Evaluatie van een instrument voor kwaliteitsbeoordeling van competentie-assessments.

Saskia Wools
Universiteit Utrecht,
Cito, Arnhem,
2007

Samenvatting

Om de kwaliteit van competentie-assessments te kunnen bepalen is een beoordelingsinstrument ontwikkeld. In deze studie wordt dit beoordelingsinstrument geëvalueerd middels een kwalitatief onderzoek. De evaluatie richt zich op drie deelthema's: ontwerp, hanteerbaarheid en betrouwbaarheid. Om tot suggesties voor verbetering van het instrument te komen is door 11 experts een Centraal Schriftelijk en Praktisch Examen (CSPE) Transport en Logistiek beoordeeld. Aan de hand van de ervaringen tijdens het beoordelen van het examen en de uiteindelijke beoordelingen is gebleken dat experts zich grotendeels konden vinden in de gekozen kwaliteitscriteria. Verder bleek het beoordelingsinstrument voor een aantal experts moeilijk te hanteren omdat het moeilijk was de benodigde informatie te selecteren. Tevens is gebleken dat de beoordelaarsovereenstemming van het beoordelingsinstrument nog onvoldoende is.

In de afgelopen jaren zijn opleidingen steeds vaker vormgegeven volgens principes van het competentiegericht onderwijs. Door de groeiende belangstelling voor deze onderwijsvorm stijgt ook de vraag naar hierbij passende toetsvormen. In het competentiegericht onderwijs wordt een leeromgeving gecreëerd waarin competent handelen centraal staat (Mulder, 2003). Dit betekent dat kennis, vaardigheden en attitudes geïntegreerd ingezet worden in de beroepstaak (Roelofs, 2006). Het ontwerpen van toetsvormen die hierbij aansluiten is volgens Van Berkel, Hofman, Kinkhorst en Te Lintelo (2003) één van de lastigste elementen van competentiegericht onderwijs omdat tegemoet moet worden gekomen aan het integratieve karakter van competenties. Om te kunnen beoordelen of studenten competent zijn is het dan ook wenselijk om niet alleen gebruik te maken van reeds gebruikelijke toetsvormen maar om deze ook te combineren met nieuwe toetsvormen (Hendriks & Schoonman, 2006). Deze nieuwe toetsvormen kunnen bijvoorbeeld situatiebeoordelings-testen zijn, maar ook computersimulaties, proeven van bekwaamheid of de beoordeling van een portfolio.

Nieuwe toetsvormen kunnen samengevat worden onder de noemer 'assessments'. De term assessment is overgenomen uit het bedrijfsleven waar het vooral binnen de Human Resource Management (HRM) gebruikt wordt. In het onderwijs worden assessments gebruikt voor het beoordelen van competenties in het algemeen (bijvoorbeeld reflecterend vermogen) alsmede voor het beoordelen van specifiek gedrag (bijvoorbeeld het voeren van een slechtnieuws gesprek) (Hendriks & Schoonman, 2006). Door deze dubbele gebruiksmogelijkheid kan het begrip assessment gezien worden als paraplubegrip waaronder meerdere toetsvormen geplaatst kunnen worden zoals de eerder genoemde nieuwe toetsvormen, die ook wel assessmenttypes genoemd worden (Brown & Knight, 1995). Een aantal auteurs (Dierick & Dochy, 2001; Dochy, Segers & Sluijsmans, 1999) is van mening dat naast assessmenttypes ook assessmentmodes, zoals peer-assessment en co-assessment, onder het paraplubegrip vallen. Een nauwkeurige begripsbepaling is dan ook vereist.

Roelofs en Straetmans (2006) introduceren een driedeling om de diverse assessments te categoriseren: hands-on instrumenten, simulaties en hands-off instrumenten. De grenzen tussen deze drie categorieën lijken op het eerste gezicht duidelijk, maar lopen in werkelijkheid in elkaar over. *Hands-on instrumenten* worden gebruikt voor het beoordelen van prestaties in reële werksituaties. Deze taken worden uitgevoerd in de complexe, vrijwel onveranderde werksituatie. Deze vorm van assessment is in zekere zin onvoorspelbaar en de moeilijkheidsgraad is niet volledig onder controle. Bij een *simulatie* moet een kandidaat in

een min of meer nagebootste omstandigheid een vaardigheid demonstreren. In de simulaties worden aspecten uit de beroepspraktijk gemanipuleerd of weggelaten en er is dus sprake van een versimpelde versie. *Hands-off instrumenten* confronteren de kandidaat met complexe of beroepskritische situaties. In sommige varianten wordt naast de oplossing ook het oplossingsproces beoordeeld. Vaak staat de afweging die een kandidaat in staat stelt om tot een juiste beslissing te komen centraal. De complexiteit van de werkelijkheid is in deze assessmentvorm gereduceerd tot voor de taak essentiële informatie.
De assessments die in dit onderzoek centraal staan passen binnen bovenstaande categorisering. Het gaat hier dus om assessmenttypes zoals beschreven door Brown en Knight (1995) en niet om assessmentmodes. Daarnaast voldoen deze assessments aan de zes kenmerken die genoemd zijn door O'Neil en Abedi (1996):

Het assessment…

1. vraagt studenten iets te tonen, creëren, produceren,
2. vraagt naar probleemoplossende vaardigheden en hogere orde denken,
3. bevat taken die betekenisvolle activiteiten voorstellen,
4. bevat realistische toepassingen,
5. vereist nieuwe instructie- en beoordelingsvormen,
6. maakt gebruik van menselijke beoordelingen.

Kwaliteit van assessments.

In de afgelopen jaren is het gebruik van assessments toegenomen. Inmiddels worden assessments niet meer alleen incidenteel als formatief beoordelingsinstrument ingezet. Steeds vaker hebben de assessments een selecterende, classificerende of certificerende functie. Dit betekent dat assessments worden ingezet om tot beslissingen over mensen te komen met betrekking tot opleiding of werk (Vermetten, Daniëls & Ruijs, 2000). Deze zogenaamde high stake beslissingen vereisen kwalitatief goede meetinstrumenten. Door de unieke aard van assessments kunnen er echter vraagtekens gezet worden bij het gebruik van de traditionele kwaliteitscriteria om de kwaliteit van assessments te bepalen. Een aantal auteurs pleit voor een verruiming van traditionele criteria als betrouwbaarheid en validiteit (Cronbach, 1989; Kane, 1992; Messick, 1994). Daarnaast zijn er auteurs die nieuwe criteria willen introduceren (Frederiksen & Collins, 1989; Haertel, 1991; Linn, Baker & Dunbar 1991). Volgens deze auteurs doen deze verruimde of nieuwe criteria meer recht aan de eigenheid van assessments en houden de criteria meer rekening met de specifieke eigenschappen van deze nieuwe toetsvormen (Dierick, Dochy & van de Watering, 2001). In dit onderzoek is binnen een kader van traditionele kwaliteitscriteria ruimte gecreëerd om nieuwe criteria als transparantie, cognitieve complexiteit en authenticiteit (Baartman, Bastiaens, Kirschner & van der Vleuten, 2007) een plaats te geven.

Kwaliteitsbepaling

Wanneer men de kwaliteit van meetinstrumenten wil bepalen worden hiervoor beoordelingsinstrumenten ingezet. De term beoordelingsinstrument kan leiden tot enige verwarring. In dit artikel wordt 'beoordelingsinstrument' gebruikt om te verwijzen naar het instrument waarmee de kwaliteit van assessments wordt bepaald. Om te verwijzen naar instrumenten waarmee studenten beoordeeld kunnen worden, wordt zoals eerder genoemd het begrip assessment gebruikt.
In Nederland wordt de kwaliteit van psychologische tests vastgesteld door de COTAN (Commissie Testaangelegenheden Nederland). De COTAN maakt voor deze kwaliteitsbepaling gebruik van hun eigen beoordelingssysteem (2004). Dit systeem is bedoeld voor psychologische testen, maar inmiddels worden ook onderwijstoetsen op deze manier beoordeeld. Hierbij moet opgemerkt worden dat het dan wel gaat om papieren toetsen met meerkeuzevragen, ook wel paper based tests (pbt's) genoemd. In 2004 is er een bewerking van het COTAN instrument verschenen om naast de kwaliteit van pbt's ook de kwaliteit van computer based tests (cbt's) vast te stellen (Keuning, 2004). Deze laatste

bewerking gaat echter nog steeds uit van toetsen met meerkeuze vragen waardoor beide systemen niet geschikt zijn om de kwaliteit van assessments te bepalen. Zoals gezegd is er inmiddels toenemende behoefte aan een duidelijk kwaliteitskader voor assessments. Een voordeel van een beoordelingssysteem zoals gebruikt wordt door de COTAN ligt in de operationalisatie van de kwaliteitscriteria. De criteria worden in meetbare elementen weergegeven waardoor het relatief eenvoudig en navolgbaar is om de kwaliteit van een assessment te bepalen. Verder is het wenselijk om met een nieuw beoordelingsinstrument aan te sluiten bij een bestaand instrumentarium. Dit komt naast de herkenbaarheid ook de vergelijkbaarheid ten goede.
Vanuit deze uitgangspunten is er voorafgaand aan dit onderzoek een bewerking van het COTAN systeem ontwikkeld ten behoeve van de kwaliteitsbepaling van competentie-assessments. Het Beoordelingsinstrument: Kwaliteit van Competentie-Assessment combineert reeds bestaande kwaliteitscriteria voor pbt's en cbt's met de nieuwe inzichten op dit gebied voor assessments.

Het COTAN beoordelingssysteem

Voordat ingegaan wordt op het beoordelingsinstrument voor assessment wordt eerst het COTAN beoordelingssysteem besproken. Het COTAN beoordelingssysteem richt zich zoals gezegd op pbt's. Deze toetsen worden ter beoordeling voorgelegd op het moment dat de toets klaar is voor gebruik. Vervolgens vindt de beoordeling van de kwaliteit plaats door de toets aan twee onafhankelijke beoordelaars voor te leggen. De beoordelaars ontvangen naast de daadwerkelijke toets ook de handleiding en technische verantwoording van de toets, de antwoordformulieren en scoringssleutels, en van toepassing zijnde artikelen, rapporten en dissertaties. De beoordelaars beoordelen het assessment middels het beoordelingssysteem van COTAN en eventuele verschillen tussen de twee beoordelaars worden besproken. Indien nodig wordt een derde onafhankelijke beoordelaar ingezet. Vervolgens is er de mogelijkheid voor de toetsconstructeur te reageren op de beoordeling. De COTAN bestudeert de reactie van de toetsconstructeur en bepaalt of deze gevolgen heeft voor het gegeven eindoordeel. Daarna wordt de definitieve eindbeoordeling vastgesteld en gepubliceerd.
In het beoordelingssysteem van COTAN worden voor de beoordeling vijf hoofdcategorieën onderscheiden. Per categorie worden enkele basisvragen gesteld, en aan de hand van deze vragen wordt bepaald of aan een bepaald minimum is voldaan. Vervolgens wordt aan de hand van een per categorie variërend aantal criteria vastgesteld of aan het hoofdonderdeel een 'onvoldoende', 'voldoende' of 'goed' wordt toegekend.

De beoordeling van een test leidt tot een waardering op de volgende hoofdonderdelen:

1. Uitgangspunten van de testconstructie.
   Een gebruiker moet kunnen beoordelen of de test past bij het doel waarvoor een test gezocht wordt. Het is daarom van belang een heldere omschrijving van de meetpretentie van de test te geven. Daarnaast moet de keuze van testinhoud en de wijze waarop begrippen worden gemeten verduidelijkt worden.
2. Kwaliteit van het testmateriaal en de handleiding
   De afname en instructie van een test moeten gestandaardiseerd zijn zodat de variatie door instructieverschillen of toetsleider geminimaliseerd is. Daarnaast moet er een duidelijke handleiding beschikbaar zijn.
3. Normen
   Een ruwe score krijgt pas betekenis wanneer er een vergelijking mogelijk is met een norm. Hierbij kan er sprake zijn van domeingerichte interpretatie of normgerichte interpretatie.
4. Betrouwbaarheid
   Er worden verschillende vormen van betrouwbaarheid onderscheiden: paralleltestbetrouwbaarheid, interne-consistentiebetrouwbaarheid, test-

hertestbetrouwbaarheid en interbeoordelaarsbetrouwbaarheid. Verder dient de generaliseerbaarheid vastgesteld te worden.

5. Validiteit
   Het betreft hier de mate waarin een test aan zijn doel beantwoordt, oftewel kan men uit de testscores de conclusies trekken die bedoeld zijn. Er wordt hierbij aangesloten op de klassieke driedeling in validiteit: begripsvaliditeit, criteriumvaliditeit en inhoudsvaliditeit.

Beoordelingsinstrument: Kwaliteit voor Competentie-Assessment

Zoals gezegd is het beoordelingsinstrument voor assessments een bewerking van het COTAN beoordelingssysteem. Dit betekent dat de structuur van beide beoordelingsinstrumenten hetzelfde is maar dat de criteria inhoudelijk verschillen. Het beoordelingsinstrument voor assessments is bedoeld om de kwaliteit van assessments vast te stellen. Om het beoordelen niet onnodig complex te maken is gekozen om de kwaliteit van enkelvoudige assessmentinstrumenten te bepalen en niet van zogenaamde competentie assessment programma's (CAP's) (Baartman, et al., 2007). Wanneer assessments gecombineerd ingezet worden in een CAP kan van elk onderdeel afzonderlijk de kwaliteit bepaald worden aan de hand van dit beoordelingsinstrument.

De procedure

De procedure die gevolgd wordt om tot een beoordeling te komen middels het beoordelingsinstrument voor assessment is vrijwel identiek aan de procedure van COTAN. Voor de beoordeling van de kwaliteit van het assessment worden onafhankelijke beoordelaars benaderd. Deze beoordelaars kunnen gezien worden als experts op het gebied van toetsing en assessment. Het materiaal van het assessment en eventuele ondersteunende informatie worden naar de beoordelaar verzonden en de beoordelaar baseert zijn oordeel op deze toegezonden informatie. Om tot dit oordeel te komen wordt het beoordelingsinstrument ingezet. Het beoordelingsinstrument bevat binnen de vijf hoofdcategorieën een aantal vragen die de beoordelaars dienen te beantwoorden met goed, voldoende of onvoldoende. Wanneer deze antwoorden gegeven zijn kan er middels een beoordelingsvoorschrift een score toegekend worden. Deze score correspondeert vervolgens met een eindoordeel voor een hoofdcategorie.
In het instrument zijn ter ondersteuning van de beoordelaars uitwerkingen van de criteria opgenomen. In deze uitwerkingen staat aan welke voorwaarden door het assessment moet zijn voldaan voor een goed, voldoende of onvoldoende. De beoordelaar maakt per criterium met behulp van de uitwerking en de informatie over het assessment een afweging met betrekking tot het oordeel.

Inhoudelijke omschrijving

Hier volgt een korte omschrijving van de hoofdcategorieën van het beoordelingsinstrument. In bijlage I is opgenomen welke criteria onder de verschillende hoofdcategorieën geplaatst zijn.

1. Uitgangspunten van testconstructie
   In deze hoofdcategorie wordt nagegaan of de ontwerper van het assessment de gemaakte keuzes gespecificeerd en verantwoord heeft. Dit betreft, overeenkomstig met het COTAN beoordelingssysteem, keuzes als de functie van het assessment en de doelgroep waarvoor het assessment bedoeld is. Daarnaast wordt in deze hoofdcategorie een competentiebeschrijving beoordeeld, een onderscheidend element ten opzichte van het oorspronkelijke beoordelingssysteem.
2. Kwaliteit van testmateriaal en de handleiding
   Deze hoofdcategorie valt net als in het COTAN beoordelingssysteem in twee delen uiteen: kwaliteit van het testmateriaal en kwaliteit van de handleiding. In het eerste deel zijn de criteria over drie onderwerpen verdeeld, te weten: inhoud, ontwerp en vormgeving. Vooral in deze hoofdcategorie zijn 'nieuwe' beoordelingscriteria gebruikt.

De criteria om de inhoud van het assessment te beoordelen richten zich namelijk op de authenticiteit van het assessment. De authenticiteit van het assessment wordt afgemeten aan de mate waarin het assessment representatief is voor de criteriumsituatie. De criteriumsituatie is de situatie zoals die redelijkerwijs in de praktijk te verwachten is na het afronden van de opleiding (Gulikers, Bastiaens, Kirschner, 2004; Gulikers, 2006). De criteria onder het onderwerp ontwerp richten zich voornamelijk op de herhaalbaarheid van beslissingen en vergelijkbaarheid van het assessment (Baartman et al., 2007). Daarnaast wordt nagegaan of het ontwerp de minimalisering van beoordelaarsverschillen in de hand werkt. Ten slotte komt ook de vormgeving aan de orde, waarbij de duidelijkheid van het assessment getoetst wordt.
In de categorie kwaliteit van de handleiding wordt ingegaan op de informatie-voorziening aan verschillende partijen. Er wordt bepaald of er voldoende informatie verschaft wordt voor toetsgebruikers, bijvoorbeeld docenten of schoolleiding. Daarnaast wordt beoordeeld of er voldoende informatie voor assessoren beschikbaar is. In tegenstelling tot het COTAN beoordelinssysteem wordt in dit beoordelings-instrument ook ruim aandacht geschonken aan de informatievoorziening richting kandidaten.

3. Normen en standaarden
   In het beoordelingsinstrument worden relatief normeren en absoluut normeren behandeld omdat beide manieren van normeren aan de orde kunnen zijn bij assessments. Wanneer relatief genormeerd wordt, wordt er vooraf een bepaald slagingspercentage vastgesteld (Hambleton & Pitoniak, 2006). Bij absoluut normeren wordt de zak/slaag beslissing genomen aan de hand van vooraf vastgestelde standaarden (Haertel & Loiré, 2004). In de praktijk komen ook mengvormen voor waarbij de op standaarden gebaseerde grensscore achteraf aangepast wordt op basis van de prestaties van de populatie. In het beoordelingsinstrument is ruimte om assessment die gebruik maken van alle drie de vormen van normeren te beoordelen. In het oorspronkelijke COTAN-beoordelingsinstrument wordt echter alleen ingegaan op criteria die betrekking hebben relatief normeren.
4. Betrouwbaarheid
   In de hoofdcategorie betrouwbaarheid wordt rekening gehouden met het gebruik van polytome items en bijbehorende betrouwbaarheidindices. Dit is noodzakelijk omdat assessments voornamelijk bestaan uit taken met een open karakter. Het COTAN systeem was daarentegen meer gericht op dichotome items. In het nieuwe beoordelingssysteem is naast aandacht voor polytome items ook aandacht voor beoordelaarsovereenstemming. Beoordelaars spelen immers vaak een belangrijke rol tijdens assessments.
5. Validiteit
   Het gedeelte over validiteit heeft de grootste verandering ondergaan. Er is gekozen voor de beoordeling van de validiteit van een assessment aan de hand van het doorlopen valideringsproces. Als proces van valideren wordt de Argument Based Approach (Kane, 2004; Kane, 2006) aangehangen. In deze benadering wordt eerst een interpretatief argument geformuleerd waarin wordt beschreven welke redenering gevolgd is. Vervolgens worden de stappen in deze redenering met empirische bewijzen onderbouwd in het zogenaamde validiteitsargument.

Onderzoeksvragen

Het onderzoek zal een formatieve evaluatie (Patton, 2002) behelzen van het instrument zoals dat er nu ligt. Het primaire doel van het onderzoek is het verbeteren van het instrument. Dit doel wordt nagestreefd door een kwalitatieve studie uit te voeren waarbij ook een aantal kwantitatieve analyses wordt uitgevoerd. Voor dit onderzoek is aan een aantal experts op het gebied van assessment gevraagd om de kwaliteit van een assessment te beoordelen met behulp van het beoordelingsinstrument. De evaluatie die op basis van de

beoordelingen is uitgevoerd richt zich op drie overkoepelende onderzoeksthema's: het ontwerp van het instrument, de hanteerbaarheid van het instrument, en de betrouwbaarheid van het instrument.

Het ontwerp van het instrument

In het instrument zijn keuzes gemaakt met betrekking tot de nieuwe of verruimde criteria. Hiermee is getracht een beoordelingssysteem te ontwikkelen waarmee assessments beoordeeld kunnen worden. In dit onderzoek zal in eerste instantie nagegaan worden of experts het eens zijn met deze keuzes. Dit komt aan de orde onder de volgende hoofdvraag en deelvragen:

- In hoeverre kunnen experts zich vinden in de gekozen kwaliteitscriteria voor het beoordelen van assessments?
    - Zijn er cruciale kwaliteitsaspecten te noemen die niet in de gehanteerde criteria aan de orde komen?
    - Is er een prioritering aan te brengen in de gehanteerde criteria? En zijn er op basis van deze prioritering criteria aan te wijzen die zouden kunnen vervallen?
    - Zijn experts het eens met de keuze voor de argument-based approach als benadering voor het hoofdonderdeel validiteit?

Hanteerbaarheid van het instrument

Om het assessment goed te kunnen beoordelen is het van belang dat experts alle vragen kunnen beantwoorden. Om de vragen te kunnen beantwoorden zal vaak bekeken moeten worden of bepaalde informatie al dan niet aanwezig is. Daarnaast zal de kwaliteit van deze informatie beoordeeld moeten worden. In dit onderdeel van het onderzoek wordt nagegaan hoe tot een beoordeling gekomen is en of experts erin slagen de benodigde informatie te selecteren. Voorafgaand aan het onderzoek is door de ontwerpers van het beoordelingsinstrument vastgesteld welke informatie de experts idealiter zouden selecteren.

- Is het instrument voor experts hanteerbaar?
    - Is het voor experts mogelijk om uit de beschikbare informatie over het assessment de juiste te selecteren?
    - Vinden experts de omvang van het beoordelingsinstrument aanvaardbaar?
    - Kiezen verschillende experts dezelfde informatie voor het beantwoorden van vragen over het assessment?

Betrouwbaarheid van het instrument

De uitspraken over de kwaliteit van een assessment die voortvloeien uit een beoordeling met het instrument dienen per beoordelaar weinig te variëren. Het is met name van belang dat beoordelaars tot dezelfde oordelen komen op de verschillende hoofdcategorieën. Deze oordelen zullen uiteindelijk naar de toetsgebruikers gecommuniceerd worden en dienen dus het meest betrouwbaar te zijn.

- Is het instrument betrouwbaar?
    - Is de beoordelaarsovereenstemming voldoende om van een betrouwbaar instrument te spreken?

Methode

Participanten

Voor de evaluatie van het beoordelingsinstrument zijn 15 experts benaderd. Deze 15 experts zijn geselecteerd op basis van hun expertise op het gebied van assessment. Daarnaast is in verband met de bereidheid om mee te werken ervoor gekozen experts te benaderen die hun

belangstelling hadden uitgesproken of die reeds bekend waren met het Cito. De experts zijn afkomstig uit verschillende disciplines, te weten, onderwijskunde, psychologie en psychometrie. Door deze experts uit diverse vakgebieden te benaderen is geprobeerd het instrument vanuit meerdere perspectieven te evalueren. Twaalf experts hebben naar aanleiding van een uitnodiging toegezegd te zullen participeren in het onderzoek. Tijdens het onderzoek is één expert afgevallen.

Materiaal

De dataverzameling vond plaats in twee fases. In de eerste fase werd een assessment beoordeeld. Hiervoor is een te beoordelen assessment met bijbehorende toetsverantwoording verspreid. Tijdens het beoordelen konden de experts hun gedachtegang rapporteren aan de hand van een zogenaamd 'hardop-denkprotocol' in de vorm van een formulier waarop per vraag de gemaakte afwegingen gerapporteerd konden worden. In de tweede fase van de dataverzameling werd de uitgevoerde beoordeling van het assessment middels een open interview toegelicht.

Het assessment

Er is gekozen om een CSPE (centraal schriftelijk en praktisch examen) uit 2006 aan de experts voor te leggen. Het CSPE Transport en Logistiek BB 2006 is een examen dat door leerlingen van de basisberoepsgerichte leerweg (BB) van het VMBO gemaakt wordt. Het examen bestaat uit praktijkopdrachten die in een nagebootste omgeving moeten worden uitgevoerd. Daarnaast wordt via korte multiplechoicetoetsen voor het examen relevante kennis getoetst.
Het CSPE Transport en Logistiek past binnen de eerder gestelde definiëring van assessment. Ten eerste is het te plaatsen binnen de categorisering van Roelofs en Straetmans (2006) aangezien hier sprake is van een simulatie. Er worden immers taken uitgevoerd in een gefingeerde praktijksituatie. Verder voldoet het aan vrijwel alle gestelde kenmerken van assessment. Er wordt gevraagd om bepaalde handelingen te demonstreren, er worden betekenisvolle taken voorgelegd, de toepassingen zijn realistisch, er wordt aangesloten op praktijkgerichte instructievormen en ten slotte wordt er gebruik gemaakt van een assessor voor het beoordelen van de taken. Alleen het tweede kenmerk 'het assessment vraagt naar probleemoplossende vaardigheden en hoger orde denken' is niet van toepassing op het CSPE. Dit wordt onder andere veroorzaakt door het niveau van het assessment en de kandidaten uit de basisberoepsgerichte leerweg van het VMBO.
Het CSPE dat ter beoordeling aan de experts werd voorgelegd voldeed verder aan een aantal eisen die voor dit onderzoek van belang waren. In de eerste plaats kwamen alle onderdelen van het beoordelingsinstrument aan bod. Dit betekent dat het assessment in ieder geval over alle in het beoordelingsinstrument gevraagde onderdelen beschikte. Door hier op toe te zien was vooraf bekend dat op alle hoofdcategorieën van het beoordelingsinstrument een beoordeling mogelijk was. Daarnaast betrof het een assessment dat niet in één oogopslag als kwalitatief goed of slecht afgedaan kon worden omdat juist de afwegingen en redeneringen van belang waren voor de evaluatie van het instrument. Wanneer een assessment zonder afwegingen als goed, voldoende of onvoldoende beoordeeld kon worden was dit van weinig waarde voor het onderzoek. Ten slotte ging het om een goed gedocumenteerd assessment dat vanaf papieren documentatie te beoordelen is. Het was immers irreëel om van experts te verwachten dat zij een assessment lijfelijk bij zouden wonen. Het was daarnaast van belang dat bepaalde onderdelen nog eens rustig nagelezen konden worden zodat de afwegingen weloverwogen gemaakt konden worden.
Voor het CSPE Transport en Logistiek BB 2006 was geen verantwoording beschikbaar. Deze is daarom door de onderzoekers geschreven. Hierbij werd tegemoet gekomen aan de voor het onderzoek gestelde eisen aan het assessment. Door de verantwoording te construeren was het mogelijk de indeling van het beoordelingssysteem grotendeels aan te houden. Op deze manier was de informatie voor experts beter gestructureerd. Daarnaast was het mogelijk om de validiteit van het instrument aan te tonen aan de hand van een

interpretatief argument (Kane, 2004), waardoor ook het onderdeel validiteit goed beoordeelbaar zou zijn.
Een aantal onderdelen van de verantwoording is bewust slecht gedocumenteerd. Op deze manier was na te gaan hoe experts omgaan met ontbrekende informatie. Er kon onderzocht worden of experts bij afwezige informatie een onvoldoende toekennen of aangeven dat een bepaald criterium niet beoordeelbaar is. In tabel 1 is te zien welke elementen in de verantwoording ontbraken of zeer summier aanwezig waren.

Tabel 1: ontbrekende informatie in de informatie over het CSPE Transport en Logistiek

Onderwerp:	Hoofdcategorie:	
Lokale betrouwbaarheid	4. Betrouwbaarheid	Over de lokale betrouwbaarheid van het CSPE Transport en Logistiek is in de verantwoording geen enkele informatie verschaft.
Normen	3. Normen & Standaarden	In deze hoofdcategorie is het afhankelijke van de gebruikte normeringsmethode welke subcategorieën beoordeeld worden. De experts kiezen welke subcategorieën (normen, standaarden, panel) zij willen beoordelen. In het geval van het CSPE Transport en Logistiek is er geen sprake van relatieve normering en de vragen onder 'normen' hoeven in principe niet beantwoord te worden.
Handleiding voor kandidaten	2b. Kwaliteit van de handleiding	Er was geen expliciete handleiding voor kandidaten aanwezig. In het assessment stonden wel 2 zinnen als instructie voor de kandidaat. Dit was echter de enige informatie die verschaft werd.

Onderzoeksmateriaal

Onder de experts zijn twee formulieren verspreid. Ten eerste een beoordelingsformulier. Hierop konden de scores van de beoordeling worden bijgehouden. Tevens werd op dit formulier aangegeven wat het eindoordeel is op een bepaald hoofdonderdeel. Het tweede formulier is een zogenaamd 'hardop-denk-protocol'. Dit formulier werd ingevuld tijdens het beoordelen van het assessment. Hierop kon de expert bijhouden hoe tot een oordeel gekomen was en welk bewijs daarvoor doorslaggevend was. Daarnaast konden eventuele inhoudelijke opmerkingen over het instrument gegeven worden. Het 'hardop-denk-protocol' kon tevens als geheugensteun dienen om tijdens het interview op terug te vallen.
Na de beoordeling vond met elke expert een open interview plaats. Dit gesprek vond plaats aan de hand van een interviewleidraad. In het interview werd ingegaan op de ervaringen met het beoordelingsinstrument en werd gevraagd naar achtergrondinformatie over de expert. Verder kwamen onderwerpen die door de expert in de formulieren genoemd werden aan de orde. Ten slotte was er ruimte voor een toelichting van de beoordelingen en eventuele vragen naar aanleiding van het hardop-denk-protocol

Procedure

De experts ontvingen na bevestiging een onderzoeksmap met daarin de benodigde materialen. In de onderzoeksmap was een instructie bijgevoegd waarin de gang van zaken rondom het onderzoek werd toegelicht. Er werd niet nader ingegaan op het te beoordelen assessment. Wel werd voor aanwijzingen voor het beoordelen verwezen naar de instructie

van het beoordelingsinstrument. Na de beoordeling van een expert volgde het (telefonische) interview.

Analyse

Voor de analyse is een data analyse tabel (tabel 2) gemaakt. Deze tabel geeft aan welke data gebruikt wordt voor het beantwoorden van de onderzoeksvragen.

Tabel 2: data analyse tabel

Ontwerp		
	In hoeverre kunnen experts zich vinden in de gekozen kwaliteitscriteria voor het beoordelen van assessments?	
1.1	Cruciale eigenschappen	Interview
1.2	Prioritering/Criteria vervallen	Interview
1.3	Argument Based Approach	Interview
Hanteerbaarheid		
	Is het instrument voor experts hanteerbaar?	
2.1	Mogelijk informatie te selecteren	Hardop-denk-protocol
2.2	Omvang beoordelingsinstrument	Hardop-denk-protocol
2.3	Dezelfde informatie	Interview
Betrouwbaarheid		
	Is het instrument betrouwbaar?	
3.1	Beoordelaarsovereenstemming	Beoordelingsformulier

Ter beantwoording van de onderzoeksvragen over het ontwerp en de hanteerbaarheid van het instrument zijn kwalitatieve analyses uitgevoerd. Voor de analyse zijn de interviews geprotocolleerd en de hardop-denk-protocollen werden verzameld. Vanuit de interview- en hardop-denk-protocollen zijn alle uitspraken van de experts met betrekking tot het instrument verzameld door telkens de hoofdzin uit een frase over te nemen. Alle uitspraken zijn op kaarten gezet zonder daarbij te vermelden van wie de uitspraak afkomstig was. Deze kaarten zijn vervolgens op basis van de onderzoeksvragen geclusterd. Een aantal kaarten paste niet binnen deze clusters en zijn vervolgens onder het bijpassende overkoepelende onderzoeksthema geplaatst. Na deze eerste clustering is een controle van de analyse uitgevoerd door een tweede persoon. Naar aanleiding van deze controle zijn in overleg 29 uitspraken in andere clusters geplaatst. In tabel 3 is te zien hoe de uitspraken in eerste instantie door beide beoordelaars over de clusters verdeeld waren.

Tabel 3: verdeling van de uitspraken over de clusters

Cluster:	Afkorting	1a	1b	2ab
Ontwerp: algemeen	OA	12	24	16
Ontwerp: aspecten die niet aan de orde komen	ON	12	15	14
Ontwerp: prioritering/criteria vervallen	OP	10	4	5
Ontwerp: mening validiteit	OV	15	10	14
Hanteerbaarheid: algemeen	HA	14	17	14
Hanteerbaarheid: mogelijk informatie te selecteren	HS	14	9	13
Hanteerbaarheid: omvang (tijd & hoeveelheid items)	HO	14	11	15
Totaal aantal uitspraken:		91	90	91

1a: clustering 1e beoordelaar
1b: clustering 2e beoordelaar
2ab: clustering na overleg tussen beoordelaars

De veranderingen die in de clustering zijn aangebracht hadden verschillende achtergronden. Ten eerste was de context van de uitspraken bij de 2e beoordelaar niet bekend en dit leidde tot verschillen in de clustering. Verder is na overleg besloten om alle opmerkingen die over de scoringsmethodiek van het beoordelingsinstrument gingen onder te brengen onder het cluster 'Ontwerp: algemeen', wat ook leidde tot een aantal veranderingen. Op dezelfde manier is besloten alle opmerkingen die experts gemaakt hebben over 'validiteit' te plaatsen binnen 'Ontwerp: mening validiteit', zelfs als dit uitspraken waren die de hanteerbaarheid van het onderdeel validiteit betroffen. Ten slotte bleken er verschillen te zitten in de interpretatie van de clusters 'ontwerp: aspecten die niet aan de orde komen' en 'ontwerp: prioritering/criteria vervallen' waardoor verschillende uitspraken in eerste instantie verschillend geclusterd waren.
Voor de beantwoording van de onderzoeksvragen over de hanteerbaarheid van het beoordelingsinstrument zijn naast de interviews ook de hardop-denk-protocollen gebruikt. In de hardop-denk-protocollen is de experts gevraagd aan te geven welke informatie zij gebruikt hebben om tot een oordeel te komen. Tijdens de analyse van deze hardop-denk-protocollen bleek echter dat experts deze formulieren vooral gebruikt hebben om opmerkingen over het beoordelingsinstrument te plaatsen. Er is niet goed te herleiden welke informatie tot een bepaald oordeel heeft geleid en welke afweging gemaakt is.
Voor het thema 'betrouwbaarheid' zijn de beoordelingsformulieren kwantitatief verwerkt. Voor de analyse van deze data is het programma RA 2006 gebruikt. Met deze analyse wordt op twee manieren aangegeven welke beoordelaars afwijkend zijn ten opzichte van de overige beoordelaars. Om deze afwijkende beoordelaars aan te kunnen wijzen werden Rbt (Correlatie beoordelaarsscore en totaalscore) en RSI (ranking similarity index) berekend. Met Rbt wordt de correlatie tussen de individuele scores van beoordelaars en tussen de gemiddelde scores van alle beoordelaars weergegeven. De RSI signaleert ook afwijkende beoordelaars. De beoordelaar met de laagste waarde is de meest afwijkende beoordelaar (Heuvelmans & Sanders, 1993).
Naast gegevens over afwijkende beoordelaars werden Gowers coëfficiënten berekend om de beoordelaarsovereenstemming te bepalen. Deze coëfficiënten zijn gebaseerd op de absolute range van de beoordelingsschaal waardoor Gowers coëfficiënt geïnterpreteerd kan worden als maat voor de gemiddelde overeenstemming tussen beoordelaars per object (Zegers, 1989). Er is gekozen om Gowers coefficienten te berekenen omdat deze maat ongevoelig is voor variantie wat vanwege het kleine aantal waarnemingen wenselijk is.

Resultaten

Voordat wordt ingegaan op de evaluatie van het beoordelingsinstrument komen de uitkomsten van de beoordelingen van het assessment aan de orde.

Beschrijvende statistieken

Het assessment is door elf experts beoordeeld. In bijlage I is te zien hoe de experts de vragen uit het beoordelingsinstrument beantwoord hebben. Het gemiddelde oordeel staat ook weergegeven. In deze bijlage is te zien dat een aantal experts geen keuze kon maken tussen twee scores. Voor de berekening van het gemiddelde is telkens volgens het principe 'voordeel van de twijfel' de hoogste score meegenomen. Daarnaast is in deze tabel te zien dat een aantal experts heeft aangegeven dat items niet van toepassing waren. Verder konden niet alle criteria door alle beoordelaars beoordeeld worden. Dit is ook in de bijlage terug te vinden.
In het beoordelingsinstrument wordt een eindoordeel gegeven in 5 hoofdcategorieën. De tweede hoofdcategorie valt echter uiteen in twee delen. Er worden daarom per expert zes eindoordelen weergegeven in tabel 4 op de volgende pagina.
De eindoordelen kwamen via de beoordelingsvoorschriften in het beoordelingsinstrument tot stand. Bij de berekening van de eindoordelen is wederom uitgegaan van het principe 'voordeel van de twijfel' wanneer er meer dan één score aan een vraag was toegekend. Er is

een M (missing) weergegeven wanneer één of meer van de vragen als 'niet beoordeelbaar' was benoemd. Uit tabel 4 blijkt dat experts de kwaliteit van het assessment op veel punten onvoldoende vonden. Met name de uitgangspunten van testconstructie en de kwaliteit van de handleiding zijn erg slecht beoordeeld.
Het was niet mogelijk om op basis van de dataverzameling indices van betrouwbaarheid te berekenen. Er waren te weinig waarnemingen, waarbinnen te weinig variantie optrad om Guttman's Lambda of interne consistentie te berekenen. Er wordt in de bespreking van de deelvragen wel ingegaan op de beoordelaarsovereenstemming, waar wel gegevens over bekend zijn.

Tabel 4: eindoordelen op hoofdcategorieën

Experts / Hoofdcategorieën	A	B	C	D	E	F	G	H	I	J	K
1. Uitgangspunten van testconstructie	1	1	1	2	1	2	2	1	1	1	1
2a. Kwaliteit van het testmateriaal	3	1	2	3	1	1	M	1	M	1	2
2b. Kwaliteit van de handleiding	1	1	1	1	2	2	M	1	1	2	1
3. Normen en Standaarden	1	M	1	3	1	2	M	2	1	2	M
4. Betrouwbaarheid	1	M	1	2	1	2	M	3	M	1	M
5. Validiteit	3	3	1	1	1	2	M	3	M	1	M

1: Onvoldoende 2: Voldoende 3: Goed M: Missing

Resultaten per deelvraag

Voor de evaluatie van het beoordelingsinstrument zijn onderzoeksvragen gesteld die passen binnen drie thema's. De resultaten van het onderzoek worden per thema besproken. In de interviews met de verschillende experts zijn niet telkens precies dezelfde onderwerpen aan bod geweest. Daarom wordt in de weergave van de resultaten expliciet aangegeven wanneer een aantal experts een bepaalde visie delen. In het volgende gedeelte van de tekst zullen de resultaten van het onderzoek per deelvraag aan de orde komen. Er zijn echter ook resultaten die niet binnen deelvragen te plaatsen zijn. Deze resultaten zullen aan het eind van de overkoepelende onderzoeksthema's besproken worden.

I. Ontwerp

Deelvraag 1

Zijn er cruciale kwaliteitsaspecten te noemen die niet in de gehanteerde criteria aan de orde komen?

Drie experts, (B, C en H), hebben aangegeven geen criteria te missen. Zij zijn van mening dat de eigenschappen die beoordeeld moeten worden om een oordeel te kunnen vellen over de kwaliteit van het assessment in voldoende mate in de criteria aan bod komen. Een aantal andere experts heeft opmerkingen geplaatst over criteria die ontbreken of elementen van het assessment die ze middels het beoordelingsinstrument niet konden beoordelen. Expert D zegt hierover: 'ik mis een criterium om de taakaspecten, de omstandigheden van het assessment te beoordelen'. Hierbij gaat het niet om de vormgeving, maar om het ontwerp van het assessment. Expert E miste vooral criteria om formatieve aspecten van assessments te beoordelen. Hierbij valt te denken aan criteria die nagaan of er rekening wordt gehouden met gevolgen voor het leren van de student, of een criterium om te bepalen of 'het assessment voldoende zekerheid geeft voor het functioneren van de lerende in de toekomst'. Verder is opgemerkt dat er geen criterium is dat nagaat of er aan het beoogde doel van de toets wordt voldaan. In het geval van competentie-assessment zou expert E namelijk graag willen weten: 'wordt daadwerkelijk competentie gemeten?'. Expert E zou ook graag zien dat er criteria over nieuwe kwaliteitseisen als transparantie en consequentiële validiteit aan het

instrument worden toegevoegd. Expert A is ten slotte van mening dat er meer aandacht kan zijn voor aanvaardbaarheid en praktische bruikbaarheid.

Deelvraag 2

> *Is er een prioritering aan te brengen in de gehanteerde criteria? En zijn er op basis van deze prioritering criteria aan te wijzen die zouden kunnen vervallen?*

Experts A en D geven aan dat zij het eens zijn met de keuze voor de vijf hoofdcategorieën als belangrijke aspecten van kwaliteit voor assessment. Expert E en G doen een voorstel voor een inhoudelijke invulling van deze prioritering. Expert G: 'De belangrijkste vraag moet zijn naar representativiteit of naar generaliseerbaarheid'. Expert C gaat inhoudelijk in op één van de criteria uit het onderdeel Betrouwbaarheid en vraagt zich af: 'zou de feitelijke hoogte van de coëfficiënten niet zwaarder moeten wegen?'.
Geen van de experts vindt dat er criteria zouden kunnen vervallen op basis van een mogelijke prioritering. Expert B is wel van mening dat het instrument in zijn algemeenheid te gedetailleerd is. Experts E en I zijn echter van mening dat het instrument niet te specifiek is en dat er 'volstrekt geen overbodige vragen' zijn.

Deelvraag 3

> *Zijn experts het eens met de keuze voor de Argument Based Approach als benadering voor het hoofdonderdeel validiteit?*

Vier experts (A, E, H, I) hebben aangegeven de keuze voor de Argument Based Approach te onderschrijven. Expert A noemt deze benadering bijvoorbeeld 'heel legitiem'. Expert B heeft aangegeven deze benadering erg moeilijk te vinden. Volgens deze expert werd dat onder andere veroorzaakt doordat de tekst erg abstract en weinig concreet was. Ten slotte was expert C erg negatief over deze benadering van het valideringsproces. De expert vond dat de benadering erg zwak was en dat er te weinig empirisch bewijs verlangd werd.
Drie experts hebben inhoudelijk gereageerd op de algemene teksten van het hoofdonderdeel Validiteit. Expert A is van mening dat het onderscheid tussen validiteit en het proces van valideren niet expliciet genoeg is. De twee andere experts gaan inhoudelijk in op de uitwerking van het model van Kane (2004, 2006). Beide experts zijn van mening dat het model van Kane op een andere manier moet worden uitgewerkt dan in het beoordelings-instrument het geval is. Expert I vindt dat het competentiedomein een overbodig domein is en dat dit inhoudelijk hetzelfde is als het testdomein. Expert D vindt juist dat het onmogelijk is om domeinen over te slaan en is van mening dat elk domein altijd aanwezig is maar 'het competentiedomein kan eventueel gelijkvallen met bijvoorbeeld het testdomein'.
Er is ook een aantal experts (A, C en D) die aangegeven hebben dat zij het gevoel hadden dat een aantal criteria uit het onderdeel validiteit al eerder aan de orde was geweest. Expert C is hier fundamenteel op tegen: 'de testconstructeur wordt twee keer beloond of gestraft voor hetzelfde, dat is onjuist'.

Overige resultaten Ontwerp

Een aantal experts heeft opmerkingen gemaakt die niet binnen de gestelde deelvragen onder te brengen zijn. De opmerkingen betreffen wel het ontwerp van het instrument. Expert G is van mening dat de kwaliteit van assessments niet met een beoordelingsinstrument zoals dat van COTAN beoordeeld kan worden. 'Het COTANsysteem is voor psychologische testen en bij assessment past een andere beoordeling vanwege de gevarieerdheid. Daarom moet het COTANsysteem losgelaten worden.'
Twee andere experts maken opmerkingen over de mogelijkheid om criteria te beoordelen met goed, voldoende en onvoldoende. Expert H en I geven aan dat zij graag een extra antwoordmogelijkheid wilden en als voorbeelden noemen zij zeer onvoldoende, matig en ruim voldoende. Expert I voegt hier aan toe dat 'experts niet kunnen aangeven waarom zij niet gescoord hebben, of ze hebben wel gescoord maar hadden eigenlijk liever niet willen scoren'.

Daarnaast vraagt expert A zich af of binnen deze hoofdcategorieën een ‘compensatorisch model’ geschikt is. Impliciet geeft deze expert daarmee aan voorstander te zijn van het aanbrengen van een prioritering. Deze expert is het immers niet eens met het middelen van oordelen op verschillende onderdelen. Een gewogen gemiddelde dat door het aanbrengen van een prioritering kan ontstaan zou dan een uitkomst zijn.
Ten slotte hebben drie experts aangegeven in hoeverre zij vonden dat het oordeel dat uit het beoordelingsinstrument kwam overeenkomstig was met het oordeel wat zij intuïtief zouden geven. Expert C geeft aan dat zijn gevoel niet overeen kwam met de beoordeling. Het assessment kwam negatiever uit dan hij verwacht had. Expert E en H geven aan dat het oordeel overeenkwam met hun beeld van het assessment.

II. Hanteerbaarheid

Deelvraag 1

Is het voor experts mogelijk om uit de beschikbare informatie over het assessment de juiste te selecteren?

Experts hebben in hun hardop-denk-protocollen niet aangegeven welke informatie zij geselecteerd hebben om tot een oordeel te komen. Bij deze deelvraag wordt daarom ingegaan op de ervaringen bij het selecteren van bewijs die experts in het interview genoemd hebben.
Vier experts (A, B, E en H) geven aan dat het lastig was om de juiste bewijzen te selecteren. Dit wordt met name veroorzaakt door het zoeken naar informatie. Expert H zegt hierover: “Soms moet je lang zoeken omdat je bang bent iets over het hoofd te zien.” Daarnaast geeft deze expert aan dat er in het beoordelingsinstrument een andere terminologie wordt aangehouden dan in het te beoordelen materiaal. Dit levert ook problemen op bij het selecteren van het juiste bewijs. Het zoeken van bewijzen leverde in een enkel geval ook frustratie op omdat “het zoeken naar informatie vertragend werkt,” aldus expert B.

Deelvraag 2

Vinden experts de omvang van het beoordelingsinstrument aanvaardbaar?

Veel experts geven aan dat zij een dagdeel (tussen 4 en 6 uur) nodig hadden. Experts C en D geven aan een paar dagen nodig gehad te hebben omdat zij het beoordelingsinstrument grondiger bekeken hebben. Experts B en G vonden dat het te veel tijd vergde om het assessment te beoordelen. De andere experts vonden de tijd die nodig was acceptabel.
De experts die aangaven dat het beoordelen van het assessment te veel tijd kostte, zijn ook van mening dat er te veel criteria te beoordelen waren. Expert G vindt het beoordelingsinstrument ‘tegelijk massief en pietepeuterig’ en zou het beoordelingsinstrument liever minder gedetailleerd zien.

Deelvraag 3

Kiezen verschillende experts dezelfde informatie voor het beantwoorden van vragen?

Uit de interviews is af te leiden dat experts hun informatie voornamelijk uit de verantwoording van het assessment hebben gehaald. Daarnaast hebben zij hun antwoorden gebaseerd op het beeld dat zij hadden van het assessment zoals dat bijgevoegd was. Welke specifieke informatie geselecteerd is, is vanwege het ontbreken van informatie in de hardop-denk protocollen niet bekend. Er is wel bekend hoe experts reageren op de eerder genoemde ontbrekende informatie in een assessment
Het criterium over lokale betrouwbaarheid is door vijf experts beantwoord terwijl hier in de toetshandleiding geen informatie over verschaft is. Opvallend is dat dit door twee experts zelfs als ‘goed’ is beoordeeld. Experts B, D, F en I hebben het criterium niet beoordeeld of hebben ‘nvt’ ingevuld. De overige experts hebben het criterium als ‘onvoldoende’ beoordeeld.
Bij het hoofdonderdeel Normen en Standaarden konden experts kiezen welke onderdelen van belang waren. Hoewel de criteria onder het kopje Normen niet ingevuld hoefden te

worden omdat er in het assessment geen sprake was van relatieve normering is dit door vijf experts toch gedaan. Vier experts hebben deze criteria niet beantwoord of 'nvt' op het beoordelingsformulier ingevuld.
Hoewel de informatie voor kandidaten minimaal was, zijn de criteria hierover door alle experts beoordeeld. Slechts één expert (B) heeft op de vraag of er een handleiding voor kandidaten aanwezig was geantwoord dat dit niet het geval was. Drie experts hebben aangegeven dat er informatie aanwezig was maar over de summiere omvang in het hardopdenk-protocol een opmerking geplaatst. De overige experts hebben alle vragen over de handleiding beantwoord.

Overige resultaten Hanteerbaarheid

Ook over het onderwerp Hanteerbaarheid zijn opmerkingen gemaakt die niet binnen de gestelde deelvragen passen. Twee experts (D en E) gaan in op de expertise die nodig is om de beoordeling uit te voeren. Expert E geeft aan vanwege het gebruikte vakjargon veel moeite te hebben met de hoofdcategorieën betrouwbaarheid en validiteit. Expert D wijst op het feit dat er erg specifieke kennis van assessments vereist is.
Verder gaven de experts E, H en I aan dat zij problemen hadden met het maken van afwegingen. Zij geven aan dat de vragen veel interpretatie vergen. Daarom is het niet altijd duidelijk of een onvoldoende, voldoende of goed moet worden toegekend. Daarnaast zegt expert E: 'Soms is er meer dan één criterium. Dan is de afweging wat voor mij het zwaarst weegt lastig.'

III. Betrouwbaarheid

Deelvraag 1

Is de beoordelaarsovereenstemming voldoende om van een betrouwbaar instrument te spreken?

Aan de hand van RSI kunnen beoordelaars aangewezen worden die afwijkend zijn. Afwijkend wil zeggen dat de beoordelaars niet consequent streng of mild zijn maar telkens wisselen. Met Rbt kan ook aangewezen worden welke beoordelaars het meest afwijken ten opzichte van het gemiddelde. Deze beoordelaars wijken echter af naar één kant, en zijn dus of strenger of milder dan de andere beoordelaars. Per hoofdcategorie wordt in tabel 5 aangegeven welke beoordelaars het meest en het minst afwijkend waren.

Tabel 5: Afwijkende beoordelaars

Hoofdcategorieën	Meest afwijkend			Minst afwijkend		
	Expert	RSI	Rbt	Expert	RSI	Rbt
1 Uitgangspunten van testconstructie	K	0.17	0.30	D	0.78	1.00
2a Kwaliteit van het testmateriaal	K	-0.23	-0.38	D & E	0.40	0.93
2b Kwaliteit van de handleiding	B	0.22	0.42	D	0.62	0.88
3 Normen en standaarden	H	-0.17	-0.08	C	0.43	0.91
4 Betrouwbaarheid	F	-0.26	-0.70	C	-0.07	.087
5 Validiteit	J	0.45	0.76	B	0.65	0.93

In de tabel is te zien dat vooral experts C en D niet vaak afwijken. Bij de beoordelaars die het meest afwijkend zijn is niet één afwijkende beoordelaar aan te wijzen, hoewel expert K twee keer als meest afwijkend wordt aangemerkt. Aangezien expert K vanaf de derde categorie (normen en standaarden) geen criteria meer beoordeeld heeft is het niet duidelijk of deze expert in de laatste drie categorieën ook het meest afwijkend zou zijn. Verder is te zien dat expert F de hoofdcategorie betrouwbaarheid strenger heeft beoordeeld dan de andere experts.
De gemiddelde Gower geeft weer in hoeverre er tussen beoordelaars per object overeenstemming is. Een gemiddelde Gower van 1.00 betekent dat er perfecte overeenstemming is, 0.00 betekent dat er geen overeenstemming is. In tabel 6 is de gemiddelde Gower per

hoofdcategorie aangegeven. Hieruit is onder andere af te leiden dat de beoordelaarsovereenstemming voor de hoofdcategorie betrouwbaarheid het laagst is. Dit betekent dat de experts het over de beoordeling van deze hoofdcategorie het minst eens zijn.

Tabel 6: gemiddelde gower per hoofdcategorie

Hoofdcategorieën	Gemiddelde Gower
1 Uitgangspunten van testconstructie	.7218
2a Kwaliteit van het testmateriaal	.6898
2b Kwaliteit van de handleiding	.7171
3 Normen en standaarden	.6087
4 Betrouwbaarheid	.5345
5 Validiteit	.7095

Conclusie & Discussie

In deze evaluatiestudie is het Beoordelingsinstrument: Kwaliteit van Competentie-Assessment nader bekeken. Experts uit verschillende disciplines hebben het beoordelingsinstrument ingezet om de kwaliteit van het CSPE Transport en Logistiek BB 2006 te bepalen. Deze studie was gericht op een drietal thema's met bijbehorende onderzoeksvragen. De onderzoeksvragen worden hier telkens kort beantwoord, gevolgd door een aantal conclusies.

In hoeverre kunnen experts zich vinden in de gekozen kwaliteitscriteria voor het beoordelen van assessments?

De experts vinden dat het beoordelingsinstrument een grotendeels compleet overzicht van kwaliteitscriteria omvat. Er is echter nog weinig aandacht voor formatieve aspecten van assessments. Hier is tijdens de constructie van het beoordelingsinstrument bewust voor gekozen. Wanneer high stake beslissingen (Vermetten, Daniels & Ruis, 2000) genomen worden door middel van assessments is de kwaliteit van het assessment van essentieel belang. Als een assessment daarentegen wordt ingezet om het leergedrag van kandidaten te sturen hoeven niet alle kwaliteitseisen even streng te zijn. Het is dus de vraag in hoeverre het noodzakelijk is ook formatieve elementen van assessments via een streng beoordelingsinstrument te beoordelen.

Verder worden er nog te weinig 'nieuwe' criteria (Baartman et al., 2007) in het instrument gebruikt, waarbij valt te denken aan representativiteit, aanvaardbaarheid en praktische bruikbaarheid. De meeste experts vinden dat er nog een aantal items rondom deze criteria moeten worden toegevoegd. De lengte van het instrument moet echter wel in het oog worden gehouden. In de omvang van het instrument ontstaat wellicht ruimte wanneer er een oplossing gezocht wordt voor de waargenomen dubbellingen in de criteria binnen het onderdeel validiteit.

De ervaringen met de criteria binnen het onderdeel validiteit, waar de meest ingrijpende wijzigingen ten opzichte van het COTAN systeem zijn doorgevoerd, zijn overwegend positief. Desalniettemin werd er ook kritiek geuit op de implementatie van de Argument Based Approach (Kane, 2004), dit richtte zich onder andere op de complexiteit van de benadering. Deze kritiek kan weggenomen worden door meer aandacht te besteden aan de uitwerking van deze benadering van validiteit, bijvoorbeeld door het zorgvuldiger uitwerken van de begeleidende teksten.

De experts kunnen zich dus over het algemeen vinden in de gekozen kwaliteitscriteria alhoewel er ook nog een aantal verbeteringen gedaan kunnen worden. Er zijn in ieder geval geen criteria aan te wijzen die zouden moeten vervallen. Verder zijn er ook geen elementen van assessments die niet meegenomen worden in de beoordeling. Bij de beantwoording van deze onderzoeksvraag is wel gebleken dat de experts veel verschillen van mening over het ontwerp van het beoordelingsinstrument. Een aantal experts vindt de omvang te groot, anderen vinden juist dat er criteria toegevoegd moeten worden. Ook over de

scoringsmethodiek zijn experts het niet eens. Men kan zich afvragen of het mogelijk is om een beoordelingsinstrument te ontwerpen waarin de mening van alle experts verwerkt is.

Is het instrument voor experts hanteerbaar?
Uit het onderzoek komt naar voren dat het instrument voor een aantal experts goed hanteerbaar is. Er zijn ook experts die de beoordeling door verschillende oorzaken niet goed konden uitvoeren. Een aantal experts had erg veel moeite met het vinden van de benodigde informatie. Er waren ook experts die moeite hadden om het beoordelingsinstrument te hanteren vanwege een gebrek aan expertise op een bepaald (deel)gebied. Het is gebleken dat er met name een behoorlijk kennisniveau nodig is op het gebied van psychometrische eigenschappen van assessments. Het was daarnaast onduidelijk voor experts hoe het beoordelingsinstrument gehanteerd diende te worden wanneer informatie in het assessment ontbrak. Er zal binnen de criteria aandacht besteed moeten worden aan het specificeren van de instructie met betrekking tot het ontbreken van informatie in assessments. Experts kiezen immers voor verschillende oplossingen wanneer dit zich voordoet, wat logischerwijs niet wenselijk is.
Een aantal experts had moeite met het vinden van benodigde informatie. Dit is zorgelijk aangezien de informatie bewust zeer gestructureerd gepresenteerd is om het de experts niet onnodig moeilijk te maken. Het is onduidelijk of experts er in zullen slagen een minder gestructureerd assessment te beoordelen. Opgemerkt moet worden dat de door COTAN ter beoordeling voorgelegde pbt's ook ongestructureerd verspreid worden. Hoewel dit niet onderzocht is, is het onwaarschijnlijk dat het probleem dat zich voordoet bij het selecteren van informatie zich beperkt tot het beoordelen van assessments. Het zal inherent zijn aan het beoordelen van de kwaliteit van toetsen volgens een bepaald format. Wanneer toetsconstructeurs niet ditzelfde format aanhouden zal er altijd gezocht moeten worden naar informatie.
Deze deelvraag kon vanwege het ontbreken van informatie in de hardop-denk-protocollen slechts gedeeltelijk beantwoord worden. Het bleek voor experts moeilijk te zijn de hardop-denk-protocollen in te vullen. Het is mogelijk dat de factor tijd hierin een rol heeft gespeeld maar gezien het feit dat geen enkele beoordelaar hier in de interviews op in is gegaan, lijkt dit onwaarschijnlijk. Een andere mogelijkheid is dat het voor experts niet goed navolgbaar was welke informatie uiteindelijk leidde tot een bepaald oordeel en dat zij daarom hierover geen aantekeningen hebben gemaakt. Een verklaring hiervoor zou kunnen zijn dat de beoordeling meer intuïtief gedaan wordt en dat de gemaakte afweging daarom lastig expliciet te maken is.

Is het instrument betrouwbaar?
Uit de analyse van de beoordelingsformulieren is gebleken dat er geen beoordelaars aangewezen kunnen worden die structureel afwijkend zijn. Geen van de beoordelaars is consequent strenger of milder. Daarnaast is er geen beoordelaar aan te wijzen die meer dan andere beoordelaars als afwijkend wordt gekenmerkt. Dit betekent dat alle beoordelaars in de analyses meegenomen konden worden. Kijkend naar deze analyses over alle beoordelaars is gebleken dat de beoordelaarsovereenstemming voor de hoofdcategorie betrouwbaarheid het laagst is. Nunnaly & Bernstein (1994) stellen richtlijnen op voor de gewenste beoordelaarsovereenstemming, die ook in het beoordelingsinstrument zijn overgenomen. Wanneer we deze richtlijnen hier volgen blijkt dat de beoordelaarsovereenstemming voor alle hoofdcategorieën als onvoldoende kan worden geclassificeerd. De experts hebben namelijk bij geen enkele hoofdcategorie een beoordelaarsovereenstemming van boven de .80 behaald. Het instrument is dus in deze vorm nog onvoldoende betrouwbaar om tot een gedegen oordeel over een assessment te komen.
In de interviews kwam naar voren dat een aantal experts ervaarde dat er ruimte was voor interpretatie tijdens het beoordelen. Dit kan er toe geleid hebben dat experts criteria verschillend geïnterpreteerd hebben. Dit kan een verklaring zijn van de relatief lage beoordelaarsovereenstemming. Verder bleek dat niet alle criteria door de experts zijn

beoordeeld. Het hoge aantal missing values heeft bijgedragen aan de lage beoordelingsovereenstemming. Verder dient opgemerkt te worden dat het aantal beoordelaars niet erg groot was, evenals het aantal criteria in bepaalde hoofdcategorieën. Door deze kleine aantallen experts en items is het mogelijk dat de beoordelings-overeenstemming niet geheel accuraat is.

Verbeterpunten

Zoals gezegd zal in het ontwerp van het instrument meer aandacht besteed moeten worden aan de uitwerking van de Argument Based Approach. Wanneer daarnaast ook de uitwerkingen van andere criteria verbeterd wordt zal dit niet alleen het ontwerp van het beoordelingsinstrument ten goede komen, maar ook de hanteerbaarheid. De verbetering van de overige criteria zal zich met name richten op de operationalisatie zodat er minder ruimte is voor een eigen interpretatie van de criteria. Naast een betere uitwerking van de reeds aan-wezige criteria zullen ook een aantal criteria toegevoegd moeten worden, bijvoorbeeld met betrekking tot praktische bruikbaarheid en representativiteit.
Hoewel de uitwerking van criteria zal bijdragen aan een betere hanteerbaarheid van het beoordelingsinstrument, zijn er nog meer verbetersuggesties op dit vlak te noemen. Wanneer het beoordelingsinstrument bijvoorbeeld gesimplificeerd wordt zal het voor meer experts mogelijk zijn alle criteria te beoordelen. Daarnaast is het van belang dat in het instrument opgenomen wordt hoe experts om dienen te gaan met ontbrekende informatie. Dit zal overigens niet alleen de hanteerbaarheid ten goede komen maar ook de beoordelaars-overeenstemming zal hierdoor waarschijnlijk toenemen.
De betrouwbaarheid van het beoordelingsinstrument is in deze studie afgemeten aan de beoordelaarsovereenstemming. Deze maat voor betrouwbaarheid kan toenemen wanneer de hanteerbaarheid van het beoordelingsinstrument verbetert. Verder is het mogelijk dat de overeenstemming tussen experts toeneemt na een training voor beoordelaars, hier zal echter eerst onderzoek naar gedaan moeten worden. Verder zal afgewogen moeten worden in hoeverre een perfecte overeenstemming haalbaar is. Er wordt van experts gevraagd een afweging te maken over de verschillende criteria, bij deze afweging zullen ervaring, expertise en persoonlijke overwegingen altijd een rol blijven spelen.

Vervolgonderzoek

Deze studie behelsde voornamelijk een kwalitatieve evaluatie van het beoordelingsinstrument. Het is nog niet duidelijk wat de kwantitatieve eigenschappen van het instrument zijn. Zo is er bijvoorbeeld niet onderzocht of experts tot dezelfde oordelen kwamen bij een herhaalde beoordeling. Verder is er slechts één assessment aan de experts voorgelegd zodat het onduidelijk is of de ervaringen met het beoordelen veroorzaakt werden door dit specifieke assessment. Deze ervaringen kunnen ook afhankelijk zijn van het CSPE Transport en Logistiek BB. Verder is in deze studie geen onderzoek gedaan naar de validiteit van het beoordelingsinstrument. Er is alleen nagegaan in hoeverre experts het eens zijn met de verschillende criteria en of alle domeinen van assessments gedekt zijn. Tevens moet worden opgemerkt dat de focus van dit onderzoek lag bij het verzamelen van aanbevelingen en verbetersuggesties voor het beoordelingsinstrument.
Op basis van de suggesties die door de verschillende experts zijn gedaan dient het beoordelingsinstrument aangepast te worden. Na deze aanpassing dient het instrument nogmaals onderzocht te worden. Er kan dan nader ingegaan worden op kwantitatieve aspecten van het beoordelingsinstrument. Zo kan bijvoorbeeld de beoordelaarsovereenstemming over meerdere beoordelingen berekend worden of kunnen meerdere assessments beoordeeld worden.
Daarnaast kan het zinvol zijn om te onderzoeken in hoeverre een beoordelaarstraining effect heeft op de mate van overeenstemming tussen de beoordelaars. Er zou tevens gekeken kunnen worden of een dergelijke beoordelaarstraining invloed heeft op de mate waarin experts het beoordelingsinstrument hanteerbaar vinden.

Het is tevens wenselijk om nader onderzoek te doen naar de invulling van de argument based approach als valideringsproces voor competentie assessment. Er kan dan ook gekeken worden naar de verschillende bezwaren ten opzichte van deze benadering die in het huidige onderzoek naar voren zijn gekomen.
Ten slotte dient opgemerkt te worden dat dit onderzoek en beoordelingsinstrument in gaan op de kwaliteitsbepaling van assessments, zoals gedefinieerd in dit onderzoek. Het onderzoek richtte zich dus alleen op enkelvoudige assessments die passen binnen de categorisering hands-on, hands-off en simulatie (Roelofs & Straetmans, 2006). Er is in het onderwijsveld echter nog veel behoefte aan hanteerbare instrumenten die ingezet kunnen worden om de kwaliteit van competentiebeoordeling in het algemeen te kunnen bepalen. Deze instrumenten zullen in de aankomende tijd ontwikkeld moeten worden. Tijdens deze ontwikkeling zal onderzocht moeten worden in hoeverre het mogelijk is om een geïntegreerd beoordelingsinstrument te ontwikkelen, waarmee de kwaliteit van alle studietoetsen beoordeeld kan worden ongeacht de aard van de toets. Het huidige beoordelingsinstrument zou daarvoor als aanzet kunnen dienen.

Literatuur

Baatman, L.K.J., Bastiaens, T.J., Kirschner, P.A., Van der Vleuten, C.P.M. (2007). Evaluating assessment quality in competence-based education: A qualitative comparison of two frameworks. *Educational Research Review, 2,* 114-129.

Brown, S., & Knight, P. (1995). *Assessing Learners in Higher Education.* Londen: Kogan Page.

COTAN. (2004). *Beoordelingssysteem voor de Kwaliteit van Tests.* NIP

Cronbach, L.J. (1989) Construct validation after thirty years. In R.L. Linn (Eds.), Intelligence: Measurement, theory and public policy. (147-171).

Dierick, S., & Dochy, F. (2001). New lines in edumetrics: new forms of assessment lead to new assessment criteria. *Studies in Educational Evaluation, 27*, 307-329.

Dierick, S., Dochy, F., & Van de Watering, G. (2001). Assessment in het hoger onderwijs: over de implicaties van nieuwe toetsvormen voor de edumetrie. *Tijdschrift voor Hoger Onderwijs, 19,* 2-18.

Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: a review. *Studies in Higher Education, 11*, 146-166.

Frederiksen, J.R., & Collins, A. (1989) A system approach to educational testing. Educational researcher, 18, (9), 27-32.

Gulikers, J. (2006). *Authenticity is in de eye of the beholder. Beliefs and perceptions of authentic assessment and the influence on student learning*. Heerlen: Open Universiteit.

Gulikers, J.T.M., Bastiaens, T.J., & Kirschner, P.A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research & Design, 52,* 67-87.

Haertel, E.H. (1991) New forms of teacher assessment. Review of research in education, 17, 3-29.

Haertel, E. H., & Loiré, W. A. (2004). Validating Standards-Based test score interpretations. *Measurement, 2,* 61-103.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting Performance Standards. In R. L. Brennan (ed.), *Educational Measurment 4th edition.* (pp. 433-470). Westport: Praeger Publishers

Hendriks, P., & Schoonman, W. (2006). *Handboek Assessment I, Gedragsproeven. Ontwikkeling, Implementatie en Evaluatie.* Assen: Van Gorcum.

Heuvelmans, A.P.J.M., & Sanders, P.F. (1993) Beoordelaarsovereenstemming. In T.H.J.M. Eggen, & P.F. Sanders (eds.), *Psychometrie in de praktijk*. (pp. 443 – 470). Arnhem: Cito.

Kane, M. (1992) An argument-based approach to validity. Psychological Bulletin, 112, 527-535.

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

Kane, M. T. (2006). Validation. In R. L. Brennan (ed.), *Educational Measurment 4th edition.* (pp. 17-64). Westport: American Council on Education and Praeger Publishers.

Keuning, J. (2004) *De ontwikkeling van een beoordelingssysteem voor het beoordelen van 'Computer Based Tests'.* Arnhem: Cito.

Linn, R.L., Baker, E., & Dunbar, S. (1991) Complex, performance-based assessment: Expectations and validation criteria. Educational Researcher, 16,1-21.

Messick, S. (1994) The interplay of evidence and consequences in the validation performance assessments. Educational Researcher, 23, (2), 13-22.

Mulder, M. (2003). Ontwikkelingen in het competentiedenken en competentiegericht beroepsonderwijs. In M. Mulder, R. Wesselink, H. Biemans, L. Nieuwenhuis, & R. Poell (Eds.), *Competentiegericht Beroepsonderwijs. Gediplomeerd, maar ook bekwaam?* (p. 15-32). Houten: Wolters Noordhof.

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

O'Neil, H. F., & Abedi, J. (1996). Reliability and Validity of a State Metacognitive Inventory: Potential for Alternative Assessment. *The Journal of Educational Research, 89,* 234-245.

Patton, M. Q. (2002) *Qualitative Research & Evaluation Methods 3rd Edition.* Thousand Oaks: Sage.

Roelofs, E. (2006). Een procesmodel voor de beoordeling van competent handelen. *Tijdschrijft voor Hoger Onderwijs, 24,* 152-167.

Roelofs, E. & Straetmans, G. (Eds.) (2006). *Assessment in Actie. Competentiebeoordeling in opleiding en beroep.* Arnhem: Cito.

Van Berkel, A., Hofman, K., Kinkhorst, G., & Te Lintelo, L. (2003). *Competentie-assessment. De ontwikkeling en toepassing van self-, peer- en expert-assessments in het hbo: een praktijkvoorbeeld.* Utrecht: Uitgeverij Lemma.

Vermetten, Y., Daniëls, J., & Ruijs, L. (2000). *Inzet van Assessment: Waarom, wat, hoe, wanneer en door wie? Beslismodel voor een beargumenteerde keuze van assessmentvormen in onderwijs en opleiding.* Heerlen: Open Universiteit Nederland.

Zegers, F. E. (1989). Het meten van overeenstemming. *Nederlands tijdschrift voor de psychologie, 44,* 145 – 156.

Bijlage I: Beoordelingscriteria en toegekende scores van de experts.

1: onvoldoende 2: voldoende 3: goed x of y: de beoordelaar heeft twee scores gegeven
NB: Niet beoordeelbaar voor deze beoordelaar NVT: de beoordelaar vindt dat dit criterium niet van toepassing is op dit assessment

Beoordelingscriteria: \ Experts:	A	B	C	D	E	F	G	H	I	J	K
1. Uitgangspunten van testconstructie Eindoordeel:	**1**	**1**	**1**	**2**	**1**	**2**	**2**	**1**	**1**	**1**	**1**
1.1 Is aangegeven wat de doelgroep(en) is (zijn) van het assessment?	3	3	3	3	2	3	3	3	3	3	2
1.2 Is aangegeven wat de functie of het gebruiksdoel van het assessment is?	2	3	3	2	2	2	2	1	1	1	3
1.3 Is aangegeven welke competentie(s) het assessment beoogt te meten?	2	3	3	2	2	1 of 2	2	1	1	1	2
1.4 Wordt de relevantie van de inhoud van het assessment voor de te meten competentie(s) aannemelijk gemaakt?	2	1	1	1	1	1 of 2	2	1	1	1	2
1.5 Worden theorieën en concepten die aan het assessment ten grondslag liggen besproken?	1	1	2	1	1	1	2	1	NB	3	1
2A. Kwaliteit van het Testmateriaal Eindoordeel:	**3**	**1**	**2**	**3**	**1**	**1**	**NB**	**1**	**NB**	**1**	**2**
2.1 Is het assessment representatief voor de criteriumsituatie?	3	1	1	2	1	1 of 2	NB	1	NB	2	3
2.2 Is het assessment erop gericht de competentie als geheel te meten/beoordelen?	2	1	2	2	1	1 of 2	NB	2	NB	2	3
2.3 Is het assessment gestandaardiseerd?	2	2	2	3	2	2	NB	3	3	3	2
2.4 Maakt het assessment gebruik van een beoordelings- of observatieschema dat de kans op verschillende beoordelingen van beoordelaars minimaliseert?	2	2	2	2	1	2	NB	2	3	3	2

2.5 Is de vormgeving van het assessment zodanig dat het de uitvoering ondersteunt of in ieder geval niet belemmert?	3	3	3	3	2	2	NB	3	3	3	2
2.6 Is het assessment geschikt voor gebruik door verschillende groepen kandidaten?	3	2	3	3	2	2	NB	3	3	3	3
2B. Kwaliteit van de handleiding Eindoordeel:	**1**	**1**	**1**	**1**	**2**	**2**	**NB**	**1**	**NB**	**1**	**2**
2.7 Is er een handleiding voor de assessoren beschikbaar?	3	3	3	3	3	3	NB	3	3	3	3
2.8 Is er een beschrijving van het assessment beschikbaar voor de kandidaten?	2	3	3	3	1	3	NB	3	1	3	3
2.9 Wordt er informatie gegeven over de gebruiksmogelijkheden en beperkingen van het assessment?	2	1	2	2	2	2	NB	2	1	3	2
2.10 Wordt er informatie gegeven over de vereiste condities bij de afname van het assessment?	2	3	2	3	3	2 of 3	NB	3	3	3	3
2.11 Wordt er informatie gegeven over de vereiste deskundigheid voor afname en interpretatie van het assessment?	1	1	2	1	1	1 of 2	NB	1	1	1 of 2	2
2.12 Wordt er informatie gegeven over de interpretatie van de scores?	2	3	3	1	2	1 of 2	NB	1	3	2	3
2.13 Zijn de aanwijzingen voor de assessor volledig en duidelijk?	2	2	2	3	2	2	NB	3	3	3	3
2.14 Is de afnameprocedure van het assessment duidelijk beschreven?	3	1	3	2	2	2	NB	3	3	3	3
2.15 Is voor kandidaten het doel van het assessment vooraf duidelijk?	2	1 of 2	2	1	1	2	NB	2	2	2	1
2.16 Zijn voorafgaand aan het assessment de beoordelingscriteria voldoende geëxpliciteerd?	1	3	1	1	1	2	NB	2	1	2	1

2.17 Is in de handleiding de afnameprocedure bij de kandidaten voldoende beschreven?	1	1 of 2	1	1	1	2	NB	2	1	NB	2
3. Normen en Standaarden Eindoordeel:	**1**	**NB**	**1**	**3**	**1**	**2**	**NB**	**2**	**1**	**2**	**NB**
3.1 Is (Zijn) de gebruikte normgroep(en) representatief?	1	1	1	NVT	2	2	NB	NVT	1	NVT	NB
3.2 Is de steekproefgrootte (per normgroep) toereikend?	1	1	1	NVT	1	NVT	NB	NVT	1	NVT	NB
3.3 Worden gemiddelden, standaardafwijkingen en gegevens over de scoreverdeling vermeld?	2	3	2	NVT	2	NVT	NB	NVT	1	NVT	NB
3.4 Worden de betekenis en de beperkingen van de normschaal duidelijk gemaakt en is het type normschaal in overeenstemming met het doel van het assessment?	1	NB	2	2	2	NVT	NB	NVT	1	NVT	NB
3.5 Is de standaardbepalingsmethode op een verantwoorde manier geselecteerd en is deze geschikt voor het te beoordelen assessment?	1	3	3	3	2	2 of 3	NB	3	1	1	NB
3.6 Is de standaardbepalingsprocedure voldoende omschreven en is deze zorgvuldig uitgevoerd?	1	3	2 of 3	3	2	2	NB	2	2	1	NB
3.7 Zijn er voor de validatie van de standaardbepalingsprocedure validiteitsbewijzen verstrekt?	1	NB	3	3	1	2	NB	3	2	1	NB
3.8 Zijn de beoordelaars naar behoren getraind?	1	3	3	3	2	2	NB	2	2	1	NB
3.9 Is het panel voldoende groot en van voldoende kwaliteit?	2	3	3	2	2	3	NB	2	2	1	NB
3.10 Zijn de beoordelingen van de beoordelaars voldoende consistent?	2	NB	2	3	2	1 of 2	NB	2	2	1	NB
4. Betrouwbaarheid Eindoordeel:	**1**	**NB**	**1**	**2**	**1**	**2**	**NB**	**3**	**NB**	**1**	**NB**
4.1 Zijn de betrouwbaarheidsgegevens berekend voor de steekproeven waarvoor het assessment gebruikt wordt?	1	2	3	3	1	3	NB	3	3	1	NB

4.2 Zijn de betrouwbaarheidsgegevens in overeenstemming met de aard van de gegevens die verzameld zijn?	2	2	3	2	2	1 of 2	NB	3	3	1	NB
4.3 Zijn de betrouwbaarheidsgegevens correct berekend?											
4.3.1 Is de betrouwbaarheidscoëfficiënt of generaliseerbaarheidscoëfficiënt correct berekend?	3	3	3	3	1	2	NB	3	NVT	1	NB
4.3.2 Is de (lokale) betrouwbaarheid correct berekend?	3	NVT	1	NVT	1	NVT	NB	3	NVT	1	NB
4.3.3 Is de consistentie of accuraatheid van classificaties correct berekend?	3	3	3	3	2	NVT	NB	3	NVT	1	NB
4.4 Zijn de resultaten voldoende gelet op het beoogde type beslissingen dat met behulp van het assessment wordt genomen?	1	2 of 3	2	3	1	2 of 3	NB	2	3	1	NB
Validiteit Eindoordeel:	**3**	**3**	**1**	**1**	**1**	**2**	**NB**	**3**	**NB**	**1**	**NB**
5.1 Is er een interpretatief argument beschikbaar?	3	3	3	3	3	3	NB	3	3	2	NB
5.2 Bevat het interpretatief argument de juiste gevolgtrekkingen?	2	2	3	3	2	3	NB	3	3	2	NB
5.3 Zijn de gevolgtrekkingen aannemelijk?	2	2	1	3	2	2	NB	3	1	2	NB
5.4 Zijn de in het valideringsproces aangeleverde bewijzen voldoende?	2	1	1	1	2	2	NB	3	NB	1	NB