

Is dit **assessment** kwalitatief goed genoeg?

□ Saskia Wools

Competentie assessments kunnen een selecterende, classificerende of certificerende functie hebben. Dit betekent dat assessments worden ingezet om tot beslissingen over mensen te komen met betrekking tot opleiding of werk. Deze zogenaamde 'high stake' beslissingen vereisen kwalitatief goede meetinstrumenten. Dit artikel beschrijft de ontwikkeling van een beoordelingsinstrument voor competentie assessment.

Inleiding

Om de kwaliteit van toetsen en assessments te evalueren zijn standaarden en kwaliteitscriteria te gebruiken die onder andere zijn voorgesteld door Baartman, Bastiaens, & Kirschner (2005). Het nadeel van deze standaarden en kwaliteitscriteria is dat ze vrij abstract blijven en er geen beoordelingsvoorschrift is toegevoegd. Dit betekent enerzijds dat het moeilijk is om de criteria naar een specifieke toets te vertalen. Anderzijds leidt een beoordeling met deze criteria niet tot een eindoordeel over de kwaliteit van een toets. Aan de hand van een beoordeling is de ultieme kwaliteitsvraag 'Is deze toets kwalitatief goed genoeg?' niet te beantwoorden.

Om toch tot een eindoordeel over de kwaliteit van competentie assessments te kunnen komen is in navolging van het COTAN beoordelingsinstrument voor studietoetsen en tests (Evers, 2004) een beoordelingsinstrument voor competentie assessments ontwikkeld. Dit artikel gaat in op dit door de auteur ontwikkelde beoordelingsinstrument. Eerst wordt beschreven welke aspecten van een competentie assessment beoordeeld worden. Daarna komt aan de orde hoe

een beoordeling met het beoordelingsinstrument tot stand komt. Vervolgens wordt een onderzoek naar de ervaringen van gebruikers van het beoordelingsinstrument beschreven. Het artikel sluit af met een korte beschrijving van de geplande vervolgstappen in de ontwikkeling van het beoordelingsinstrument.

De ingrediënten van het beoordelingsinstrument

In het beoordelingsinstrument worden vijf hoofdcategorieën onderscheiden die hier kort besproken worden. Hierbij is de volgorde aangehouden die door de COTAN is geïntroduceerd en waarbij de categorie 'Kwaliteit van Testmateriaal en Handleiding' uiteenvalt in 2a 'Kwaliteit van het Testmateriaal' en 2b 'Kwaliteit van de Handleiding'.

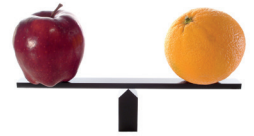
1. Uitgangspunten van Testconstructie

Een toekomstig gebruiker van een assessment moet kunnen beoordelen of het assessment past bij het gebruiksdoel dat de gebruiker voor ogen heeft. Een theoretische omschrijving van de competenties die met het assessment beoordeeld kunnen worden is dan ook noodzakelijk. Dit geldt ook voor een duidelijke omschrijving van de beslissingen die aan de hand van het assessment genomen kunnen worden.

2.A. Kwaliteit van het Testmateriaal

In deze hoofdcategorie zijn de criteria over drie onderwerpen verdeeld, te weten: inhoud, ontwerp en vormgeving. Een specifiek kenmerk van assessment is de authentieke inhoud en context van de voorgelegde taken. Dit betekent dat de taken in een assessment veel gelijkenis vertonen met taken uit de toekomstige beroepspraktijk (Gulikers, Bastiaens, Kirschner, 2004). Het is van belang om te beoordelen in hoeverre taken authentiek zijn en of dit overeenkomt met het gebruiksdoel van het assessment. De expert beoordeelt daarom of de taken representatief zijn voor de toekomstige beroepssituatie. Het beoordelingsinstrument vereist daarnaast een bepaalde mate van

In het beoordelingsinstrument worden vijf hoofdcategorieën van kwaliteitscriteria onderscheiden



standaardisatie van het assessment zodat de herhaalbaarheid van beslissingen gegarandeerd kan worden. Er zijn echter vormen van assessments waarbij standaardisatie vrijwel onmogelijk is, maar waar op andere manieren herhaalbaarheid en vergelijkbaarheid van beslissingen gerealiseerd wordt. Dit kan bijvoorbeeld door mogelijke antwoordpatronen op te stellen. Verder wordt ook vanuit het oogpunt van herhaalbaarheid en vergelijkbaarheid nagegaan of het ontwerp van het assessment de minimalisering van beoordelaarverschillen in de hand werkt. Ten slotte komt ook de vormgeving aan de orde, waarbij getoetst wordt of de vormgeving van het assessment niet onnodig ingewikkeld is zodat voorkomen wordt dat kandidaten nodeloos de fout in gaan.

2.B. Kwaliteit van de Handleiding

In de hoofdcategorie 'kwaliteit van de handleiding' wordt ingegaan op de informatievoorziening aan verschillende partijen. Uitgangspunt hierbij is dat het assessment transparant is voor alle betrokkenen (Baartman, Bastiaens, Kirschner & van der Vleuten, 2006). Docenten of toetscommissies moeten weten hoe het assessment ingezet kan worden. Daarnaast moeten assessoren precies weten wat er van hun verwacht wordt en hoe zij zich tijdens het assessment moeten gedragen. Is het bijvoorbeeld toegestaan dat assessoren de kandidaten bevragen tijdens het assessment? Ten slotte is het ook noodzakelijk dat de kandidaten gedetailleerde informatie over het assessment ontvangen. Zij moeten bijvoorbeeld vooraf weten wat er tijdens het assessment verwacht wordt en welke beoordelingscriteria gehanteerd worden.

3. Normen en Standaarden

De normen of standaarden van een assessment worden gebruikt om te bepalen of de ruwe score van een kandidaat goed genoeg is. Het is dan ook van groot belang dat dit zorgvuldig wordt vastgesteld. In het beoordelingsinstrument wordt zowel met relatieve als met absolute normeringen rekening gehouden. Het beoordelingsinstrument biedt daarom zowel criteria om de kwaliteit van een normeringsonderzoek en normgroepen te beoordelen alsook voor beoordeling van een vastgestelde standaard en standaardbepalingsprocedure. In de beoordeling wordt meege-

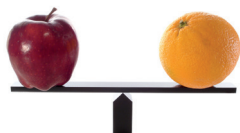
nomen dat het belangrijk is om een voor de toets geschikte methode te kiezen en dit te verantwoorden. Zo ligt het bij certificerende competentie assessments meer voor de hand om gebruik te maken van standaarden om uitspraken te kunnen doen over de bekwaamheid ten opzichte van een onveranderlijke standaard. Het is dan immers de bedoeling om aan te geven of iemand een bepaalde competentie verworven heeft en niet om aan te geven of iemand bij de 10% meest competente studenten behoort.

4. Betrouwbaarheid

Betrouwbaarheid is een maat waarmee wordt aangegeven in hoeverre metingen vrij zijn van de invloed van toevallige factoren. Voor competentie assessments is het van belang dat vooral betrouwbaarheidsmaten gebruikt worden die passen bij de specifieke taken die in deze toetsvorm gebruikelijk zijn (zoals praktijkopdrachten of essayvragen). Daarnaast is er bij competentie assessments een grote rol voor beoordelaars weggelegd. In het beoordelingsinstrument wordt dan ook veel aandacht besteed aan de borging en aantoonbaarheid van de kwaliteit van beoordelaars of assessoren.

5. Validiteit

De validiteit van een assessment kan worden omschreven als de mate waarin het assessment aan zijn doel beantwoordt. Bieden de toegekende scores het soort informatie waarin we geïnteresseerd zijn en kunnen deze scores gebruikt worden bij het nemen van goede beslissingen (bijvoorbeeld zakken of slagen)? Om dit te kunnen aantonen is het aan de toetsconstructeurs om argumenten te leveren die de experts ervan overtuigen dat de toets of het assessment tot valide beslissingen leidt. Daarnaast leveren zij empirisch bewijs om hun argumenten te ondersteunen (Kane, 2004). Het is vervolgens een taak van de beoordelende expert om de argumenten te wegen en te beslissen of het aannemelijk is dat het assessment tot valide scores en beslissingen leidt. Eén van de elementen die hierbij doorslaggevend kunnen zijn, is bijvoorbeeld in hoeverre het inhoudsdomein door de taken gedekt wordt.



Het is noodzakelijk dat de kandidaten gedetailleerde informatie over het assessment ontvangen

Scoringsvoorschrift

Het eindoordeel voor hoofdcategorie 2B, 'Kwaliteit van de Handleiding' is gelijk aan het laagst toegekende oordeel op één van de drie onderdelen. Het vaststellen van de oordelen voor de verschillende onderdelen gebeurt door de scores bij elkaar op te tellen.

Informatieverstreking:

11 – 12 punten	Goed
8 – 10 punten	Voldoende*
4 – 7 punten	Onvoldoende

Informatie voor assessoren:

5 – 6 punten	Goed
4 punten	Voldoende*
2 – 3 punten	Onvoldoende

Informatie voor kandidaten:

8 – 9 punten	Goed
6 – 7 punten	Voldoende*
3 – 5 punten	Onvoldoende

* De beoordeling 'voldoende' wordt niet toegekend indien aan minstens een van de criteria een onvoldoende toegekend is.

teitscriteria onderscheiden. Per categorie worden enkele basisvragen gesteld. Aan de hand van deze vragen wordt bepaald of aan een bepaald minimum is voldaan. Vervolgens wordt per hoofdcategorie een verschillend aantal criteria beoordeeld. Alle criteria in het instrument zijn voorzien van een uitgebreide toelichting. De expert weet daardoor hoe het criterium geïnterpreteerd moet worden en wat goed genoeg is. In de toelichting wordt tevens beschreven hoe een criterium beoordeeld moet worden wanneer slechts gedeeltelijk aan de eisen is voldaan. De expert kent vervolgens een 'onvoldoende' (1 punt), 'voldoende' (2 punten) of 'goed' (3 punten) aan de criteria toe. Als alle criteria beoordeeld zijn kunnen deze met behulp van de bijgeleverde scoringsvoorschriften samengevoegd worden tot een oordeel over de hoofdcategorie. De inhoudelijke eisen die aan een criterium gesteld zijn, zoals het absolute minimum, het verschil tussen een voldoende en goed en de invulling van scoringsvoorschriften, zijn op basis van literatuuronderzoek geformuleerd. Er ligt echter een arbitraire beslissing van de onderzoeker aan de invulling ten grondslag en hier is tijdens de evaluatie van het beoordelingsinstrument dan ook op ingegaan. In figuur 1 is een voorbeeld van een scoringsvoorschrift opgenomen.

Figuur 1 Scoringsvoorschrift van hoofdcategorie 'Kwaliteit van de Handleiding'

De gebruiksaanwijzing van het beoordelingsinstrument

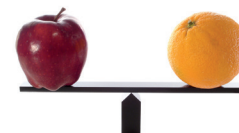
De beoordeling van een assessment vindt plaats volgens dezelfde procedure als die van de COTAN (zie bijvoorbeeld www.cotan.nl). Samenvattend houdt dit in dat een assessment met al het bijbehorende materiaal en een wetenschappelijke verantwoording wordt voorgelegd aan experts op het gebied van toetsing en toetsconstructie. Het oordeel over de kwaliteit van het assessment wordt gevormd op basis van de voorgelegde materialen. Het is dus niet de bedoeling dat de experts een assessment bijwonen: het betreft een 'papieren beoordeling'.

Om tot een oordeel te komen wordt gebruikgemaakt van een beoordelingsinstrument. Hierin worden voor de beoordeling vijf hoofdcategorieën van kwaliteitscriteria onderscheiden.

De ervaringen met het beoordelingsinstrument

Na de ontwikkeling van het beoordelingsinstrument is door de auteur van het beoordelingsinstrument een evaluatieonderzoek uitgevoerd. In dit onderzoek is aan twaalf assessment experts gevraagd om het CSPE (centraal schriftelijk en praktisch examen) Transport en Logistiek voor de Basisberoepsgerichte Leerweg uit 2006 te beoordelen (zie ook: Van Hest, Zeelenberg & Rietvorst, 2008). Dit examen bestaat uit zowel praktijkopdrachten als korte kennistoetsen. Een voorbeeld van een praktijkopdracht is ter illustratie opgenomen in figuur 2. Na het beoordelen van het CSPE hebben de experts hun ervaringen tijdens een vraaggesprek met de onderzoeker toegelicht. Aan de hand van de uitgevoerde beoordelingen en de vraaggesprekken is een evaluatie van het beoordelingsinstrument uitgevoerd die zich richtte op drie onderwerpen: het ontwerp van het instrument, de hanteerbaarheid van het instrument en de betrouwbaarheid van het instrument.

Experts hebben uiteenlopende meningen hebben over de opgenomen beoordelingscriteria



Ontwerp

Het eerste deel van het onderzoek richtte zich op het ontwerp van het beoordelingsinstrument. In dit kader bleek dat de experts uiteenlopende meningen hadden over de opgenomen beoordelingscriteria en de invulling hiervan. Een aantal experts wilde criteria opnemen om ook formatieve assessments te beoordelen. Anderen vonden dat er al te veel criteria opgenomen waren en dat het instrument korter moest. Er waren tevens experts die de criteria momenteel te streng vonden en dat het oordeel daardoor lager uitviel dan verwacht. Een tweede punt betrof de keuze voor het COTAN-format en de indeling in vijf hoofdcategoryën. De meeste experts (11 van de 12) zijn van mening dat deze indeling een goed raamwerk biedt om alle criteria onder te kunnen brengen.

Hanteerbaarheid

De tweede onderzoeksvraag richtte zich op de hanteerbaarheid van het instrument. Het bleek dat hier nog veel werk te verrichten is. De teksten kunnen verder gesimplificeerd worden en een verbetering van de lay-out kan ook bijdragen aan een betere bruikbaarheid van het instrument. Het instrument kan ook geautomatiseerd worden waardoor experts niet meer zelf de eindscores hoeven uit te rekenen. Er trad echter ook een probleem op dat meer fundamenteel voor expertbeoordelingen lijkt te zijn. Het is namelijk de vraag hoeveel expertise van een expert verwacht mag worden. Vooral ten aanzien van de meer psychometrische onderdelen van het beoordelingsinstrument is een zekere basiskennis vereist. Het is dan ook een uitdaging om een beoordelingsinstrument te ontwikkelen dat toegankelijk is voor experts die minder thuis zijn in de psychometrie, zoals inhoudelijk vakdeskundigen. Tegelijk moet recht gedaan worden aan de complexiteit van kwaliteitsbepaling van toetsen.

Betrouwbaarheid

De derde en laatste onderzoeksvraag was gericht op de betrouwbaarheid van het beoordelingsinstrument, gedefinieerd als beoordelaarovereenstemming. De beoordelaarovereenstemming was voor het beoordelingsinstrument in de huidige vorm nog onvoldoende. Het instrument is dus in deze vorm nog onvoldoende betrouwbaar om tot een gedegen oordeel over een

ONDERDEEL C MELDEN CALAMITEITEN

tijsduur 20 minuten

Gisteren is een zending diepvrieskippen binnengekomen die niet is opgeslagen in de vriescellen. Deze calamiteit moet je telefonisch melden aan jouw chef.

Bedenk dat je moet handelen volgens de 'Instructie voor de warehousemedewerker van Galaxy BV' (bijlage 1).

Voor deze opdrachten heb je nodig bijlage 1 en uitwerkbijlage 8.

Bereid dit telefoongesprek voor door de Instructie (bijlage 1) en uitwerkbijlage 8 goed te bestuderen.

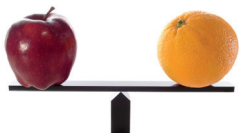
2p	21	Vraag aan de examinerator naar het telefoonnummer waarop je chef te bereiken is. Bel je chef op en leg hem het probleem uit.
5p	22	Met je chef heb je afspraken gemaakt over wat je moet doen. Een collega van je moet de afspraken uitvoeren. Vul de telefoonnotitie in, zodat het voor je collega duidelijk is wat hij moet doen.

Figuur 2 Praktijkopdracht uit het CSPE Transport & Logistiek – BB 2006

assessment te komen. Een verklaring hiervoor kan de beperkte expertise van de beoordelaars op het gebied van psychometrie zijn.

Conclusie en vervolgstappen

Tijdens de ontwikkeling en evaluatie van het beoordelingsinstrument bleek dat een aantal inhoudelijke aspecten van kwaliteit van competentie assessments nog onvoldoende uitgewerkt is om tot een eenduidig beoordelingsinstrument te komen. Daarnaast bleek dat de hanteerbaarheid van het instrument zoals dat er nu ligt te wensen over laat. Vanuit deze ervaringen is binnen het RCEC (Research Center voor Examinering & Certificering) een promotieonderzoek gestart dat zich richt op het uitwerken van specifieke aspecten van de kwaliteit van competentie assessments. Tijdens dit onderzoek wordt onder andere de in het beoordelingsinstrument gebruikte methode voor valideren verder uitgewerkt. Er zal ook aandacht zijn voor het bepalen van kwaliteit van gecombineerde toetsen (bijvoorbeeld in een competentie assessment programma). Uiteindelijk zullen de inzichten die verworven worden tijdens het promotieonderzoek, verwerkt worden in een verbeterde versie van het beoordelingsinstrument. Bij de ontwikkeling van deze tweede versie zullen



Een aantal inhoudelijke aspecten van kwaliteit van competentie assessments is nog onvoldoende uitgewerkt

de resultaten van de evaluatiestudie met betrekking tot hanteerbaarheid en betrouwbaarheid uiteraard meegenomen worden.

Literatuur

Baartman, L.K., Bastiaens, T.J. & Kirschner, P.A. (2005).

Kwaliteitscriteria voor Competentie Assessment Programma's. *EXAMENS*, 2(2), 13-15.

Baartman, L.K., Bastiaens, T.J., Kirschner, P.A. & van der Vleuten, C.P. (2006). The wheel of competency assessment: Presenting quality criteria for competency assessment programs. *Studies in Educational Evaluation*, 32, 153-170.

Evers, A. (2004). Hoe verantwoord meten we?

EXAMENS, 1(1), 31-34.

Gulikers, J.T.M., Bastiaens, T.J. & Kirschner, P.A. (2004). A

five-dimensional framework for authentic assessment. *Educational Technology Research & Design*, 57, 67-87.

Kane, M.T. (2004). Certification Testing as an Illustration of Argument-Based Validation. *Measurement*, 2, 135-170.

Van Hest, E., Zeelenberg, R., & Dietvorst, P. (2008). Het centraal schriftelijk en praktisch examen (cspe). *EXAMENS*, 5(1), 11-16.

□ Mw.drs. S. Wools is als promovendus werkzaam bij het Onderwijskundig Onderzoeks Centrum van Cito. Haar promotieonderzoek richt zich op de kwaliteitsbeoordeling van assessments en wordt uitgevoerd in opdracht van het RCEC (Research Center voor Examinering en Certificering). E-mail: saskia.wools@cito.nl.

In het volgende nummer van **EXAMENS** komen onder andere volgende onderwerpen aan bod:

- Beoordelen met het oog op de toekomst: tien kenmerken van duurzaam beoordelen
- Beoordelingsschalen in competentiegerichte vmbo-examens
- Terecht of niet...? – Typefouten in examenvragen
- Het aanscherpen van de exameneisen voor de vmbo-basisvakken
- Een interview met Willem Hofstee: Hofstee beoordeelt de beoordelaar

EXAMENS 1, 2010 verschijnt in maart

Ook de SOG examens



staan onder toezicht van de

Examen
kamer
kwaliteit door onafhankelijkheid