgens geven we een beknopt overzicht van de geschiedenis van psychologische testen.

DEFINITIE

Psychometrie vormt de wetenschappelijk studie van (de kwaliteit van) psychologische testen (Rust & Golombok, 2014). De focus ligt daarbij op psychologische testen en assessment, maar psychometrie is ook toepasbaar op andere domeinen zoals criminologie, economie, of educatiewetenschappen.

1.1 Het belang van psychologische testen

Psychologische testen worden dagelijks gebruikt. De beslissingen die genomen worden op basis van dergelijke testen, kunnen een grote impact hebben op mensen. Omdat die impact zo groot kan zijn, is het belangrijk dat je de basisprincipes van psychologische metingen begrijpt. Of je nu later als psycholoog of als onderzoeker aan de slag gaat – maar ook in je dagelijkse leven – je zult altijd geconfronteerd worden met psychologische testen.

1.1.1 VOOR PSYCHOLOGEN

Misschien overweeg je wel een carrière waarin je zelf psychologische testen zult afnemen in de praktijk, bijvoorbeeld als klinisch psycholoog of arbeidspsycholoog. Daarbij zul je beslissingen nemen, gebaseerd op resultaten van psychologische testen. Voor klinisch psychologen vormen psychologische testen bijvoorbeeld een belangrijk instrument om diagnoses te stellen. Om te bepalen of een persoon een

persoonlijkheidsstoornis heeft zal een psycholoog bijvoorbeeld psychologische testen afnemen, in combinatie met andere technieken (bv. interviewdata). Daarbij is het nuttig om het onderscheid te maken tussen *testing* en *assessment*. Assessment is een brede term en omvat zowel het gebruik van psychologische testen als andere technieken om beslissingen te nemen (bv. interviews en observatie). Als meerdere testen en technieken gecombineerd worden om een beslissing te nemen over een persoon (bv. om een diagnose te stellen), dan spreken we van assessment. Testing daarentegen gaat over het gebruik van één psychologische test om een beslissing te nemen (Coaley, 2014). We focussen in dit handboek op testing, en niet zozeer op de combinatie van verschillende bronnen van informatie voor assessment.

Een ander voorbeeld uit de praktijk is een arbeidspsycholoog die moet beslissen of iemand een geschikte kandidaat is voor een vacante functie. In die situatie zal een arbeidspsycholoog veelal gebruikmaken van psychologische testen die de persoonlijkheid of de integriteit van sollicitanten meten. Men verwacht bijvoorbeeld dat iemand die hoog scoort op bepaalde persoonlijkheidstrekken, later ook beter zal presteren als werknemer. Hoe sollicitanten scoren op die persoonlijkheidstest, kan daarom bepalen of ze aangeworven worden of niet.

In de voorgaande voorbeelden neem je als psycholoog telkens beslissingen gebaseerd op de resultaten van psychologische testen. Maar als je geen goede test gebruikt, dan loop je het risico dat je verkeerde beslissingen neemt, die grote gevolgen kunnen hebben voor de patiënten, cliënten, werknemers, en anderen die getest worden. Om de resultaten van een test op een goede manier te kunnen interpreteren, is het dan ook cruciaal dat je inzicht hebt in de psychometrische kenmerken van de test.

1.1.2 VOOR ONDERZOEKERS

Ook voor wetenschappelijke onderzoekers vormen psychologische testen een belangrijk instrument. Metingen vormen altijd de basis van onderzoek, ongeacht of het onderzoek op basis van experimenten, vragenlijsten, of interviews betreft. Psychologische testen worden binnen onderzoek gebruikt om verschillen tussen personen en/of verschillen binnen personen te bestuderen. Verschillen tussen personen – bijvoorbeeld hoe twee personen verschillen van elkaar op een

persoonlijkheidstrek – noemen we interindividuele verschillen. Verschillen binnen personen – bijvoorbeeld hoe het stressniveau van één persoon verschilt op twee verschillende tijdstippen – noemen we intra-individuele verschillen.

Stel je bijvoorbeeld voor dat men een experimenteel onderzoek opzet om de effectiviteit van antidepressiva te testen. In dat onderzoek zal men moeten bepalen hoe depressief participanten zich voelen. De meest voor de hand liggende manier om dat te doen, is door de participanten een psychologische test te laten invullen die depressiviteit meet. Vervolgens kunnen we de testscores die zijn verkregen op die psychologische test, vergelijken tussen groepen van participanten die verschillende dosissen van de antidepressiva kregen. Maar als we een slechte test gebruiken om depressiviteit te meten, dan zullen de conclusies die we trekken uit dat onderzoek, waarschijnlijk fout zijn (Flake & Fried, 2020). Het is dan ook belangrijk voor onderzoekers om aandacht te besteden aan de psychometrische kenmerken van psychologische testen die ze gebruiken.

1.1.3 VOOR DE MAATSCHAPPIJ

In het dagelijkse leven worden we allemaal geconfronteerd met psychologische testen, doordat we zelf frequent psychologische testen afleggen. Als student leg je bijvoorbeeld examens af, wat een test is die gebruikt wordt om te bepalen of je bepaalde kennis of vaardigheden beheerst. Als consument vul je vragenlijsten in die peilen naar je tevredenheid met dienstverlening of producten. We maken daarnaast frequent gebruik van de resultaten van psychologische testen om zelf beslissingen te nemen. Zo ga je misschien rekening houden met de antwoorden van andere consumenten die een product beoordelen om te kiezen welk product je zal kopen. In dat opzicht vond de voorbije decennia een enorme evolutie plaats door de opkomst van internet en smartphones die een impact heeft op het gebruik van psychologische testen en psychometrie (Rust & Golombok, 2014). Zo is het tegenwoordig bijvoorbeeld niet meer noodzakelijk om mensen een persoonlijkheidstest te laten invullen om inzicht te krijgen in hun persoonlijkheid. Socialemediaplatformen, zoals Facebook, kunnen op basis van de gegevens van hun gebruikers en dankzij *machine learning*-algoritmes accurate voorspellingen doen over de persoonlijkheid van gebruikers. De gegevens die we zelf creëren op dergelijke platformen, of die verzameld worden via onze smartphone, kunnen zo gebruikt worden als een soort psychologische test. Als dergelijke gegevens

gebruikt worden om beslissingen mee te nemen – bijvoorbeeld welke berichten je te zien krijgt op een socialemediaplatform – dan is het ook belangrijk om stil te staan bij de psychometrische kwaliteit van die gegevens.

VERDIEPING

Wist je dat psychologische testen in sommige gevallen kunnen beslissen over leven en dood? Een voorbeeld daarvan is de wetgeving over de toepassing van de doodstraf in bepaalde staten van de Verenigde Staten. Zo stelt de wet in North Carolina dat de doodstraf niet mag worden uitgevoerd bij personen met een mentale achterstand. Mentale achterstand wordt binnen diezelfde wet omschreven als een intelligentieniveau dat significant onder het gemiddelde ligt. De vraag is natuurlijk vanaf wanneer intelligentie significant lager dan het gemiddelde is, en hoe men dat kan bepalen. Om die complexe vragen te beantwoorden, maakt men gebruik van psychologische tests.

Zo stelt de wet in North Carolina dat er sprake is van een intelligentieniveau dat significant onder het gemiddelde ligt als men een score van 70 of lager behaalt op een individueel afgenomen, wetenschappelijk erkende, en gestandaardiseerde intelligentietest die afgenomen werd door een erkend psychiater of psycholoog. Met andere woorden, personen met een intelligentiequotiënt (IQ) kleiner of gelijk aan 70 kunnen in North Carolina niet ter dood worden veroordeeld (*State Statutes Prohibiting the Death Penalty for People with Intellectual Disability (Pre-Atkins)*, 2023).

1.2 Wat zijn psychologische testen?

Cronbach (1960) definieerde een psychologische test als een systematische procedure om het gedrag van twee of meer personen te vergelijken. We zullen die definitie hanteren, maar breiden ze uit door te stellen dat psychologische testen ook gebruikt kunnen worden om het gedrag van een individu op verschillende momenten in de tijd te vergelijken. Met andere woorden, een psychologische test kan gebruikt worden om zowel inter- als intra-individuele verschillen te bestuderen. Een voordeel van die definitie is dat ze opgaat voor verschillende soorten testen. Sommige omschrijvingen van psychologische testen beperken zich vooral tot 'pen en papier'-testen. Bijvoorbeeld de Mini-IPIP is een psychologische test die bestaat uit twintig vragen of items die vijf persoonlijkheidstrekken meten ('extraversie', 'aangenaamheid', 'conscientieusheid', 'neuroticisme', en 'intelligentie/verbeelding') (Donnellan et al., 2006). Zoals je kunt zien in figuur 1.2, moeten personen die de test afleggen, zichzelf beoordelen. Bij elk van de twintig items moeten ze één van de beschikbare antwoordopties kiezen, om zo aan te geven in welke mate elk item een goede omschrijving vormt van zichzelf. Voor elk van de vijf persoonlijkheidstrekken wordt vervolgens een testscore berekend door de scores die bij de geselecteerde antwoorden horen, op te tellen. Die testscores geven zo voor elke persoon weer in welke mate ze extravert, aangenaam, consciëntieus, neurotisch, en intelligent/verbeeldend zijn.

Hieronder staan beweringen over menselijk gedrag. Ga na in hoeverre elke bewering op jou van toepassing is. Geef vervolgens aan hoe oneens of eens je bent met deze bewering over jezelf. Denk daarbij aan hoe je op dit moment in het algemeen bent, niet hoe je in de toekomst zou willen zijn.

Beschrijf zo eerlijk mogelijk hoe je jezelf ziet in verhouding tot anderen van jouw geslacht en leeftijd. Je antwoorden worden vertrouwelijk en anoniem bewaard, zodat je zo eerlijk mogelijk kunt antwoorden. Lees elke bewering over jezelf als-jeblieft zorgvuldig en kies dan een van de volgende antwoorden: Heel oneens, enigszins oneens, neutraal (nier oneens, niet eens), enigzins eens, heel eens.

	Heel oneens (1)	enigzins oneens (2)	Neutraal (niet oneens, niet eens) (3)	enigzins eens (4)	Heel eens (5)
Zorg voor leven in de brouwerij	○	○	○	○	○
Leef mee met de gevoelens van anderen	○	○	○	○	○
Doe klusjes meteen	○	○	○	○	○
heb regelmatig stem-mingswisselingen	○	○	○	○	○
Heb een levendige fantasie	○	○	○	○	○
Praat niet veel	○	○	○	○	○
Ben niet geïnteresseerd in de problemen van anderen	○	○	○	○	○

Figuur 1.2 Voorbeeld van een psychologische test: Enkele items uit de Mini-IPIP (Donnellan et al., 2006; Goldberg, 1992).

De Mini-IPIP is duidelijk een psychologische test, maar andere methoden om systematisch stalen van gedrag te verkrijgen kunnen ook als een psychologische test beschouwd worden. Bijvoorbeeld in een laboratoriumexperiment vraagt men soms aan participanten om te reageren op bepaalde stimuli door zo snel mogelijk op een knop te duwen. Volgens Cronbachs (1960) definitie is die meting van de snelheid waarmee een knop wordt ingedrukt eveneens een psychologische test.

Uit Cronbachs (1960) definitie van psychologische testen volgt dat testen een aantal kenmerken hebben (Gregory, 2004):

1. Ze leveren een staal van gedrag op.
2. Ze gebruiken systematische procedures.
3. Ze maken gebruik van een testscore.
4. Ze maken gebruik van normen of standaarden.
5. Ze worden gebruikt om gedrag buiten de test te voorspellen.

We zullen die kenmerken één voor één bespreken.

DEFINITIE
Een psychologische test is een systematische procedure om het gedrag van twee of meer personen te vergelijken, of om het gedrag van een persoon op verschillende momenten te vergelijken.

1.2.1 EEN STAAL VAN GEDRAG

Elke test moet worden beschouwd als een staal of steekproef van gedrag (Gregory, 2004). Aangezien elke test bestaat uit een eindig aantal items of vragen, kan een test maar een indruk geven van gedrag (of van kennis, vaardigheden ...) gebaseerd op een beperkt aantal antwoorden. Omdat een test typisch wordt gebruikt om inferenties (interpretaties, afleidingen ...) te maken op basis van een selectie van items is de keuze van die items cruciaal. Een goede test meet een representatief staal van gedragingen (met andere woorden, geeft een goed beeld van 'typisch' gedrag van de deelnemer). Veronderstel bijvoorbeeld dat een test naar intelligentie alleen

vragen over rekenen zou bevatten. In dat geval zouden deelnemers aan de test waarschijnlijk opmerken dat intelligentie meer is dan alleen rekenen en dat de intelligentietest dus geen representatief staal van intelligentie oplevert. Men kan dan ook in vraag stellen of je conclusies kunt trekken over iemands intelligentie op basis van die test. Het is dan ook een uitdaging voor testontwikkelaars om een goed evenwicht te vinden tussen de nood aan een selectie van vragen die een representatief staal van gedrag opleveren en praktische beperkingen zoals de tijdsduur van de test. Bijvoorbeeld de bedoeling van een woordenschattest is om de woordenschat van een persoon te bepalen door definities te vragen van een beperkt aantal zorgvuldig gekozen woorden (Gregory, 2004). Daarbij is men eigenlijk niet zozeer geïnteresseerd of een deelnemer de 35 woorden uit bijvoorbeeld de woordenschat subtest van de WAIS (Wechsler Adult Intelligence Scale) correct kan omschrijven. Men wil uitspraken doen over de algemene woordenschat van die persoon, niet in de prestaties van die persoon op de test zelf. Indirect geeft het aantal goed omschreven woorden in de test een beeld van de algemene woordenschat van de deelnemer.

De keuze voor items in een test is dus heel belangrijk, maar we moeten daarbij opmerken dat die items niet noodzakelijk hoeven te gelijken op het gedrag dat de test beoogt te voorspellen (Gregory, 2004). Essentieel is dat een goede test de onderzoeker in staat stelt om een bepaald gedrag van de deelnemer te voorspellen (zie 1.2.5 Voorspellen van non-testgedrag). De items in een test hoeven daarvoor niet te gelijken op het gedrag dat men wil voorspellen. Veronderstel dat uit onderzoek zou blijken dat 'nee'-antwoorden op de vraag 'ga je graag uit eten op restaurant' zou toelaten om depressieve neigingen van mensen te voorspellen, dan is dat ogenschijnlijk niet ter zake doende item toch een nuttige indicator voor depressie. Daaruit blijkt dat uitmaken of een item een goede voorspeller is van een bepaald gedrag vooral een empirische kwestie is die onderwerp is van wetenschappelijk onderzoek.

1.2.2 SYSTEMATISCHE PROCEDURES

Het gebruik van systematische procedures in psychologische tests vinden we terug op drie niveaus:

1. testontwikkeling
2. testafname
3. testscoring en interpretatie.

Op het eerste gezicht zou het kunnen lijken dat je in principe in staat zou zijn om een test te ontwikkelen op voorwaarde dat je voldoende 'afweet' van de kenmerken of de eigenschappen die je met je test wenst te meten. Maar een evident, logisch verband tussen de inhoud van een item en het te meten construct volstaat niet. Testontwikkelaars moeten items selecteren op grond van een wetenschappelijke studie van een groot aantal items (de itempool). Op basis van onderzoek moeten ze vervolgens de beste items selecteren uit die itempool, en nagaan of die geselecteerde items een test vormen die goed scoort op psychometrische kenmerken (bv. betrouwbaarheid, predictieve validiteit, testaccuraatheid). Veelal zijn meerdere studies nodig om aan te tonen dat de test goed scoort op die psychometrische kenmerken in verschillende steekproeven, en dus veralgemeend kunnen worden naar een bredere populatie. En in principe zou dat onderzoek regelmatig herhaald moeten worden om aan te tonen dat de test psychometrisch kwalitatief blijft met het verstrijken van de tijd. Zoals we in latere hoofdstukken zullen zien, is het eigenlijk nodig dat je bij elke nieuwe steekproef waarin een test wordt afgenomen, de psychometrische kwaliteit van een test nagaat, aangezien indicatoren van die kwaliteit (bv. betrouwbaarheid) afhankelijk zijn van de samenstelling van de steekproef. Het uiteindelijke doel van al dat onderzoek is altijd hetzelfde: aantonen dat de items in de test effectief datgene meten wat ze geacht worden te meten.

Systematische procedures zijn ook vereist op het vlak van de testafname. Als we mensen op basis van testscores met elkaar willen vergelijken, dan willen we zeker zijn dat verschillen in testscores te wijten zijn aan verschillen in het gemeten gedrag en niet aan verschillen in de wijze waarop de test werd afgenomen. Beeld je bijvoorbeeld eens in dat men wil nagaan wat de *digit span* van iemand is (Gregory, 2004). Een 'digit span'-test gaat na hoeveel cijfers iemand maximaal kan herhalen.

Een niet-gestandaardiseerde benadering om dat te testen zou kunnen zijn dat men instructie geeft aan de testleider om alsmaar langer wordende reeksen cijfers voor te lezen die de deelnemer dan moet herhalen. Het aantal cijfers in de langste reeks zou dan de digit span van de deelnemer vormen. Maar die instructie is vaag en het ontbreekt aan uniformiteit: verschillende onderzoekers zouden tot heel verschillende versies van die test kunnen komen. Zo is het bijvoorbeeld veel eenvoudiger om 1-2-3-4-5-8-8-8-9-9-9 of 0032-2-629-20-56 na te zeggen dan bijvoorbeeld 7-2-8-1-9-4-6-3-0-5-4. Het is daarom noodzakelijk om dergelijke proeven te standaardiseren, zodat elke onderzoeker precies dezelfde cijferreeksen gebruikt en bovendien de cijfers ongegroepeerd voorleest op een vastgelegd ritme van bijvoorbeeld 1 cijfer per seconde. Bovendien moet de onderzoeker ook weten hoe te reageren op een onverwachte vraag zoals 'kunt u dat nog eens zeggen?' (Meestal is het voorgeschreven antwoord 'nee'). Soms gaat de testontwikkelaar vrij ver in het standaardiseren van de afname procedures. Zo kan een bepaalde stemintonatie aangegeven worden, of een neutrale gelaatsuitdrukking bij het noteren van de responsen van de deelnemer. Hoewel dat wat overdreven kan lijken, kan de impact van kleine gedragingen van de testleider erg groot zijn. Door een 'bedenkelijk' gezicht te trekken, of door instemmend te knikken kan de testleider de deelnemer motiveren, angstig of zenuwachtig maken.

Voorts moeten responsen omgezet worden in cijfers: testscores. Alle tests omvatten regels of procedures die aangeven hoe een waargenomen antwoord omgezet moet worden in een cijfer of categorie. Op dat vlak zijn er sterke verschillen tussen testen. Sommige testen hanteren objectieve scoring, een proces waarbij de omzetting van antwoorden in een cijfer gebeurt door het gegeven antwoord te vergelijken met een lijst met alle mogelijke antwoorden en de daarbijhorende scores. Het gevolg daarvan is dat twee deelnemers die hetzelfde antwoord geven, automatisch dezelfde score op de test verkrijgen. In andere testen wordt subjectieve scoring toegepast; dat is een proces waarbij de persoon die de antwoorden scoort maar beschikt over een stel richtlijnen. Een voorbeeld van subjectieve scoring vinden we bij 'de inktvlekkentest' of Rorschachtest (zie figuur 1.3), waarbij de deelnemer moet zeggen wat hij of zij 'ziet' in een reeks getoonde inktvlekken. Er zijn zeer veel (misschien zelfs oneindig?) mogelijke manieren om te beschrijven wat je ziet in dergelijke inktvlek. Het is dan ook onmogelijk om al die mogelijke antwoorden in een lijst op te nemen en er scores aan te verbinden. In dat geval bestaat de antwoordsleutel uit een stel richtlijnen die aangeven hoe scores worden gegeven aan bepaalde 'typen' antwoorden. Om ervoor te zorgen dat verschillende testleiders dezelfde antwoorden verwerken tot gelijkaardige

scores, zijn zeer precieze richtlijnen noodzakelijk. Desondanks is hier meer ruimte voor verschillende interpretaties dan bij objectieve scoring. Het gevolg daarvan is dat twee deelnemers die eenzelfde antwoord geven, niet noodzakelijk dezelfde score op de test verkrijgen.

Figuur 1.3 Eerste inktvlek in de Rorschachtest (bron: https://web.archive.org/web/20070820233339/http://ar.geocities.com/test_de_rorschach/).

DEFINITIE

Objectieve scoring is een proces waarbij de omzetting van de respons in een cijfer gebeurt door de gegeven respons te vergelijken met een lijst waarin alle mogelijke responsen worden opgesomd en waar dan de bijbehorende score in kan worden opgezocht.

Subjectieve scoring is een proces waarin de persoon die de scoring uitvoert, maar beschikt over een stel richtlijnen die gehanteerd moeten worden bij het scoren van responsen.