Evaluatie van een beoordelingssysteem voor de kwaliteit van competentie-assessments

S. Wools, P. F. Sanders, T. J. H. M. Eggen, L. K. J. Baartman, en E. C. Roelofs

Samenvatting

In dit artikel wordt de ontwikkeling van een beoordelingssysteem voor de kwaliteit van competentie-assessments beschreven. De ontwikkeling is gestuurd door principes van ontwerponderzoek. In de beschrijving van het onderzoek wordt volgens een indeling in drie fases achtereenvolgens de ontwerpopdracht geformuleerd, het prototype beschreven en wordt er gerapporteerd over een uitgevoerd evaluatieonderzoek. Tijdens het evaluatieonderzoek werd het beoordelingssysteem voor de kwaliteit van competentie-assessments door elf assessmentexperts gebruikt om de kwaliteit van een Centraal Schriftelijk en Praktisch Examen Transport en Logistiek te beoordelen. In het artikel wordt geconcludeerd dat het beoordelingssysteem in de huidige vorm nog niet geschikt is voor de beoordeling van de kwaliteit van competentie-assessments. Want hoewel assessmentexperts het in principe eens zijn met gemaakte ontwerpkeuzes is het noodzakelijk om de gebruikersvriendelijkheid te verhogen. Daarnaast zal ook de beoordelaarsovereenstemming verhoogd moeten worden om een goede beoordeling van kwaliteit mogelijk te maken.

1 Inleiding

Door de invoering van competentiegericht onderwijs in het mbo en hbo is het gebruik van competentie-assessments in de afgelopen jaren toegenomen. Competentie-assessments worden nu naast studietoetsen en tests gebruikt om te bepalen of een kandidaat bepaalde competenties verworven heeft. Onder studietoetsen worden hier toetsen verstaan waarmee de kennis en/of de vaardigheid van bijvoorbeeld rekenen of taal getoetst wordt. Met tests worden hier meetinstrumenten bedoeld waarmee psychologische constructen zoals intelligentie gemeten worden. Onder competentie-assessments vallen veel uiteenlopende toetsvormen die door Roelofs en Straetmans (2006) ingedeeld worden in drie categorieën: hands-on instrumenten, simulaties en hands-off instrumenten. Hands-on instrumenten worden gebruikt voor het beoordelen van taken die kandidaten moeten uitvoeren in reële werksituaties. Bij simulaties moeten kandidaten de uitvoering van taken demonstreren in situaties die de reële werksituaties zo getrouw mogelijk nabootsen. Hoewel kandidaten bij hands-off instrumenten wel geconfronteerd worden met beroepskritische situaties, gebruiken deze meetinstrumenten niet de reële werksituaties of nabootsingen daarvan.

Competentie-assessments worden steeds vaker ingezet bij het nemen van belangrijke beslissingen over personen (Vermetten, Daniëls, & Ruijs, 2000). Aan de kwaliteit van competentie-assessments moeten daarom ook hoge eisen gesteld worden. In Nederland zijn eisen waar meetinstrumenten zoals tests en studietoetsen aan zouden moeten voldoen, opgesteld door de Commissie Testaangelegenheden Nederland (COTAN). Door de Inspectie van het Onderwijs zijn eisen, zogenaamde standaarden, opgesteld voor de examens die afgenomen worden in het middelbaar beroepsonderwijs (Inspectie van het Onderwijs, 2009)**.** Voor beide beoordelingssystemen geldt echter dat de beoordelingscriteria onvoldoende recht doen aan de specifieke eigenschappen van competentie-assessments (Dierick, Dochy, & Van de Watering, 2001). Vandaar dat in de literatuur zowel gepleit wordt voor een verruiming van psychometrische criteria zoals betrouwbaarheid en validiteit (Cronbach, 1989; Kane, 1992; Messick, 1994) als voor het gebruik van nieuwe criteria (Baartman, Bastiaens, Kirschner, & Van der Vleuten, 2006; Frederiksen & Collins, 1989; Haertel, 1991; Linn, Baker, & Dunbar, 1991).

Het doel van het onderzoek is te komen tot een beoordelingssysteem voor de kwaliteit van competentie-assessments. De ont-

wikkeling van het beoordelingssysteem is gestructureerd volgens de principes van ontwerponderzoek. Onderwijskundig ontwerponderzoek is geschikt om het ontwerp, de ontwikkeling en de evaluatie van onderwijskundige programma's, producten of interventies systematisch te bestuderen (Plomp, 2007). Hierbij ligt de nadruk op het vergroten van kennis over de kenmerken van de programma's, producten of interventies en het leveren van inzichten in de gevolgde ontwerp- en ontwikkelprocessen. Tijdens het beschreven onderzoek zijn de principes van ontwerponderzoek gevolgd om het ontwerpproces te structureren en om een systematische verslaglegging mogelijk te maken. Dit artikel is een weergave van zowel het gevolgde ontwikkelproces, als van de evaluatie van het beoordelingssysteem voor de kwaliteit van competentie-assessments. Allereerst wordt ingegaan op de invulling van het ontwerponderzoek voor de ontwikkeling van het beoordelingssysteem voor de kwaliteit van competentie-assessments. In de resultatensectie worden vervolgens de drie verschillende fases die tijdens het ontwerpproces doorlopen zijn één voor één beschreven. In deze beschrijving zal onder andere ingegaan worden op de afwegingen die de ontwikkeling van het beoordelingssysteem gestuurd hebben. Daarnaast zal het ontwikkelde prototype gepresenteerd worden. Vervolgens wordt een beschrijving van het uitgevoerde evaluatieonderzoek gegeven. Het artikel sluit af met een conclusie ten aanzien van het ontwikkelde beoordelingssysteem en een algemene discussie over kwaliteitsbeoordeling. Ten behoeve van de leesbaarheid zal in dit artikel het beoordelingssysteem voor de kwaliteit van competentie-assessments aangeduid worden als beoordelingssysteem voor competentie-assessments.

2 Methode

In dit onderzoek zijn de principes van ontwerpgeoriënteerd onderzoek ingezet om de ontwikkeling van het beoordelingssysteem voor competentie-assessments te sturen. Plomp (2007) onderscheidt drie fases die tijdens ontwerpgeoriënteerd onderzoek doorlopen moeten worden. Tijdens de verkenningsfase worden de specificaties van het product in kaart gebracht, wordt een literatuurstudie gedaan naar wat reeds bekend is en wordt er een theoretisch kader voor het onderzoek vastgesteld. In de ontwikkelfase wordt volgens een iteratief proces een prototype ontwikkeld, waarbij telkens formatieve evaluatie wordt ingezet om de verschillende onderdelen van het product te verbeteren. Ten slotte wordt in de evaluatiefase nagegaan of het uiteindelijke product voldoet aan de eerder vastgestelde specificaties. Deze evaluatie leidt vaak tot aanbevelingen voor de verbetering van het product. In deze sectie wordt de invulling van de drie fases samengevat. In de resultatensectie zal inhoudelijk op de fases worden ingegaan.

2.1 Verkenningsfase

In de verkenningsfase is een heldere ontwerpopdracht voor het beoordelingssysteem voor competentie-assessments geformuleerd. Vervolgens is er middels een literatuurstudie in kaart gebracht welke beoordelingssystemen reeds beschikbaar zijn en in hoeverre deze geschikt zijn voor het beoordelen van competentie-assessments. Naar aanleiding van dit overzicht zijn ontwerpspecificaties opgesteld waaraan het nieuwe beoordelingssysteem zou moeten voldoen. In de resultatensectie wordt een samenvatting van de beschikbare instrumenten gegeven en wordt ingegaan op de geschiktheid van deze instrumenten voor de beoordeling van competentie-assessments. Daarnaast worden ook de uiteindelijke ontwerpspecificaties beschreven.

2.2 Ontwikkelfase

Het beoordelingssysteem voor competentie-assessments bestaat uit vijf hoofdonderdelen. Voor elk van deze hoofdonderdelen werd een ontwerpcyclus doorlopen. Tijdens deze cyclus werd op basis van bestudeerde literatuur bepaald welke criteria in het beoordelingssysteem moesten worden opgenomen. Daarna werden de criteria en toelichtingen op deze criteria geformuleerd. Vervolgens werd door twee experts feedback gegeven op zowel inhoudelijke keuzes als taalkundige aspecten van de criteria. Voor elk hoofdonderdeel werd dit proces een aantal keren

doorlopen totdat er een definitief prototype van het beoordelingssysteem voor competentie-assessments was ontwikkeld. In de resultatensectie wordt voor elk van de hoofdonderdelen besproken welke criteria zijn opgenomen.

2.3 Evaluatiefase

In de evaluatiefase hebben een aantal experts het prototype van het beoordelingssysteem voor competentie-assessments gebruikt voor de beoordeling van een competentie-assessment. Tijdens de evaluatie is onderzocht wat experts van het ontwerp en de hanteerbaarheid van het beoordelingssysteem vonden en is nagegaan in hoeverre zij overeenstemden in hun oordelen.

In de evaluatiefase zijn 14 experts op het gebied van toetsing en assessment benaderd om deel te nemen aan het onderzoek. In totaal waren 11 experts bereid aan het onderzoek mee te werken. Deze groep van experts bestond uit lectoren op het gebied van competentiebeoordeling en assessment, leden van de COTAN, leden van universitaire vakgroepen met specialismen op het gebied van beoordeling en assessment, toetsdeskundigen van Cito en een toetsdeskundige van een kenniscentrum voor het beroepsonderwijs.

De participerende experts hebben het beoordelingssysteem voor competentie-assessments getest door zelf een assessment te beoordelen. Het betrof het Centraal Schriftelijk en Praktisch Examen (CSPE) Transport en Logistiek uit 2006. Dit is een examen dat bij leerlingen van de basisberoepsgerichte leerweg van het vmbo afgenomen is. Tijdens dit examen voeren de leerlingen drie praktijkopdrachten uit, die worden afgewisseld met korte toetsen waarin de kennis van leerlingen middels meerkeuzevragen getoetst wordt. De beschikbare informatie over het CSPE is ten behoeve van het onderzoek verwerkt in een toetsverantwoording zodat het voor de experts mogelijk was om de kwaliteit van het examen te beoordelen. De toetsverantwoording werd conform de indeling van het beoordelingssysteem gepresenteerd om de belasting van de experts zo laag mogelijk te houden. Daarnaast werd bepaalde informatie in de toetsverantwoording bewust slecht gedocumenteerd of achterwege gelaten om na te gaan hoe de experts hiermee om zouden gaan.

Om na te gaan in hoeverre experts overeenstemmen in hun oordeel kan de interbeoordelaarsbetrouwbaarheid berekend worden. Een voorbeeld van een veelgebruikte maat voor interbeoordelaarsbetrouwbaarheid is de intraklassecorrelatiecoëfficiënt (ρ; Heuvelmans & Sanders, 1993). Maten voor interbeoordelaarsbetrouwbaarheid hebben echter als nadeel dat ze niet gevoelig zijn voor absolute verschillen in de oordelen van experts. Dit kan leiden tot een onderschatting van de mate waarin experts overeenstemmen wanneer de relatieve oordelen van experts gelijk zijn. Om deze onderschatting te voorkomen wordt de mate van overeenstemming tussen experts in dit onderzoek gekwantificeerd met behulp van de Gower-coëfficiënt (Zegers, 1989). Deze maat gaat uit van absolute verschillen tussen beoordelingen. Deze coëfficiënt geeft dus weer in hoeverre de experts daadwerkelijk dezelfde scores hebben toegekend en neemt, wanneer dit het geval is, de waarde 1 aan.

In de tweede fase van de gegevensverzameling waren 9 van de 11 experts beschikbaar om de uitgevoerde beoordeling van het assessment in een open (telefonisch) interview toe te lichten. In het interview werden onderwerpen met betrekking tot het *ontwerp* van het beoordelingssysteem en de *hanteerbaarheid* ervan aan de orde gesteld. In relatie tot het onderwerp *ontwerp* werd ingegaan op de volledigheid van het beoordelingssysteem en de nieuwe invulling van het begrip validiteit die in het beoordelingssysteem gehanteerd wordt. Ten aanzien van de *hanteerbaarheid* werd ingegaan op de mogelijkheid om een gefundeerd oordeel te geven over een competentie-assessment en de omvang van het beoordelingssysteem. Vervolgens zijn van alle interviews transcripten gemaakt. Uit deze transcripten zijn tekstfragmenten geselecteerd die zonder de context van het interview betekenisvol waren. Deze tekstfragmenten zijn vervolgens door twee onderzoekers onafhankelijk van elkaar in zeven vooraf opgestelde categorieën ingedeeld. De categorieën zijn gebaseerd op de operationalisatie van de begrippen *ontwerp* en *hanteerbaarheid* zoals die ook in het interview gebruikt is. Wanneer een tekstfragment door de

onderzoekers verschillend was ingedeeld, is in overleg besloten in welke categorie deze uitspraak uiteindelijk geplaatst werd. In Tabel 1 zijn zowel de vragen die ter operationalisatie van de begrippen ontwerp en hanteerbaarheid geformuleerd zijn, als de uiteindelijke categorieën weergegeven. Tevens is aangegeven hoeveel tekstfragmenten er in een bepaalde categorie zijn ingedeeld.

3 Resultaten

In deze paragraaf worden de uitkomsten van de verschillende fases beschreven. In de beschrijving van de verkenningsfase zal een samenvatting van de literatuurstudie gegeven worden. Deze beschrijving gaat in op de noodzaak voor een nieuw beoordelingssysteem voor competentie-assessments en op de ontwerpkeuzes die gemaakt zijn voor het prototype dat is ontwikkeld. Vervolgens zal in de beschrijving van de ontwikkelfase het ontwikkelde prototype beschreven worden. Ten slotte zal in de evaluatiefase een beschrijving van de uitkomsten van het evaluatieonderzoek gegeven worden.

3.1 Verkenningsfase

Om te bepalen of een bestaand beoordelingssysteem geschikt is om competentie-assessments te beoordelen, zijn zowel nationale als internationale systemen geïnventariseerd. De systemen worden in deze paragraaf vergeleken met vooraf opgestelde ontwerpeisen voor het beoordelingssysteem voor competentie-assessments, te weten:

– De beoordeling richt zich op toetstechnische kwaliteit van competentie-assessments. Een aantal systemen voor kwaliteitsbeoordeling richt zich op processen of procedures die gevolgd worden tijdens het toetsconstructieproces of de afname van toetsen. In het hier voorgestelde beoordelingssysteem zal echter alleen toetstechnische kwaliteit beoordeeld worden.
– Er wordt een enkelvoudig instrument beoordeeld. Dit betekent dat toetsen die alleen in combinatie met andere toetsen afgenomen kunnen worden, niet beoordeeld kunnen worden.
– Het beoordelingssysteem is geschikt om competentie-assessments te evalueren. Dit betekent dat het beoordelingssysteem recht moet doen aan toetsen waarin met

Tabel 1

Classificatiecategorieën voor tekstfragmenten uit de interviews

Onderzoeksvragen	Categorieën	Aantal fragmenten
In hoeverre kunnen experts zich vinden in de gekozen kwaliteitscriteria voor het beoordelen van assessments?	Ontwerp: algemeen	16
Zijn er cruciale kwaliteitsaspecten te noemen die niet in de gehanteerde criteria aan de orde komen?	Ontwerp: aspecten die niet aan de orde komen	14
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?	Ontwerp: Prioritering / criteria vervallen	5
Zijn experts het eens met de keuze voor de argument-based approach als benadering voor het hoofdonderdeel validiteit?	Ontwerp: mening validiteit	14
Is het instrument voor experts hanteerbaar?	Hanteerbaarheid: algemeen	14
Is het voor experts mogelijk om uit de beschikbare informatie over het assessment de juiste te selecteren?	Hanteerbaarheid: mogelijk informatie te selecteren	13
Vinden experts de omvang van het beoordelingsinstrument aanvaardbaar?	Hanteerbaarheid: omvang (tijd & hoeveelheid items)	15
Totaal		91

name open vragen, simulaties of praktijkgerichte taken worden voorgelegd. In deze toetsen heeft de beoordelaar meestal een grote rol.
– De beoordeling is concreet. Het betreft een analytisch beoordelingssysteem dat leidt tot een eindoordeel over de kwaliteit. Dit betekent dat de criteria verwoord zijn in concrete eisen waar een toets aan moet voldoen.

Roorda (2007) laat zien dat er zowel nationale als internationale initiatieven zijn waarbij kwaliteitscriteria voor toetsen ontwikkeld werden. De beoordelingssystemen variëren in het object van de beoordeling: er zijn beoordelingssystemen die de processen of procedures tijdens constructie, afname en verwerking van toetsen beoordelen en er zijn systemen die de toets op toetstechnische eisen beoordelen. Aangezien dit één van de ontwerpeisen betrof, is ervoor gekozen alleen de beoordelingssystemen die geschikt zijn om toetstechnische eisen te beoordelen mee te nemen in de inventarisatie. Verder verschillen de beoordelingssystemen in de mate waarin zij bepaalde kwaliteitseisen opleggen. Sommige beoordelingssystemen geven slechts richtlijnen aan, terwijl andere systemen harde eisen geven waaraan de toets moet voldoen. Bij het inventariseren van beschikbare beoordelingssystemen voor kwaliteitsbeoordeling is daarom een onderscheid gemaakt tussen richtlijnen (*guidelines*), standaarden (*standards*) en beoordelingsinstrumenten (*review systems*). Bij deze indeling zijn richtlijnen het minst en beoordelingsinstrumenten het meest streng. Opgemerkt moet worden dat de standaarden van de Inspectie van het Onderwijs niet in deze inventarisatie zijn meegenomen omdat in deze standaarden slechts één standaard betrekking heeft op de kwaliteit van toetsen en examinering. Vanwege de beperkte omvang van deze standaard is besloten om dit niet als een beoordelingssysteem voor kwaliteitsbeoordeling van toetsen te beschouwen.

Richtlijnen

Systemen die richtlijnen voor kwaliteit van toetsen bevatten, beschrijven wenselijk gedrag van toetsconstructeurs en toetsgebruikers. Zij bevatten echter geen kwaliteitscriteria waarmee een toets kan worden beoordeeld. Voorbeelden van richtlijnen zijn de "International Guidelines for Test Use" die zijn uitgebracht door de International Test Commission (Bartram, 2001) en de "Code of Fair Testing Practices in Education" (2004) uitgebracht door de Joint Committee of Testing Practices (2004)

Standaarden

Naast systemen met richtlijnen zijn er ook systemen die standaarden presenteren. De afspraken en criteria die zijn opgenomen in standaarden dienen door professionals opgevolgd te worden. Het bekendste internationale voorbeeld van standaarden is de "Standards for Educational and Psychological Testing", een gezamenlijk initiatief van de American Educational Research Association (AERA), American Psychological Association (APA), en de National Council on Measurement in Education (NCME; 1999. De verzameling criteria die Baartman en collega's (2006) onderscheiden kunnen ook als een verzameling standaarden gezien worden. Deze criteria zijn specifiek bedoeld voor competentie-assessments en zullen daarom ook in deze inventarisatie meegenomen worden.

Beoordelingsinstrumenten

Ten slotte worden beoordelingsinstrumenten onderscheiden. Deze beoordelingsinstrumenten bevatten concrete criteria die beoordeeld worden en leiden meestal tot een eindoordeel over de kwaliteit van een toets. Om tot dit eindoordeel te komen wordt een scoringsregel gehanteerd waarmee de gehanteerde criteria geaggregeerd kunnen worden tot één eindoordeel. Een internationaal voorbeeld van een beoordelingsinstrument is dat van de European Federation of Psychologists' Association (Lindley, Bartram, & Kennedy, 2004). In Nederland heeft de Commissie Testaangelegenheden Nederland (COTAN) een beoordelingssysteem uitgebracht. In tegenstelling tot alle voornoemde systemen leiden deze laatste systemen na beoordeling tot een uitspraak over de kwaliteit van de toets. In dit onderzoek is gebruik gemaakt van het COTAN beoordelingssysteem uit 2000, inmiddels is er een herziene versie verschenen

waarin delen van dit onderzoek verwerkt zijn (COTAN, 2010).

Vergelijking van systemen

De genoemde systemen voor kwaliteitsbeoordeling van toetsen zijn vergeleken op grond van de vastgestelde ontwerpeisen. Tabel 2 geeft deze vergelijking weer. In de tabel is met een plus (+) of min (–) aangegeven of een systeem voor kwaliteitsbeoordeling aan een bepaalde ontwerpeis voldoet.

Zoals eerder vermeld, zijn alle beoordelingssystemen die zijn opgenomen in Tabel 2 geschikt om toetstechnische kwaliteit beoordelen. Dit houdt in dat in alle systemen bijvoorbeeld gekeken wordt naar de kwaliteit van de items en de kwaliteit van de lay-out. Ook de betrouwbaarheid en de validiteit van een toets worden met alle systemen beoordeeld. Daarnaast zijn vrijwel alle beoordelingssystemen erop gericht een enkelvoudige toets te beoordelen. Alleen de methodiek van Baartman richt zich op het beoordelen van een assessmentprogramma waarin meerdere toetsen gecombineerd worden.

De methodiek van Baartman is ook de enige die zich expliciet richt op de kwaliteit van competentie-assessments. De andere systemen zijn ontworpen om de kwaliteit van schooltoetsen met meerkeuze vragen, met korteantwoordvragen of met zeer gestandaardiseerde open vragen te beoordelen.

Verder hebben veel van deze beoordelingssystemen een oorsprong bij psychologische tests. Dit heeft tot resultaat dat de beoordelingssystemen minder geschikt zijn om competentie-assessments te beoordelen. In competentie-assessments wordt bijvoorbeeld gewerkt met taken waarin kandidaten gedrag moeten demonstreren in complexe situaties. Om de kwaliteit van deze taken te beoordelen zal bijvoorbeeld de mate van authenticiteit van items of taken beoordeeld moeten worden, terwijl dit bij meerkeuzevragen veel minder relevant is. Verder zijn de beschikbare beoordelingssystemen veelal gericht op toetsen met een normgerichte interpretatie van de score. Dit houdt in dat de interpretatie van een score plaats vindt door de score van een kandidaat te vergelijken met de scores van een grote groep andere kandidaten. Deze manier van normeren is echter niet gebruikelijk voor competentie-assessments, waar de kandidaten vaak vergeleken worden met een vooraf opgestelde inhoudelijke standaard.

In Tabel 2 is tevens te zien dat alleen het COTAN- en het EFPA-beoordelingssysteem leiden tot een eindoordeel over de kwaliteit. De andere systemen geven slechts richtlijnen of mogelijke kwaliteitsaspecten aan en bevatten geen geoperationaliseerde criteria.

Uit Tabel 2 leiden we af dat de beoordelingssystemen van de COTAN en van de EFPA aan de meeste ontwerpeisen voldoen, maar dat geen enkel beoordelingssysteem aan alle eisen voldoet. Er is daarom besloten om een beoordelingssysteem aan te passen voor het beoordelen van competentie-assessments. Er is besloten om het COTAN-beoordelingssysteem hiervoor als uitgangspunt te gebruiken, omdat dit beoordelingssysteem in Nederland meer bekendheid geniet dan het EFPA-beoordelingssysteem. Het huidige COTAN-beoordelingssysteem wordt immers al gebruikt voor het beoordelen van schooltoetsen en toetsgebruikers zijn daarom al bekend met de structuur van COTAN-beoordelingen. Voor de aanpassing van het

Tabel 2

Vergelijking tussen beoordelingsinstrumenten

	Systemen voor kwaliteitsbeoordeling					
Ontwerpeisen	COTAN	EFPA	Baartman	ITC	APA	AERA, APA, NCME
Toetstechnische kwaliteit	+	+	+	+	+	+
Enkelvoudig instrument	+	+	-	+	+	+
Competentie-assessment	-	-	+	-	-	-
Leidt tot eindoordeel kwaliteit	+	+	-	-	-	-

COTAN-beoordelingssysteem zal onder andere gebruik gemaakt worden van de criteria zoals beschreven door Baartman en collega's (2006).

Het aanpassen van een beoordelingssysteem als van COTAN heeft een aantal praktische implicaties. Allereerst zal dezelfde beoordelingsmethodiek gehanteerd worden. Dit betekent dat de beoordeling zal plaatsvinden op basis van beschikbare documentatie en dat een toetsafname niet lijfelijk bijgewoond zal worden. Verder zal de COTAN indeling van criteria in vijf hoofdcategorieën ook gebruikt worden in het beoordelingssysteem voor competentie-assessments. Ook voor de toekenning van scores aan de criteria zal de COTAN-methodiek gevolgd worden (0 = *onvoldoende*, 1 = *voldoende* en 2 = *goed*).

Samengevat zijn in de verkenningsfase bestaande kwaliteitsbeoordelingssystemen met elkaar vergeleken. In deze vergelijking is rekening gehouden met vooraf vastgestelde ontwerpeisen. Op basis van de vergelijking is gekozen voor een aanpassing van het COTAN-beoordelingssysteem.

3.2 Ontwikkelfase

In de ontwikkelfase is het beoordelingssysteem voor competentie-assessments geschreven. Voor elke hoofdcategorie werd eerst bepaald, welke van de huidige COTAN-criteria ook geschikt zijn voor de beoordeling van competentie-assessments. Vervolgens werd bepaald welke nieuwe criteria nodig waren om recht te doen aan de specifieke eigenschappen van competentie-assessments en zodoende een valide beoordeling van de kwaliteit van competentie-assessments mogelijk te maken.

In het prototype van het beoordelingssysteem voor competentie-assessments zijn bijvoorbeeld criteria toegevoegd om te beoordelen of de te meten competentie voldoende omschreven is. Er wordt dus minder gelet op de beschrijving van het construct, zoals dat verwacht wordt bij psychologische tests, maar bijvoorbeeld meer op de inkadering van de competentie in het opleidingsprogramma.

Verder houdt het beoordelingssysteem voor competentie-assessments meer rekening met het open karakter van de responsen die in competentie-assessments van kandidaten verlangd worden. In een proeve van bekwaamheid zijn de handelingen die een kandidaat moet verrichten bijvoorbeeld vaak geplaatst in een authentieke situatie, waardoor deze handelingen vaak complex en minder gestandaardiseerd zijn. Het beoordelingssysteem is daarom ook geschikt om complexe observatieschema's of beoordelingsvoorschriften te beoordelen. Daarnaast wordt er rekening gehouden met het inzetten van assessoren en kunnen verschillende procedures om een cesuur vast te stellen beoordeeld worden.

Competentie-assessments leiden voor een kandidaat vaak tot een classificatie in termen van onvoldoende, voldoende, of beginner, gevorderd, expert. Daarnaast zullen er nog steeds competentie-assessments zijn die leiden tot een schoolcijfer (1 tot 10) of een andere toetsscore. Voor de beoordeling van bijvoorbeeld de betrouwbaarheid van de competentie-assessments die leiden tot een cijfer waren in het bestaande COTAN-beoordelingssysteem al criteria beschikbaar. Voor de competentie-assessments die leiden tot een classificatie zijn in het beoordelingssysteem voor competentie-assessments nieuwe criteria opgenomen die bedoeld zijn om bijvoorbeeld de betrouwbaarheid van classificatiebeslissingen te beoordelen.

Ten slotte is in het beoordelingssysteem voor competentie-assessments het validiteitsbegrip gemoderniseerd. In het COTAN-beoordelingssysteem werd validiteit gesplitst in inhouds-, begrips- en criteriumvaliditeit, zoals beschreven door Messick (1989). Voor schooltoetsen en ook voor competentie-assessments is criteriumvaliditeit echter niet altijd van belang. Dit geldt bijvoorbeeld wanneer we slechts willen beoordelen wat een leerling in de les heeft opgestoken en geen voorspelling willen doen over toekomstig gedrag. Het beoordelingssysteem voor competentie-assessments hangt daarom de meer flexibele argumentatieve benadering van validiteit aan (Kane, 2004, 2006). Deze benadering geeft meer ruimte om tijdens de validering, en ook bij de beoordeling van de validiteit, aan te sluiten bij het specifieke doelen van de toets die gevalideerd wordt.

Hieronder worden de zes hoofdcategorieën van het prototype van het beoordelingssysteem voor competentie-assessments kort

beschreven. Voor een volledig overzicht van de indicatoren behorende bij de zes criteria wordt verwezen naar Appendix 1.

Criterium 1. Uitgangspunten van de toetsconstructie
Bij dit criterium wordt nagegaan of de ontwikkelaar van het assessment de gemaakte keuzes gespecificeerd en verantwoord heeft. Dit betreft keuzes als de functie van het assessment en de doelgroep waarvoor het assessment bedoeld is. Daarnaast wordt gevraagd naar de mate van operationalisatie van de competentie(s) die het assessment geacht wordt te meten.

Criterium 2. Kwaliteit van het toetsmateriaal en de handleiding
Dit criterium heeft betrekking op de kwaliteit van het toetsmateriaal en de kwaliteit van de handleiding. De kwaliteit van het toetsmateriaal heeft betrekking op de inhoud, het ontwerp en de vormgeving van een toets. Inhoudelijk verschillen de meeste tests en toetsen van competentie-assessments doordat kandidaten bij assessments authentieke taken moeten uitvoeren in complexe beroepssituaties. Vandaar dat bij dit criterium de authenticiteit van het assessment nagegaan wordt. Die authenticiteit wordt afgemeten aan de mate waarin het assessment representatief is voor de criteriumsituatie, dat wil zeggen de situatie zoals die redelijkerwijs in de praktijk te verwachten is na het afronden van de opleiding (Gulikers, 2006; Gulikers, Bastiaens, & Kirschner, 2004). Bij dit criterium wordt ook beoordeeld of het assessment een gestandaardiseerde beoordeling van de prestaties van kandidaten mogelijk maakt. Wat betreft de vormgeving van het assessment wordt nagegaan of het assessment niet onnodig ingewikkeld vormgegeven is, zodat voorkomen wordt dat kandidaten onnodig fouten maken.

De kwaliteit van de handleiding betreft de informatievoorziening aan de onderscheiden betrokkenen bij het assessment. Vastgesteld moet worden of er voldoende informatie verstrekt wordt aan toetsgebruikers of assessoren. Daarnaast wordt in deze versie van het beoordelingssysteem uitgebreid ingegaan op de transparantie van het assessment ten opzichte van kandidaten.

Criterium 3. Normen en standaarden
In de examenpraktijk worden toetsen en assessments zowel relatief als absoluut genormeerd. In geval van relatief normeren wordt vooraf een bepaald slagingspercentage vastgesteld (Hambleton & Pitoniak, 2006), terwijl bij absoluut normeren de zak/slaag beslissing op basis van vooraf vastgestelde beheersingsstandaarden (Haertel & Loiré, 2004) genomen wordt. Mengvormen hiervan zijn ook mogelijk waarbij een vooraf vastgestelde beheersingsstandaard soms aangepast wordt naar aanleiding van tegenvallende prestaties van de kandidaten. Met het beoordelingssysteem kunnen competentie-assessments beoordeeld worden die gebruik maken van relatief of absoluut normeren. De criteria voor de beoordeling van relatief genormeerde toetsen richten zich op de grootte en de representativiteit van normgroepen. Terwijl de criteria voor absoluut genormeerde toetsen zich richten op de kwaliteit van de standaardbepalingsprocedure.

Criterium 4. Betrouwbaarheid
Bij dit criterium wordt niet alleen uitgegaan van toetsen of assessments bestaande uit gesloten vragen of taken, maar ook van assessments bestaande uit veelal complexe taken met een open karakter die door beoordelaars beoordeeld dienen te worden. In het laatste geval wordt bij de kwantificering van de betrouwbaarheid nagegaan wat de invloed van (verschillen tussen) beoordelaars op de betrouwbaarheid van het assessment is. In het beoordelingssysteem wordt gegeven de bestaande beoordelingspraktijk de voorkeur gegeven aan een kwantificering van betrouwbaarheid in termen van misclassificaties, dat wil zeggen percentages ten onrechte gezakte en ten onrechte geslaagde kandidaten (Van Rijn, Béguin, & Verstralen, 2009).

Criterium 5. Validiteit
Vergeleken met het COTAN-beoordelingssysteem heeft het criterium validiteit de grootste verandering ondergaan. In het beoordelingssysteem is gekozen voor de implementatie van de argumentatieve benadering (*argument-based approach*) voor valideren van Kane (2004, 2006). In deze benadering wordt eerst geëxpliciteerd welke inferenties

nodig zijn voor een valide beslissing en vervolgens worden deze inferenties met (empirische) bewijzen onderbouwd. In het beoordelingssysteem wordt nagegaan of de juiste inferenties beschreven zijn en of de gecombineerde bewijzen overtuigend genoeg zijn om een valide beslissing te garanderen (Wools, Eggen, & Sanders, 2010).

In de ontwikkelingsfase is het prototype van het beoordelingssysteem voor competentie-assessments ontwikkeld door oude en nieuwe beoordelingscriteria samen te voegen. De oude criteria waren afkomstig uit het bestaande COTAN-beoordelingssysteem, de nieuwe criteria zijn geformuleerd om recht te doen aan de praktijk van competentie-assessments.

3.3 Evaluatiefase

In de evaluatiefase is het beoordelingssysteem door 11 experts gebruikt om tot een oordeel over de kwaliteit van een Centraal Schriftelijk en Praktisch Examen (CSPE) Transport en Logistiek te komen. In deze paragraaf worden de resultaten van de beoordeling van het CSPE Transport en Logistiek beschreven. Daarnaast wordt ingegaan op de meningen van de experts over het ontwerp en de hanteerbaarheid van het beoordelingssysteem en op de mate van overeenstemming in de oordelen van de experts.

Resultaat beoordelingen CSPE

Het beoordelingssysteem resulteert in een (eind)beoordeling op zes criteria. De resultaten van de beoordelingen staan in Tabel 3. Een beoordeling is gebaseerd op het aantal scorepunten dat een expert aan de indicatoren behorende bij een bepaald criterium heeft toegekend. Per indicator konden maximaal twee scorepunten worden toegekend. Wanneer een expert van ten minste één indicator aangeeft dat deze niet beoordeelbaar is, kan er geen (eind)beoordeling bepaald worden. In Tabel 3 wordt een ontbrekende beoordeling weergegeven als M (*missing*).

Tabel 3 laat zien dat de experts de kwaliteit van het assessment op veel criteria als onvoldoende beoordeelden. Opmerkelijk is dat de uitgangspunten van de testconstructie door de meeste experts als onvoldoende beoordeeld worden terwijl de beschrijving van de uitgangspunten vergelijkbaar is met de informatie die jaarlijks over honderden (eind) examens verstrekt wordt. Dat de kwaliteit van de handleiding door sommige experts als voldoende beoordeeld is, is verrassend omdat een handleiding bij dit assessment in feite ontbrak. Over de kwaliteit van de andere criteria lopen de beoordelingen van de experts uiteen van *onvoldoende* tot *goed*. Dat bij expert G beoordelingen op vijf criteria ontbreken, komt doordat deze expert het niet eens was met het toekennen van scorepunten aan de specifieke indicatoren behorende bij de criteria. Expert K heeft de laatste drie criteria niet beoordeeld vanwege zijn naar eigen zeggen onvoldoende kennis van de psychometrie. De overige ontbrekende waarden worden veroorzaakt door het om uiteenlopende redenen ontbreken van een oordeel over één of twee indicatoren waardoor de gehanteerde

Tabel 3

Beoordelingen op zes criteria door 11 experts

	Experts											
Criteria	A	B	C	D	E	F	G	H	I	J	K	Aantal oordelen
1. Uitgangspunten van testconstructie	1	1	1	2	1	2	2	1	1	1	1	11
2a. Kwaliteit van het testmateriaal	3	1	2	3	1	1	X	1	M	1	2	9
2b. Kwaliteit van de handleiding	1	1	1	1	2	2	X	1	1	2	1	10
3. Normen en standaarden	1	M	1	3	1	2	X	2	1	2	X	8
4. Betrouwbaarheid	1	M	1	2	1	2	X	3	M	1	X	7
5. Validiteit	3	3	1	1	1	2	X	3	M	1	X	8

Noot. 1 = *onvoldoende*, 2 = *voldoende*, 3 = *goed*. M = 1 of 2 indicatoren ontbreken, X = geen enkele indicator beoordeeld.

scoringsregel van het beoordelingssysteem niet toegepast kon worden.

Vanwege het open karakter van de interviews zijn niet altijd dezelfde onderwerpen aan bod gekomen zodat bij de presentatie van de resultaten expliciet aangegeven wordt, wanneer experts een bepaalde visie deelden.

Ontwerp

Een aantal experts heeft opmerkingen gemaakt over indicatoren die ontbreken of over aspecten van het assessment die ze middels het beoordelingssysteem niet konden beoordelen. Zo miste expert D een indicator voor het beoordelen van de inhoud van de taak en informatie over de manier waarop uit een voorgelegde taak bewijzen voor het competent handelen van een kandidaat verzameld konden worden**.** Expert E miste vooral criteria die nagaan of er rekening wordt gehouden met gevolgen voor het leren van de student (consequentiële validiteit) en een criterium om na te gaan of de competentie ook daadwerkelijk gemeten wordt. Opgemerkt kan worden dat dit als onderdeel van validiteit beoordeeld moet worden. Twee experts geven suggesties voor nieuwe criteria of indicatoren die eisen bevatten ten aanzien van transparantie, consequentiële validiteit, de aanvaardbaarheid van verschillende betrokkenen bij het assessment en de praktische bruikbaarheid van assessments.

Een aantal experts zijn voorstander van een prioritering van criteria of indicatoren. Expert E en G zijn van mening dat het criterium *validiteit* zwaarder zou moeten wegen, terwijl expert C meer gewicht wil toekennen aan de *betrouwbaarheid*.

Ten aanzien van het criterium validiteit geven experts A, C en D aan dat een aantal indicatoren reeds eerder in het beoordelingssysteem aan de orde kwamen. Dat vindt expert C onterecht, omdat het assessment hierdoor twee keer hetzelfde 'beloont' of 'straft'. Verder geven vier experts aan (A, E, H, I) de keuze voor de argumentatieve benadering van valideren te onderschrijven. Expert B geeft aan deze benadering lastig te vinden, mede doordat deze expert de toelichting in het beoordelingssysteem erg abstract en weinig concreet vond. Expert C was erg negatief over deze benadering van het valideringsproces, omdat deze benadering te weinig nadruk zou leggen op het verzamelen van empirisch bewijs voor het aantonen van de validiteit.

Ten slotte zijn een aantal algemene opmerkingen gemaakt over het ontwerp van het beoordelingssysteem. Experts H en I geven bijvoorbeeld aan dat zij meer beoordelingsmogelijkheden zouden willen hebben dan *onvoldoende*, *voldoende*, *goed*, waarbij te denken valt aan *zeer onvoldoende*, *matig* en *ruim voldoende*. Verder hebben drie experts aangegeven dat hun beoordeling van het assessment met het beoordelingssysteem in overeenstemming was met hun eerste globale indruk van de kwaliteit van het assessment. Expert C, die volgens Tabel 3 alleen de kwaliteit van het testmateriaal voldoende vond, gaf aan dat zijn eerste globale indruk positiever was dan zijn beoordeling door middel van het beoordelingssysteem.

Hanteerbaarheid

Veel experts geven aan dat zij voor de beoordeling tussen vier en zes uur nodig hadden. Experts C en D geven aan een paar dagen nodig gehad te hebben, omdat zij het beoordelingssysteem grondiger bekeken hebben. De experts B en G waren van mening dat er te veel criteria beoordeeld moesten worden en vonden dan ook dat het beoordelen van het assessment teveel tijd vergde. Expert B gaf aan dat dit te wijten was aan het zoeken in de documentatie. De andere experts vonden de tijd die nodig was voor het beoordelen acceptabel.

Uit de interviews is af te leiden dat experts hun informatie voornamelijk uit de verantwoording van het assessment en het bijgevoegde materiaal hebben gehaald. Nagegaan is hoe experts omgaan met het beoordelen van indicatoren waar geen informatie voor beschikbaar was, zoals de lokale betrouwbaarheid, de relatieve normen en de handleiding voor kandidaten. De indicator over lokale betrouwbaarheid is door vijf experts beoordeeld, terwijl de toetsverantwoording hier geen enkele informatie over bevatte. Opvallend is dat die indicator door twee experts zelfs als *goed* is beoordeeld. Vier experts hebben de indicator niet beoordeeld of *niet van toepassing* ingevuld. De overige drie experts hebben de indicator met *onvoldoende*

beoordeeld. Bij het criterium normen en standaarden moesten experts kiezen welke onderdelen van toepassing waren. Hoewel de indicatoren bij het onderdeel normen niet beoordeeld hoefden te worden, omdat er in het assessment geen sprake was van relatieve normering is dit door vijf experts toch gedaan. Vier experts hebben de betreffende indicatoren niet beoordeeld of *niet van toepassing* ingevuld. Hoewel de informatie voor kandidaten bij de assessmenttaken minimaal was, zijn de indicatoren hierover door alle experts beoordeeld. Slechts één expert heeft op de vraag of er informatie ten behoeve van kandidaten beschikbaar was geantwoord dat dit niet het geval was. Drie experts hebben aangegeven dat er informatie beschikbaar was, maar zij hebben een opmerking gemaakt over de summiere omvang. De overige experts hebben alle indicatoren over de informatie voor kandidaten beantwoord.

Ten slotte gaan twee experts (D en E) 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 criteria betrouwbaarheid en validiteit. Expert D is van mening dat erg specifieke kennis over assessments vereist is. De experts E, H en I gaven aan dat de indicatoren veel interpretatie vergden en dat niet altijd duidelijk was of met een *onvoldoende*, *voldoende* of *goed* beoordeeld moest worden.

Betrouwbaarheid

Tabel 4 bevat per criterium de gemiddelde Gower-coëfficiënt van beoordelingen van alle paren experts voor de onderscheiden indicatoren bij een criterium. Een gemiddelde Gower-coëfficiënt van 1 betekent dat er sprake is van perfecte overeenstemming, terwijl 0 betekent dat er geen overeenstemming is. De overeenstemming tussen beoordelaars is berekend op hun toegekende scores op de indicatoren. Bij de analyse zijn experts G en K buiten beschouwing gelaten, omdat zij over respectievelijk 5 en 3 criteria geen oordeel hebben gegeven.

Tabel 4 laat zien dat de overeenstemming tussen de experts bij alle criteria relatief laag is (< 0,74) en dat de overeenstemming bij het criteria normen en standaarden en betrouwbaarheid duidelijk lager is dan bij de overige criteria. Tevens zijn per criterium de meest afwijkende paren experts geïdentificeerd, dat wil zeggen paren experts waarvan de overeenstemming zeer laag was. Tabel 4 laat zien dat verwijdering van afwijkende paren experts invloed had op de hoogte van de overeenstemming, met name voor de criteria *normen* en *standaarden* en *betrouwbaarheid*.

In de evaluatiefase van het ontwerponderzoek is het prototype van het beoordelingssysteem voor competentie-assessments door experts gebruikt om de kwaliteit van een CSPE te beoordelen. De experts gaven suggesties voor het toevoegen of verwijderen van criteria en voor het aanpassen van de gehanteerde beoordelingsmethodiek. Daarnaast bleek dat experts verschillend omgingen met het beoordelen van criteria waarover geen informatie beschikbaar was. Deze verschillende oordelen kwamen ook naar voren in de relatief lage interbeoordelaarsovereenstemming die gevonden is.

Tabel 4

Gemiddelde Gower-coëfficiënt per criterium

	Alle paren experts		Zonder afwijkende paren experts	
Criteria	*N*	Gower	*N*	Gower
1 Uitgangspunten van testconstructie	9	0,74	8	0,76
2a Kwaliteit van het testmateriaal	9	0,70	6	0,76
2b Kwaliteit van de handleiding	9	0,71	7	0,74
3 Normen en Standaarden	9	0,58	7	0,70
4 Betrouwbaarheid	9	0,53	6	0,71
5 Validiteit	9	0,70	7	0,78

4 Conclusie en discussie

In dit artikel werd de ontwikkeling van een beoordelingssysteem voor competentie-assessments beschreven. De ontwikkeling is gestuurd door principes van ontwerponderzoek. In deze conclusie zal, zoals gebruikelijk bij ontwerponderzoek (Plomp, 2007), ingegaan worden op de ervaringen die tijdens de ontwikkeling van het instrument zijn opgedaan. Centraal staat hierbij of het beoordelingssysteem in de huidige vorm geschikt is voor de beoordeling van kwaliteit van competentie-assessments. Daarnaast zullen de lessen die tijdens de ontwikkeling en evaluatie van het beoordelingssysteem geleerd zijn beschreven worden. Ten slotte zullen een aantal discussiepunten aan de orde komen waarbij tevens verwezen wordt naar algemene problemen van kwaliteitsbeoordeling, daarnaast zullen er kanttekeningen bij het beschreven onderzoek geplaatst worden.

4.1 Geschiktheid beoordelingssysteem voor competentie-assessments

Hoewel de experts vinden dat er geen fundamentele veranderingen van het beoordelingssysteem voor competentie-assessments nodig zijn, is het systeem in de huidige vorm nog niet geschikt om de kwaliteit van competentie-assessments te beoordelen. De experts zijn bijvoorbeeld van mening dat er nog inhoudelijke aanpassingen aan de criteria gedaan moeten worden. Daarnaast kan de bruikbaarheid van het beoordelingssysteem verhoogd worden door een eenvoudigere toelichting op de criteria te schrijven, door een verbeterde lay-out en door automatisering van het beoordelingssysteem zodat experts niet meer zelf de (eind)beoordelingen hoeven te berekenen. Verder is de huidige lage interbeoordelaarsovereenstemming een punt van zorg. Er zijn maatregelen denkbaar die kunnen bijdragen aan het verhogen van deze overeenstemming. Te denken valt aan een training of een verbeterde bruikbaarheid van het instrument. Het is echter niet zeker dat deze maatregelen de interbeoordelaarsovereenstemming voldoende zullen verhogen.

4.2 *Lessons learned*

In de evaluatiefase van de ontwikkeling van het beoordelingssysteem voor competentie-assessments bleek dat experts moeilijkheden ondervonden bij het beoordelen van de validiteit van het assessment. Dit heeft gedeeltelijk te maken met de toenemende complexiteit van de steeds meeromvattende theorieën over validiteit (Lissitz & Samuelsen, 2007). Er is in het denken over validiteit een patstelling ontstaan die door Borsboom, Mellenbergh, en Van Heerden (2004) omschreven is als “the theory fails to serve either the theoretically oriented psychologist or the practically inclined tester” (p.1061). Aangezien de beschrijving van validiteit in het beoordelingssysteem ook onvoldoende houvast bood voor beoordelaars, is het niet verwonderlijk dat de beoordeling problemen opleverde. Het is daarom belangrijk om de gekozen validiteitsbenadering nog verder uit te werken in heldere en eenduidige beoordelingscriteria (Wools, Eggen, & Sanders, 2010).

Naast de moeilijkheden die experts bij de beoordeling van het criterium validiteit ondervonden, gaf een aantal experts aan problemen te hebben met het beoordelen van betrouwbaarheid. De experts schrijven dit toe aan een gebrek aan expertise met betrekking tot de psychometrie. Psychometrie is echter een groot onderdeel in het beoordelingssysteem en valt niet weg te denken uit een complete beoordeling van de kwaliteit van competentie-assessments. Er zal daarom overwogen worden om de beoordeling in twee delen te splitsen, een deel waarin de psychometrische kwaliteit van een competentie-assessment beoordeeld wordt en een deel waarin de inhoud en vorm van het competentie-assessment beoordeeld worden. De kwaliteitsbeoordeling wordt dan een product van overleg tussen twee experts die beiden hun eigen expertise in kun brengen om tot een beoordeling van kwaliteit te komen.

4.3 Discussie

In het licht van het beschreven onderzoek kan de vraag gesteld worden in hoeverre de huidige assessmentpraktijk geschikt is voor de hier voorgestelde beoordeling van de kwaliteit van assessments. Zo was het selecteren van een assessment dat voorgelegd kon worden aan de assessmentexperts geen eenvou-

dige aangelegenheid. Dit was enerzijds het gevolg van de specifieke eisen die aan het assessment gesteld werden in verband met het onderzoek, maar anderzijds omdat er weinig goed gedocumenteerde assessments beschikbaar waren die op de zes criteria beoordeeld konden worden. Vanwege het gebrek aan goed gedocumenteerde assessments is het nu niet mogelijk om de kwaliteit van assessments die momenteel in het onderwijs gebruikt worden, middels het onderzochte beoordelingssysteem te documenteren. De praktijk doet echter vermoeden dat er nog een groot hiaat ligt tussen de kwaliteitseisen die in literatuur aan assessments gesteld worden (Baartman et al., 2006; Baartman, Prins, Kirschner, & Van der Vleuten, 2007; Dierick & Dochy, 2001) en de kwaliteitseisen waaraan assessments daadwerkelijk voldoen. Er zou dan ook onderzoek gedaan moeten worden naar de huidige stand van zaken met betrekking tot de kwaliteit van assessments in het onderwijs.

Naast de vraag naar de huidige stand van zaken met betrekking tot de kwaliteit van gebruikte assessments, is het tevens van belang om de kwaliteit van nieuw te ontwikkelen assessments te waarborgen. Hoewel het formuleren van kwaliteitscriteria voor assessments een stap in de goede richting is, zal dit alleen van invloed zijn op de kwaliteit van assessments wanneer de criteria nageleefd moeten worden en er dus sprake is van zogenaamde *compliance* (Wise, 2006). *Compliance* kan bereikt worden door consequenties te verbinden aan de beoordelingen van experts. In Nederland gebeurt dit al bij de verwijzing naar het speciaal basisonderwijs waar alleen toetsen die door de COTAN als voldoende zijn beoordeeld voor dit doel ingezet mogen worden (Resing, Evers, Koomen, Pameijer, & Bleichrodt, 2005). Deze eis kan ook ingevoerd worden bij het gebruik van competentie-assessments voor kwalificerende beslissingen. Het is dan uiteraard wel zaak om de juiste criteria te hanteren, zodat een valide beoordeling van toetskwaliteit gewaarborgd wordt.

Tevens kan de fundamentele vraag of het mogelijk is om via een studie van papieren documenten een uitspraak te doen over toetskwaliteit gesteld worden. De keuze voor een papieren beoordeling in dit onderzoek kwam voort uit de beslissing om de beoordelingsmethodiek van COTAN zo veel mogelijk te volgen. Dit neemt niet weg dat er in zijn algemeenheid vragen gesteld kunnen worden bij de geschiktheid van een documentstudie voor kwaliteitsbeoordeling. Voordelen van deze manier van kwaliteitsbeoordeling zijn bijvoorbeeld de beperkte belasting van experts, zij hoeven immers niet fysiek een assessment bij te wonen. Daarnaast is het via een documentstudie ook mogelijk om kwantitatieve bewijzen van bijvoorbeeld betrouwbaarheid, of de samenstelling van normgroepen, zo goed mogelijk te beoordelen. Een nadeel van een documentstudie ter kwaliteitsbeoordeling is bijvoorbeeld de toegenomen werklast voor toetsconstructeurs (toetsuitgevers, onderwijsinstellingen, docenten) om overzichtelijke documenten aan te leveren. Verder wordt via deze manier van kwaliteitsbeoordeling de daadwerkelijke implementatie van het competentie-assessment, of de context waarin het assessment wordt afgenomen niet meegewogen. De kwaliteitsbeoordeling zoals die door het beoordelingssysteem voor competentie-assessments wordt uitgevoerd leidt alleen tot een uitspraak over de kwaliteit bij implementatie zoals voorzien door de uitgever. Uit onderzoek (Gulikers, Biemans, & Mulder, 2009) blijkt echter dat de daadwerkelijke implementatie vaak bij verschillende docenten verschilt. Hierdoor kan de kwaliteit zoals onderzocht dus niet altijd gegarandeerd worden. Uiteindelijk is een mengvorm waarbij zowel naar de beschikbare documenten, als naar de implementatie van competentie-assessments in onderwijsinstellingen gekeken wordt, het meest wenselijk. Maar de vraag is of dit, gezien de werkdruk die dit voor beoordelaars zou betekenen, een haalbare oplossing is. Tot die tijd lijkt het voorwaardelijk maken van een voldoende op een documentstudie een andere mogelijkheid. Wanneer uit de documentstudie al blijkt dat de kwaliteit van competentie-assessments onvoldoende is, is het ongeacht de implementatie onverantwoord om het instrument in te zetten voor kwalificerende beslissingen over kandidaten.

In het beschreven onderzoek zijn de eerste stappen in de ontwikkeling van een beoorde-

lingssysteem voor competentie-assessments gezet. Uit de evaluatie van het ontwikkelde prototype bleek dat het ontwikkelde beoordelingssysteem nu nog niet geschikt is om in te zetten in de dagelijkse toetspraktijk. Het is noodzakelijk om het huidige prototype verder te verbeteren op basis van de opmerkingen die experts tijdens de evaluatiefase gemaakt hebben. Tevens is er in de uitgevoerde evaluatiefase slechts beperkt onderzoek gedaan naar de kwaliteit van het beoordelingssysteem voor competentie-assessments. De kleinschaligheid van het huidige ontwerponderzoek leidt er toe dat het huidige prototype slechts gezien mag worden als een eerste aanzet om te komen tot een geschikt beoordelingssysteem voor competentie-assessments.

In een latere fase is het noodzakelijk om de definitieve versie van het beoordelingssysteem uitgebreider te evalueren. Er kan bijvoorbeeld nagegaan worden in hoeverre het nieuwe beoordelingssysteem meer geschikt is voor het beoordelen van competentie-assessments dan bestaande systemen. Daarnaast zal moeten worden onderzocht of het mogelijk is om met dit nieuwe beoordelingssysteem alle voorkomende competentie-assessments naar tevredenheid te beoordelen. Ten slotte is het noodzakelijk dat de kwaliteitseisen die we opleggen aan competentie-assessments ook gelden voor het beoordelingssysteem, of zoals men in het Engels zegt "you've got to practice what you preach".

Literatuur

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

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

Baartman, L. K., Prins, F. J., Kirschner, P. A., & Vleuten, C. P. van der. (2007). Determining the quality of competence assessment programs: A self-evaluation procedure. *Studies in Educational Evaluation*, *33*, 258-281.

Bartram, D. (2001). The development of international guidelines on test use: the international test commission project. *International Journal of Testing, 1*(1), 33-53.

Borsboom, D., Mellenbergh, G. J., & Heerden, J. van. (2004). The concept of validity. *Psychological Review*, *111*, 1061-1071.

COTAN. (2010). *COTAN Beoordelingssysteem voor de kwaliteit van tests. Geheel herziene versie, mei 2009; gewijzigde herdruk mei 2010.* Amsterdam: NIP.

Cronbach, L. J. (1989) Construct validation after thirty years. In R. Linn (Ed.), *Intelligence: Measurement, theory and public policy* (pp. 147-171). Chicago, Ill: University of Illinois Press.

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., & Watering, G. van de. (2001). Over de implicaties van nieuwe toetsvormen voor de edumetrie. *Tijdschrift voor Hoger Onderwijs*, *19,* 2-18.

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

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

Gulikers, J. T. M., Bastiaens, T. J., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Design, 57,* 67-87.

Gulikers, J. T. M., Biemans, H., & Mulder, M. (2009). Developer, teacher, student and employer evaluations of competence-based assessment quality. *Studies in Educational Evaluation, 35*¸ 110-119.

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. Brennan (Ed.), *Educational Measurement 4th edition* (pp. 433-470). Westport, CT: American Council on

Education and Praeger Publishers.
Heuvelmans, A. P. J. M., & Sanders, P. J. (1993). Beoordelaarsovereenstemming. In T. Eggen & P. Sanders (red.), *Psychometrie in de Praktijk* (pp. 443-470). Arnhem, Nederland: Cito.
Inspectie van het Onderwijs. (2009). *Normenbundel Exameninstrumentarium 2009*. Amersfoort, Nederland: Inspectie van het Onderwijs.
Joint Committee on Testing Practices. (2004). *Code of Fair Testing Practices in Education.* Washington, DC: Joint Committee on Testing Practices.
Kane, M. T. (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. Brennan (Ed.), *Educational Measurement 4th edition* (pp. 17-64). Westport, CT: American Council on Education and Praeger Publishers.
Lindley, P., Bartram, D., & Kennedy, N., (2004). *EFPA review model for the description and evaluation of psychological tests*. Opgehaald op 9 december 2010, van http://www.efpa.eu/professional-development/tests-and-testing.
Linn, R. L., Baker, E., & Dunbar, S. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, *16,* 1-21.
Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, *36*, 437-448.
Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). New York: American Council on Education and Macmillan.
Messick, S. (1994). The interplay of evidence and consequences in the validation performance assessments. *Educational Researcher, 23,* 13-22.
Plomp, T. (2007). Educational design research: an introduction. In T. Plomp & N. Nieveen (Eds.), *An introduction to educational design research* (pp. 9-35). Enschede, Nederland: SLO.
Resing, W. C. M., Evers, A. V. A. M., Koomen, H. M. Y., Pameijer, N. K., & Bleichrodt, N. (2005). *Indicatiestelling Speciaal Onderwijs een Leerlinggebonden Financiering: Condities en instrumentarium.* Amsterdam: Boom.
Rijn, P. van, Béguin, A., & Verstralen, H. (2009). Zakken of slagen? De nauwkeurigheid van examenuitslagen in het voortgezet onderwijs. *Pedagogsiche Studiën, 86*, 185-195.
Roelofs, E., & Straetmans, G. (red.) (2006). *Assessment in Actie: Competentiebeoordeling in opleiding en beroep.* Arnhem, Nederland: Cito.
Roorda, M., (2007). Quality systems for tests. In R. Ramaswamy & C. Wild (Eds.), *Improving Testing: process tools and techniques to assure quality* (pp. 145-176). Hillsdale, NJ: Lawrence Erlbaum.
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, Nederland: Open Universiteit Nederland.
Wise, L. (2006). Encouraging and supporting compliance with standards for educational tests. *Educational Measurement: Issues and Practice*, *25,* 51-53.
Wools, S., Eggen, T., & Sanders, P.(2010). Evaluation of validity and validation by means of the argument-based approach. *CADMO, 8,* 63-82.
Zegers, F. E. (1989). Het meten van overeenstemming. *Nederlands Tijdschrift voor de Psychologie, 44,* 145-156.

Manuscript aanvaard op: 6 december 2010

Auteurs

Saskia Wools is promovenda en Toetsdeskundige bij Cito in Arnhem. **Piet Sanders** is directeur van het Research Center voor Examinering en Certificering. **Theo Eggen** is bijzonder hoogleraar Psychometrie aan de Universiteit Twente en senior Toetsdeskundige bij Cito in Arnhem. **Liesbeth Baartman** is postdoctoral onderzoeker aan de Eindhoven School of Education.
Erik Roelofs is senior Toetsdeskundige bij Cito in Arnhem.

Correspondentieadres: Saskia Wools, Cito, Postbus 1034, 6801 MG Arnhem, E-mail: saskia.wools@cito.nl.

Abstract

Testing an evaluation system for performance tests

In this article, the development of a review system for the evaluation of competence assessment is described. The development process followed the principles of educational design research. The description of this study follows three phases. In the first phase, the design task is elaborated. In the second phase, the prototype is described and in the third phase an evaluation study is described. In this evaluation study, eleven assessment experts tested the review system by evaluating an actual assessment. We conclude that the current review system is still in need of improvement; particularly, the workability and the interrater agreement should be improved.

Appendix 1:

Indicatoren van het beoordelingssysteem: kwaliteit van competentie-assessments

Uitgangspunten van testconstructie

1.1 Is aangegeven wat de functie of het gebruiksdoel van het assessment is?*
1.2 Is aangegeven wat de doelgroep(en) is (zijn) van het assessment?
1.3 Is aangegeven welke competentie(s) het assessment beoogt te meten?
1.4 Wordt de relevantie van de inhoud van het assessment voor de te meten competentie(s) aannemelijk gemaakt?*
1.5 Worden theorieën en concepten die aan het assessment ten grondslag liggen besproken?

Kwaliteit van het testmateriaal

Inhoud

2.1 Is het assessment representatief voor de criteriumsituatie?
2.2 Is het assessment erop gericht de competentie als geheel te meten/beoordelen?

Ontwerp

2.3 Is het assessment gestandaardiseerd?*
2.4 Maakt het assessment gebruik van een beoordelings- of observatieschema dat de kans op verschillende beoordelingen van beoordelaars minimaliseert?

Vormgeving

2.5 Is de vormgeving van het assessment zodanig dat het de uitvoering ondersteunt of in ieder geval niet belemmert?
2.6 Is het assessment geschikt voor gebruik door verschillende groepen kandidaten?

Kwaliteit van de handleiding

Basisvraag

2.7 Is er een handleiding beschikbaar?*

Informatieverstrekking

2.8 Wordt er informatie gegeven over de gebruiksmogelijkheden en beperkingen van het assessment?*
2.9 Wordt er informatie gegeven over de vereiste condities bij de afname van het assessment?
2.10 Wordt er informatie gegeven over de vereiste deskundigheid voor afname en interpretatie van het assessment?*
2.11 Wordt er informatie gegeven over de interpretatie van de scores?*

Informatie voor assessoren

2.12 Zijn de aanwijzingen voor de assessor volledig en duidelijk?
2.13 Is de afnameprocedure van het assessment duidelijk beschreven?

Informatie voor kandidaten

2.14 Is voor kandidaten het doel van het assessment vooraf duidelijk?
2.15 Zijn voorafgaand aan het assessment de beoordelingscriteria voldoende geëxpliciteerd?
2.16 Is in de handleiding de afnameprocedure bij de kandidaten voldoende beschreven?

* Deze indicator is ook in het COTAN beoordelingssysteem voor de Kwaliteit van Tests opgenomen.

Normen en Standaarden

Normen

3.1 Is (Zijn) de gebruikte normgroep(en) representatief?*

3.2 Is de steekproefgrootte (per normgroep) toereikend?*

3.3 Worden gemiddelden, standaardafwijkingen en gegevens over de scoreverdeling vermeld?*

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?*

Standaarden

3.5 Is de standaardbepalingsmethode op een verantwoorde manier geselecteerd en is deze geschikt voor het te beoordelen assessment?

3.6 Is de standaardbepalingsprocedure voldoende omschreven en is deze zorgvuldig uitgevoerd?

3.7 Zijn er voor de validering van de standaard-bepalingsprocedure validiteitsbewijzen verstrekt?

Panel

3.8 Zijn de beoordelaars naar behoren getraind?

3.9 Is het panel voldoende groot en van voldoende kwaliteit?

3.10 Zijn de beoordelingen van de beoordelaars voldoende consistent?

Betrouwbaarheid

4.1 Zijn de betrouwbaarheidsgegevens berekend voor de steekproeven waarvoor het assessment gebruikt wordt?*

4.2 Zijn de betrouwbaarheidsgegevens correct berekend?

4.3 Is de betrouwbaarheidscoëfficiënt of generaliseerbaarheidscoëfficiënt correct berekend?

4.4 Is de (lokale) betrouwbaarheid correct berekend?

4.5 Is de consistentie of accuraatheid van classificaties correct berekend?

4.6 Zijn de resultaten voldoende gelet op het beoogde type beslissingen dat met behulp van het assessment wordt genomen?

Validiteit

Basisvraag

5.1 Is er een interpretatief argument beschikbaar?

Interpretatief Argument

5.2 Bevat het interpretatief argument de juiste gevolgtrekkingen?

5.3 Zijn de gevolgtrekkingen aannemelijk?

Validiteitsargument

5.3 Zijn de in het valideringsproces aangeleverde bewijzen voldoende?