Inzet van Assessment:

Waarom, wat, hoe, wanneer en door wie ?

Document

Identificatie	
U-nummer	
Status	Concept
Soort document	
Auteur(s)	YIV, JDA, LRU
Datum afdruk	10 mei 2001
Opgeslagen	

Goedkeuring

Acroniem	**Handtekening**	**Datum**

Wijzigingshistorie

Versie	**Acroniem**	**Datum**	**Wijziging**
0.1	YJV	26-01-01	diverse
0.2	YJV	10-05-01	diverse

Distributie

Versie	Datum	Naam
0.1	26-01-01	YJV, JDA, LRU
0.2	10-05-01	YJV, JDA, LRU

Onderwijstechnologisch expertisecentrum OTEC
Open Universiteit Nederland

Inzet van Assessment: Waarom, wat, hoe, wanneer en door wie?

Beslismodel voor een beargumenteerde keuze van assessmentvormen in onderwijs en opleiding.

OTEC 2001/13

Colofon

Titel:	Inzet van Assessment: Waarom, wat, hoe, wanneer en door wie? Beslismodel voor een beargumenteerde keuze van assessmentvormen in onderwijs en opleiding.
Auteurs:	Yvonne Vermetten, Jan Daniëls, Liesbeth Ruijs
Projectleiding:	Kathleen Schlusmans
Projectondersteuning:	Mieke Haemers
Uitgifte:	OTEC
Datum druk:	19 februari 2004

Onderwijstechnologisch expertisecentrum (OTEC)
Open Universiteit Nederland

Inzet van Assessment:
Waarom, wat,hoe, wanneer en door wie?

Beslismodel voor een beargumenteerde keuze van assessmentvormen in onderwijs en opleiding.

Inhoudsopgave

Inleiding op dit rapport

Dit rapport is één van de producten van programmalijn 1 in fase 4 van het Developmentprogramma. In fase 4 liepen er binnen deze programmalijn vier deelprojecten, waarvan het huidige gericht was op de ontwikkeling van **een beslismodel voor het inzetten van assessment**. De term assessment wordt hierbij gebruikt als verzamelterm voor allerlei vormen van toetsing.

De doelstelling van het Developmentprogramma (Projectplan, 1998) werd oorspronkelijk als volgt geformuleerd:
- het realiseren van een geïntegreerde elektronische leeromgeving, als instrumentarium voor studenten en docenten
- met het oog op het implementeren van een innovatieve onderwijsaanpak, met name, competentiegericht onderwijs (CGO).

CGO werd daarbij omschreven als een onderwijsaanpak die verschillende vormen aan kan nemen. In de eindrapportage 1ste fase (Koper, e.a., 1998) wordt niettemin een eigen didactische aanpak voor CGO naar voren geschoven, namelijk competentieleren dat verloopt via studietaken. In de 'Eindrapportage deelproject onderwijsaanpak - Didactische scenario's' (Manderveld, e.a., 1999), wordt als belangrijkste verschil tussen CGO en traditioneel onderwijs vermeld: de toetsing en de docent- en studentgerichtheid. Bij traditioneel onderwijs richt de toetsing zich hoofdzakelijk op het afzonderlijk toetsen van kennis en vaardigheden en is het onderwijs sterk docentgestuurd. In CGO geldt dat de toetsing geïntegreerd is in het leerproces (en vaak samenvalt met de uitvoering van de opdrachten) en dat onderwijs en toetsing meer studentgestuurd zijn (product- en procesbeoordeling en vormen van peer- en self-assessment). Ook de rol van toetsing is anders, niet louter beoordelend of certificerend, maar ook een didactische rol: monitoring van het leerproces. Anderzijds wordt ook vermeld dat er verschillende (klassieke) toetsvormen zijn die goed aansluiten bij competentiegericht leren, mits de nodige aanpassingen hierop plaatsvinden.

Het bovenstaande leidt tot allerlei vragen over de inzet van toetsing of assessment. Bijvoorbeeld, wanneer en waarom kiezen voor een meer klassieke vorm van toetsing en wanneer meer aandacht besteden aan product- of procesaspecten in de toetsing? Wat zijn de voor- en nadelen van het inzetten van peer- en self-assessment? Levert het gelijktijdig gebruiken van een toets als middel voor beoordeling en als middel voor het bijsturen van het leerproces geen verwarring op? In het huidige rapport en beslismodel wordt gepoogd op dergelijke vragen een antwoord gegeven.

De vragen waar het hier om gaat betreffen beslissingen die *vooraf* gaan aan het daadwerkelijk modelleren van onderwijs in EML. Het beslismodel dient ter ondersteuning van de ontwerper en inhoudsdeskundige bij het maken van een onderbouwde keuze voor een bepaalde toetsvorm bij een bepaald stuk onderwijs. In de literatuur is de sterk sturende werking van assessment op het leergedrag van studenten bekend. Hiernaar wordt vaak verwezen als het 'WYTIWYG-principe': What you test is what you get. Wat er getoetst wordt heeft een directe invloed op wat en hoe studenten leren. Kiezen voor een passende assessmentmethode is dan ook een essentiële vraag bij het ontwerpen van onderwijs.

In het huidige rapport is als eerste een overzicht opgenomen met definities en trends in assessment (paragraaf 1). Hierbij is gebruik gemaakt van zowel wetenschappelijke literatuur

als vakliteratuur. De algemene trends in assessment worden geïllustreerd met een aantal voorbeelden van praktische implementaties. Paragraaf 1 mondt uit in een indeling van toetsvormen (paragraaf 2). De keuze voor de indeling is gebaseerd op het principe dat deze functioneel moet zijn ten aanzien van het beslismodel. Daarna volgt een analyse van de onderscheiden assessmentvormen met behulp van twee 'instrumenten', te weten vijf basisvragen en acht kwaliteitscriteria (paragraaf 3 en 4). Deze analyse vormt het fundament voor het uiteindelijke beslismodel (paragraaf 5). Het beslismodel vormt het sluitstuk van dit rapport, en zou in een vervolgtraject kunnen worden uitgewerkt tot een geautomatiseerd programma.

Hoofdstuk 1: Definiëring, trends, en een indeling in assessmentvormen

In deze paragraaf worden hoofdlijnen uit de (internationale) literatuur met betrekking tot assessment beschreven. In paragraaf 1.2 wordt gekeken naar de onderwijsliteratuur, in paragraaf 1.3 naar literatuur vanuit de organisatiekunde. Omdat een aantal trends in 'educational assessment' voortkomen uit toepassingen in organisaties en bedrijven wordt de literatuur uit deze hoek ook belicht. Bovendien is er in deze literatuur veel aandacht voor 'competenties': één van de peilers van het Developmentprogramma. Maar eerst wordt begonnen met de definiëring van assessment.

1.1. Definiëring van de term assessment

Assessment kan worden gedefinieerd als het verzamelen en interpreteren van informatie over de prestaties van een student (Brookhart, 1999). In essentie bestaat assessment uit het nemen van een steekproef uit hetgeen studenten doen (informatie verzamelen), daar iets uit afleiden (interpreteren), en de waarde van hun acties proberen in te schatten (Brown, Bull & Pendlebury, 1997). De steekproef kan bijvoorbeeld bestaan uit het schrijven van een essay, het invullen van een multiplechoicetest, het oplossen van problemen en rapporteren van de oplossingen, het geven van een presentatie, etc. Op basis van de steekproef worden conclusies getrokken over de prestaties van de student of over zijn of haar potentieel, kennis, vaardigheden en attitudes. Vervolgens wordt meestal een schatting van de waarde gemaakt in de vorm van een cijfer, een niveauaanduiding of in de vorm van feedback en aanbevelingen.

De term assessment is afgeleid van het Latijnse *ad sedere*: zich ergens naast zetten, op gelijke hoogte gaan zitten of ondersteunen. De implicatie van de etymologie is dat assessment primair te maken heeft met het geven van begeleiding en feedback aan de lerende. Dit is in essentie ook nu nog steeds de belangrijkste functie van assessment: het verder brengen van de lerende.

In de Angelsaksische literatuur worden de termen 'assessment' en 'evaluation' in verschillende betekenissen gebruikt. Zo verwijst de term 'assessment' in de Engelse literatuur naar het toetsen en beoordelen van de prestaties van de student, terwijl de term 'evaluation' voornamelijk gebruikt wordt voor het evalueren van de kwaliteit van onderwijs en onderwijsomgeving (kwaliteitszorg). In de Amerikaanse literatuur wordt met 'assessment' verwezen naar 'de activiteit van het meten' en met 'evaluation' naar het waarderen, het scoren van de prestaties van de student. Voor het evalueren van de kwaliteit van het instructieproces gebruikt men hier vaak de term 'accountability'.

In dit rapport wordt de term assessment in de Engelse betekenis gehanteerd, en omvat aldus het gehele proces van meten en beoordelen; de verschillende procedures of methodes die gebruikt worden voor het vaststellen en beoordelen van prestaties of eigenschappen van de lerende voor wat voor doel ook (Brown, Bull & Pendlebury, 1997). Er zijn ook auteurs (bijvoorbeeld Dochy & Segers, 1999) die de term 'assessment' uitsluitend reserveren voor niet-klassieke vormen van toetsing, ofwel voor 'alternatieve' en 'nieuwe' vormen van toetsing. In dit rapport wordt de term assessment niet in die exclusieve betekenis gebruikt, maar als synoniem voor toetsing.

1.2. Onderwijsliteratuur: Trends in 'educational assessment'

Klassieke toetsing
Traditioneel wordt in het onderwijs voornamelijk gebruik gemaakt van schriftelijke en mondelinge toetsen, die aan het eind van een vak of cursus worden afgenomen, en die gericht zijn op het meten van kennis. Deze 'klassieke' manier van toetsen kan bestaan uit gesloten vragen (bijvoorbeeld meerkeuzevragen), of uit open vragen. In het eerste geval wordt wel gesproken van 'objectieve scoring', omdat de scoring van antwoorden op volstrekt objectieve wijze kan gebeuren, en desgewenst kan worden geautomatiseerd. Hambleton (1996) spreekt in dit geval van 'selected response items'. Bij open vragen wordt gesproken van 'subjectieve scoring', omdat de antwoorden een open vorm hebben (studenten moeten deze zelf construeren), en daarom stuk voor stuk geëvalueerd moeten worden op hun kwaliteit. Hambleton (1996) verwijst naar deze categorie met de term 'constructed response items'. De termen 'objectieve scoring' en 'subjectieve scoring' houden geen waardeoordeel in.

Een voorbeeld uit de praktijk wat deze meer 'klassieke' toetsvormen illustreert is het tentamensysteem **SYS**, dat bij de Open Universiteit Nederland wordt gebruikt. SYS is een geautomatiseerd tentamensysteem waarbij met zowel gesloten (multiplechoice- of true/false items) als met open vragen wordt gewerkt. Het is bedoeld voor het geven van eindbeoordelingen van individuele studenten. Het systeem bevat itembanken en procedures voor het genereren, printen, scoren en analyseren van toetsen (Moerkerke, 1996). Een itembank van een module bevat een verzameling items (gemiddeld ongeveer 500), het antwoordmodel en het tentamenprofiel.

Aan SYS-tentamens kan worden deelgenomen op momenten die de student zelf goed uitkomen. Een student die een tentamen wil maken, schrijft zich in bij een studiecentrum van de Open Universiteit Nederland. Voor de kandidaat wordt ter plekke een tentamen gegenereerd uit een itembank. Na afronding van het tentamen, worden de antwoorden op de MC-vragen door de computer gescoord en krijgt de student een voorlopige uitslag. De open vragen worden afzonderlijk door verschillende beoordelaars aan de hand van modelantwoorden en bijbehorende criteria gescoord. Na enkele controles door het systeem, ontvangt de student de definitieve score en indien de student geslaagd is, een certificaat.

Nadelen van klassieke toetsing
In de jaren tachtig werd in de Verenigde Staten, in het kader van een streven naar een grotere objectiviteit, nagenoeg exclusief gebruik gemaakt van MC-vragen, meestal ontwikkeld door externe toetsdeskundigen. Dit leidde ertoe dat docenten en studenten zich in hun onderwijs en leeractiviteiten sterk gingen richten op de inhoud van deze toetsen, waarbij aanvankelijk feitenkennis overheerste (zie hier een demonstratie van het WYTIWYG-principe). Bij internationaal vergelijkend onderzoek bleken de resultaten van Amerikaanse studenten, vergeleken met de omringende landen, echter beneden het gemiddelde. Dergelijke bevindingen resulteerden in een groots opgezet plan tot onderwijsvernieuwing: 'Goals 2000. Educate America Act'. Er werden nieuwe onderwijsdoelen centraal gesteld, zoals '..to demonstrate competency of higher level cognitive skills'. Men hoopte deze vernieuwde onderwijsdoelstellingen te realiseren, niet alleen door het onderwijsproces te veranderen maar tevens door een forse veranderingen in de praktijk van assessment. Meerdere onderwijsdeskundigen pleitten dan ook voor een verschuiving van het toetsen van vakkennis naar het toetsen van de vaardigheid van studenten, om te redeneren, kritisch te denken en problemen op te lossen (Hambleton, 1996).

Verschuivingen in onderwijs en assessment

De bovenstaande ontwikkelingen sluiten aan bij de algemene roep in de onderwijsliteratuur om onderwijsvernieuwing met nadruk op 'deep level learning': inzichtelijk leren en het verwerven van cognitieve vaardigheden en competenties. Onderwijsdoelstellingen verschuiven hiermee van het verwerven van veel (feiten)kennis naar het verwerven van vaardigheden en competenties, met name de vaardigheid om zelf verder te kunnen leren en kennis te verwerven (ook wel 'leren leren', of 'levenslang leren').

Om deze verschuiving te verklaren wordt vaak verwezen naar een aantal maatschappelijke veranderingen, waaronder de snel evoluerende kennismaatschappij (geleerde kennis veroudert snel) en de algemene intrede van ICT in onze maatschappij, waardoor ook ICT-vaardigheden tot de kwalificaties van afgestudeerden zijn gaan behoren. Anderzijds komt de behoefte aan andere onderwijsdoelstellingen voort uit gewijzigde inzichten in de leerpsychologie ten aanzien van het verloop van het leerproces. De constructivistische stroming omschrijft daarbij het leerproces als een actief en constructief proces van de lerende, dat wordt bevorderd door het aanbieden van authentieke (realistische) problemen, en het bevorderen van samenwerkend leren, feedback en zelfreflectie.

De gewijzigde inzichten vormen een basis om naast de klassieke toetsbenadering, ook aandacht te besteden aan alternatieve assessmentvormen als een meer valide vorm van toetsing. Het assessmentproces dient meer in overeenstemming te worden gebracht met de (nieuw) nagestreefde leerdoelen, het leerproces en de leeractiviteiten van de student. Biggs (1996) spreekt in dit kader over 'constructive alignment': het laten aansluiten van assessment en instructiemethoden op de nieuwe doelstellingen zoals hogere-orde vaardigheden en competenties. Met andere woorden, nieuwe vormen van assessment zijn nodig '.. to fall into line with new curriculum specifications' (Hambleton, 1996). Dit vraagt om vormen van authentieke assessment: toetsen die sterk lijken op de complexe taken, problemen en opdrachten die ook in de realiteit van het vakgebied een rol spelen.

Maatschappelijke veranderingen en gewijzigde inzichten over het leerproces (zoals het constructivisme) zorgen er tevens voor dat de rol van de student in het onderwijs verandert van een passieve naar een actieve rol (Biggs, 1996). Dit geldt tevens voor de rol van de student in het *toets*proces. In plaats van het passief ondergaan van een weinig doorzichtig examen wordt van de student actieve participatie in het assessmentproces verwacht. Dit geldt zowel voor de inhoud en de vorm van de toets, als voor het opstellen van de criteria en beoordelingsregels van de resultaten. Concreet vindt dit plaats in vormen van self-, peer- en co-assessment. Met deze assessmentvormen wordt bovendien een specifieke onderwijsdoelstelling nagestreefd, namelijk het ontwikkelen van zelfreflectie op het eigen leerproces en de resultaten (Brown & Glasner, 1999; Dochy, 1999).

Aansluitend op de idee dat studenten hun eigen leerproces sturen en dat 'deep level learning' moet worden bevorderd, wordt assessment niet langer gezien als enkel de afsluiting van het onderwijs- en leerproces. Assessment is een integraal onderdeel geworden van het instructieproces. De nadruk komt steeds vaker te liggen op de sturende en ondersteunende functie van assessment in het leerproces (ook wel: formatieve toetsing). In dit kader wordt ook wel de term 'dynamic assessment' gehanteerd, waarmee men de nadruk wil leggen op het dynamisch sturend karakter van de resultaten van toetsing, zowel voor het instructieproces van de docent als het leerproces van de student.

Als er sprake is van een sturende rol van assessment in het leerproces is 'criterion-referenced' assessment de aangewezen manier van beoordelen. Bij 'criterion-referenced' assessment worden individuele prestaties gewaardeerd op basis van een op voorhand

vastgelegd niveau van presteren (bijvoorbeeld een aantal duidelijk omschreven criteria waaraan voldaan moet zijn). Het relateren van eigen prestaties aan heldere criteria is immers een erg leerzaam proces. Bij traditionele toetsen lag de nadruk vooral op 'norm-referenced' assessment waarbij de prestaties van de individuele student bij het scoren voornamelijk vergeleken worden met de prestaties van zijn medestudenten. De groepsprestaties en het groepsgemiddelde vormen aldus de referentie voor het beoordelen van een individuele prestatie.

De rol van ICT in assessment

ICT vervult steeds meer verschillende rollen binnen het onderwijsleerproces. Zo zal het ook een steeds belangrijkere rol gaan spelen binnen assessment. In het verleden werd vooral beroep gedaan op ICT voor het beheer van toetsen en voor het scoren en analyseren van toetsresultaten. Hierbij ging het meestal om itembanking (geautomatiseerde opslag van items), en itemanalyse (berekening van de klassieke itemparameters, moeilijkheidsgraad, discriminatie-index en itembetrouwbaarheid). Dergelijke geautomatiseerde toetsservicesystemen zijn hoofdzakelijk gericht op klassieke toetsen en het gebruik van klassieke psychometrische modellen.

De ontwikkeling van de PC tot een modern krachtig multimediaal genetwerkt werkstation biedt vooral mogelijkheden in het kader van ontwikkeling van moderne assessmentmethoden. Zo kunnen door het inschakelen van multimedia realistische stimulussituaties worden gecreëerd. Krachtige computers bieden ook de mogelijkheid tot het ontwikkelen van dynamische simulaties, tot het systematisch en stapsgewijs toetsen van probleemoplossende vaardigheden, en tot het sequentieel en adaptief toetsen in functie van het ontwikkelingsverloop van de lerende. Nauw hiermee verbonden is het verstrekken van gedifferentieerde feedback of query-gebaseerde feedback in functie van de specifieke informatiebehoefte van de lerende of zijn uitdrukkelijke vraag.

Een ander toepassingsdomein waarbij de inschakeling van ICT een nuttige rol kan vervullen is de opbouw van een elektronische portfolio. Door de uitbouw van persoonlijke dossiers op PC en het regelen van de toegang via het net, kunnen problemen zoals het maken van meerdere kopieën voor beoordeling door verschillende beoordelaars gemakkelijk worden omzeild. Dankzij het gebruik van netwerken kan bovendien de toetsafname en het geven van feedback nagenoeg tijd- en plaatsonafhankelijk gebeuren. Naast enthousiasme over nieuwe mogelijkheden is een waarschuwing op zijn plaats voor overdreven verwachtingen. Het beoordelen van authentieke producten en processen blijft voorlopig nog mensenwerk.

Drie vormen van alternatieve toetsing

De verschuivingen in onderwijs en assessment komen tot uiting in drie alternatieve assessment-vormen die steeds meer gebruikt worden in onderwijssettings: performance assessment, portfolio-assessment, en self-, peer-, en co-assessment. Hieronder wordt toegelicht wat onder ieder van deze vormen wordt verstaan. Ook wordt bij iedere vorm een recent praktijkvoorbeeld gegeven als illustratie van hoe alternatieve assessment in het onderwijs geïmplementeerd wordt.

1.2.1. Performance assessment

Performance assessment wordt gekarakteriseerd door termen als 'authentic' en 'directness' (Hambleton, 1996; Messick, 1994). De term 'authentic' duidt op het aansluiten van de opdrachten bij de leeractiviteiten of concrete opdrachten en situaties uit het toekomstige beroepenveld. De taak heeft dus zoveel mogelijk een levensechte vorm. Dit betekent dat de assessmenttaak gegrond is in het soort werk dat de mensen in dat vakgebied of discipline feitelijk doen. De term 'direct' wil zeggen dat de performance assessment beoogt om het

leerdoel ofwel criterium (bijvoorbeeld een bepaalde competentie) rechtstreeks te meten. Klassieke toetsen zoals mc-vragen meten eerder indirecte indicatoren van de beoogde competenties.

In een ideale performance assessment werken studenten aan complexe en relevante taken, waarbij ze hun voorkennis en hun recent verworven kennis en vaardigheden gebruiken voor het oplossen van realistische complexe problemen, en het creëren van producten met een substantiële omvang. Performance assessment omvat een grote verscheidenheid aan open toetsvormen zoals projecten, casustoetsen, OSCE's (objective structured clinical examinations), poster presentaties, essays, papers, etc.

Het product, en vaak ook het proces van de student wordt beoordeeld op kwaliteit. De beoordeling vindt plaats aan de hand van een scoringsschema of checklist. De feitelijke beoordeling kan de bevoegdheid zijn van een docent, maar ook van studenten en/of externe experts.

De **Overall toets** (Dochy & Segers, 1999) kan worden gezien als een illustratie van performance assessment in een academische context. Hij wordt gebruikt aan de faculteit der Economische Wetenschappen en Bedrijfskunde van de Universiteit Maastricht, waar het onderwijs is opgezet volgens het probleemgestuurd model. Om de manier van toetsen beter te laten aansluiten bij de doelstellingen van dit model en bij de eindtermen van de opleiding, is in 1990 de Overall toets (OAT) ingevoerd.

De OAT toetst in welke mate studenten problemen kunnen analyseren, en kunnen bijdragen aan de oplossing van problemen door toepassing van concepten, modellen en theorieën uit het betreffende vakgebied (synthese). Tevens wordt getoetst of men mogelijke oplossingen of te nemen beslissingen kan beargumenteren en evalueren.

Om deze doelen te toetsen kent de OAT een aantal specifieke kenmerken zowel wat betreft toetsvorm als wat betreft de organisatievorm. De OAT is een afsluitende toets die telkens na twee blokken is ingepland. Tijdens een zelfstudieperiode van twee weken bestuderen de studenten een set van artikelen. De artikelen beschrijven een probleemsituatie in zijn totaliteit en vanuit verschillende disciplines. Niet alleen de essentiële informatie die nodig is voor het oplossen van het probleem wordt in een korte casus behandeld, maar ook contextgegevens en afhankelijke variabelen worden opgenomen in uitgebreide casussen om te bewerkstelligen dat het probleem op de juiste wijze wordt geïnterpreteerd. Van de studenten wordt verwacht dat zij zelfstandig de stappen doorlopen die zij aanleren bij de analyse van problemen in de onderwijsgroepen (namelijk., definiëren, analyseren, oplossen, evalueren).

De toets heeft een open-boek-karakter en de toetsvragen richten zich op kernaspecten van de probleemsituatie. Juist/onjuist vragen toetsen of studenten het probleem goed definiëren via de herkenning van begrippen, modellen en theorieën in de nieuwe situatie. De overige stappen worden gemeten met open vragen. Studenten mogen alle materialen meebrengen die in een reële beroepssituatie ook voorhanden zouden zijn. De beperkte toetstijd voorkomt dat studenten gaan ´studeren´ in het meegebrachte materiaal.

De beoordeling vindt plaats aan de hand van antwoordmodellen. Dergelijke modellen bestaan onder andere uit een omschrijving van goede en mogelijk foutieve antwoorden, een toelichting op de antwoorden, en beoordelingscriteria wanneer voor een vraag meerdere goede antwoorden gelden. Het relatieve gewicht van de vragen is afhankelijk van een aantal criteria, zoals het type vraag of de hoeveelheid denkstappen die de student dient te zetten.

In dit praktijkvoorbeeld komt het genoemde belang van authenticiteit in toetsing duidelijk naar voren. Er wordt gezorgd voor complexe, realistische probleemsituaties waarvoor studenten een beroep moeten doen op hun verworven kennis én op probleem-oplossingsvaardigheden. De beoordeling is ook op beide aspecten gericht. Het eerder genoemde idee van 'constructive alignment' komt naar voren in het bovenstaande voorbeeld.

1.2.2. Portfolio assessment

Een ander steeds vaker gebruikt instrument in het kader van alternatieve assessment is de portfolio. Oorspronkelijk afkomstig uit het kunstonderwijs, is deze vorm van assessment erg populair geworden in de lerarenopleidingen. Een portfolio is een door de student in de loop van de tijd zorgvuldig samengestelde verzameling van bewijsmateriaal die (steekproefsgewijs) aantoont wat hij of zij geleerd heeft (Brown e.a., 1997). Een portfolio kan bestaan uit producten, werkstukken, beoordelingsrapporten van docenten (en/of medestudenten), video-opnames, e.d.. Het portfolio wordt meestal vergezeld door reflectieverslagen over het eigen leerproces en het huidige ontwikkelingsniveau, en door een uitleg van de student over wat hij of zij wil demonstreren met de verschillende onderdelen van het portfolio. Het doel van portfolio assessment is meestal het opvolgen van het verloop van het leerproces, het aantonen van wat de student in de afgelopen periode gepresteerd en geleerd heeft en waar hij nu staat. Het is dan ook vooral een geschikt instrument in het kader van formatieve toetsing, waarbij de ontwikkeling van de student ondersteund wordt. Er is daarbij vaak sprake van self-assessment, waarbij de feedback van een docent alleen een toetssteen is voor de student zelf.

Om de functie van 'tool for learning' te vervullen is het van belang dat er sprake is van 'criterion-referenced' assessment met duidelijke criteria en standaarden vanuit een vooraf vastgesteld prestatieniveau. In de mate dat men bij het verwerven van dit prestatieniveau uitgaat van het beginniveau van de student en rekening houdt met het individueel leerverloop en tempo spreekt men zelfs van 'ipsative-referenced' of 'self-referenced' assessment. Naast duidelijke criteria is het verstrekken van informatieve feedback een essentieel onderdeel van dit instrument. Bij de keuze van de items waaruit het portfolio is samengesteld is de student vrij of kunnen op voorhand door de docent/begeleider bepaalde aanwijzingen worden gegeven over inhoud en vormgeving. Hetzelfde geldt voor de beoordelingscriteria, waarin de student ook een rol kan spelen.

Gebruik in het kader van eindbeoordeling is niet uitgesloten. Gelijktijdig gebruik voor individuele ontwikkeling en voor beoordeling kan echter het verloop van een open discussie tussen docent/begeleider en student ernstig belemmeren. Bij gebruik voor beoordeling moeten verwachtingen over de samenstelling van het portfolio vooraf expliciet worden gemaakt.

Er zijn verschillende soorten portfolio's met verschillende kenmerken in functie van de beoogde doelstelling of toepassingscontext. Zo worden bijvoorbeeld volgende vormen onderscheiden (Dochy, 1999):
- een exemplarische of productportfolio met alleen de meest overtuigende objecten of items uit een bepaalde ontwikkelingsperiode, en voornamelijk bedoeld voor de buitenwereld (gebruikelijk in de kunst);
- een procesportfolio met elementen die vooral de ontwikkeling van de student op een bepaald domein, over een bepaalde periode illustreren (hier is sprake van een 'tool for learning');
- een gecombineerde portfolio die beide vormen van assessment, product- en procesevaluatie, mogelijk maakt;

- een moederportfolio die alle mogelijke elementen bevat en die voor de student als basis fungeert voor het samenstellen van een van de hoger vermelde portfolio's.

Als praktijkvoorbeeld is de introductie van een **portfolio voor docenten-in-opleiding** (dio´s) bij de postdoctorale lerarenopleiding van het ICLON (Interfacultair Centrum voor Lerarenopleiding, Onderwijsontwikkeling en Nascholing) van de Universiteit Leiden gekozen (Beijaard, Longayroux en Tanner, 1997). Uit de portfolio van dio's moet duidelijk blijken dat deze zich bewust zijn van hun eigen leerproces en ontwikkeling als docent door daarop te reflecteren en deze reflecties in het portfolio te beschrijven. Deze beschrijvingen worden aangevuld met ´bewijsmaterialen´. Het portfolio handelt over meerdere onderwerpen of thema's (bij voorkeur niet minder dan vijf en niet meer dan tien thema's). Centraal bij de beschrijving van een thema staat de reflectie van de dio op de keuze van het thema, de inhoud ervan en op welk aspect van diens ontwikkeling het betrekking heeft. Het portfolio wordt gebruikt als een instrument voor professionele ontwikkeling.

Het portfoliotraject bestaat uit twee gedeelten. In het eerste deel leert de dio onder begeleiding een portfolio samenstellen. Ook wordt overleg met mede-dio's gestimuleerd. Studenten krijgen opdrachten rondom belangrijke portfoliobegrippen (Wat is bewijsmateriaal? Hoe selecteer je het?) en ontvangen een handleiding met onder andere een definitie van portfolio, een voorbeeld van een inhoudsopgave en een lijst met te gebruiken ´bewijsmaterialen´ (bijvoorbeeld een video-opname van een les, een bij leerlingen afgenomen toets, voorbeelden van werk van leerlingen, lesevaluaties etc.). Het eindgesprek van dit traject resulteert o.a. in leerdoelen die de basis vormen voor het tweede portfoliotraject.

In het tweede gedeelte gaat de student meer zelfstandig te werk. Er wordt verwacht dat de dio het portfolio zodanig samenstelt dat het een afspiegeling is van de volle breedte van het leraarsberoep. Studenten ontvangen een lijst met relevante aspecten van het beroep. De student legt zelf accenten in het portfolio, de exacte inhoud ervan wordt bepaald door zijn of haar persoonlijke leerdoelen. Ook hier wordt overleg met anderen aangemoedigd. Dit traject wordt eveneens afgerond met een eindgesprek waarin het portfolio wordt besproken. Aanwezig hierbij zijn de desbetreffende instituutsdocent, de stagebegeleider en de dio.

In het beschreven voorbeeld zie je dus een aantal algemene punten die werden beschreven voor portfolio's duidelijk terugkomen. Er wordt een accent gelegd op persoonlijke ontwikkeling en reflectie hierop, er is een grote mate van individualiteit bij het gebruik van het portfolio ('self-referenced') en de eigen inbreng en verantwoordelijkheid van de student zijn groot.

1.2.3. Self-, peer- en co-assessment

Klassiek behoort in het onderwijs de toetsing en beoordeling tot de prerogatieven van de docent. Hij of zij bepaalt de vorm, procedure, criteria en normen van het afsluitend tentamen. Dit is echter niet meer zo vanzelfsprekend in het kader van onderwijsvernieuwing waarin meer dan ooit de nadruk wordt gelegd op het verwerven van complexe vaardigheden en waarbij zelfreflectie, zelfverantwoordelijkheid en het kunnen beoordelen van de prestaties van collega-studenten een wezenlijk onderdeel van deze competenties vormen. In deze context zijn vormen van self-, peer- en co-assessment, waarbij de student inspraak verwerft in vorm en wijze van beoordeling, aangewezen.

In essentie gaat het bij self-, peer- en co-assessment om de actieve inbreng van de student in het toetsproces. Door studenten hierin te betrekken wordt het inzicht in de verwachte prestaties en de eigen verantwoordelijkheid voor het leerproces verhoogd. Bovendien vormt

het deelnemen in het assessmentproces op zich een belangrijke leerervaring die bijdraagt aan het verwerven van zelfreflectievaardigheden en het beoordelen van de prestatie van anderen. De actieve betrokkenheid kan plaatsvinden in het kader van allerlei assessmentvormen, maar ziet men vooral terug bij alternatieve vormen zoals portfolio's. De nadruk ligt op formatieve toetsing en zelfreflectie, eerder dan op eindbeoordeling.

Bij self-assessment toetsen de studenten zichzelf. Dit kan gebeuren bij aanvang van een leereenheid, gedurende of bij het afsluiten. Bij peer-assessment beoordelen studenten elkaar. Het hoofddoel is elkaar feedback te geven en zo het leerproces te bevorderen. Peer-assessment komt veel voor in onderwijsleergroepen. Bij co-assessment beslissen docent en student gezamenlijk over de wijze van beoordelen en de geldende beoordelingscriteria. Vervolgens beoordelen studenten de eigen leerresultaten (of die van medestudenten). Het uiteindelijke oordeel wordt door de docent vastgesteld. Ter ondersteuning van de beoordelingsvaardigheid van de student is het aanbevolen om dit extra te ondersteunen door een aangepaste trainingsperiode.

Een voorbeeld waarin self- peer- en co-assessment een rol spelen is de toetsing bij de implementatie van het **'virtueel bedrijf' bij TAS-opleidingen** (Joosten & Boon, 1999). Ze worden gecombineerd met onder andere performance assessment en een portfolio.

Het door de Open Universiteit Nederland ontwikkelde onderwijsconcept Virtueel Bedrijf (VB) heeft als doel de competenties van studenten of werknemers verder te ontwikkelen. Zij krijgen hiervoor faciliteiten tot hun beschikking als 'video- en audio-conferencing, groupware, application sharing systems, personal information manager' etc… Met deze instrumenten gaan de studenten aan de slag als werknemers in een genetwerkt bedrijf, ofwel een virtueel bedrijf. In dit bedrijf werken zij aan authentieke opdrachten en aan de ontwikkeling van hun competenties.

Het VB is onder andere geïmplementeerd bij de opleiding tot basisontwerper (BO-er) bij TAS-opleidingen. Tijdens deze opleiding moeten de studenten in teamverband een (grote) opdracht uitvoeren, namelijk het maken van een basisontwerp. Hierbij worden zij beoordeeld op verschillend soorten vaardigheden (persoonlijke, vaktechnische, communicatieve en coördinerende vaardigheden).

Een belangrijk kenmerk van de toetsing is dat deze geïntegreerd is in het leerproces. Er wordt op veel verschillende manieren getoetst, door verschillende beoordelaars, met verschillende doelen en op verschillende momenten. Meer specifiek zijn de meetmomenten gekoppeld aan de zeven taken die plaatsvinden binnen de opdracht. Iedere taak levert een product op (bijvoorbeeld een plan van aanpak of een functioneel basisontwerp) en een procesverslag (bijvoorbeeld een gespreksverslag). Alle 'output' per opdracht wordt opgeslagen in de portfolio's van de individuele cursisten. In de beoordelingen wordt onder meer gelet op de groei die iemand doormaakt in de loop van de tijd.

Na iedere taak vindt assessment plaats door verschillende boordelaars, te weten:
- assessment door de gesprekspartners
- expert-assessment (vakinhoudelijk iemand, bijvoorbeeld een senior IT-er)
- peer-assessment (medecursisten beoordelen elkaar)
- self-assessment of reflectie (cursist beoordeelt zichzelf)
- coach assessment (ieder team heeft een coach, bijvoorbeeld een HRM-figuur, maar wel met inhoudelijke deskundigheid).

De verschillende beoordelaars letten op verschillende vaardigheden. Ze vullen allemaal een voor hen bestemde beoordelingslijst in. Bij de self-assessment wordt hieraan bovendien een

reflectie op eigen kunnen (o.a. een sterkte/zwakte analyse en gewenste verbeteringen) toegevoegd. De expert geeft, naast de beoordelingslijst, ook feedback aan het team over de toegekende scores. Bij de peer-assessment beoordeelt ieder teamlid de andere teamleden op een aantal vaardigheden. De coach kijkt na iedere opdracht of er voldoende vooruitgang is bij team en teamleden en of er ingegrepen moet worden. Hiertoe maakt hij een scoringsoverzicht op door het samenbrengen van de informatie uit het portfolio en de beoordelingslijsten. Hij geeft hierbij beoordelingen over de mate van beheersing van de vaardigheden. Daarnaast vult de coach een beoordelingslijst in over de kwaliteit van het portfolio (o.a. van de reflectie). Ten derde stelt de coach vast of ingrijpen nodig is. Alle beoordelingslijsten worden opgeslagen in de portfolio's.

Na de laatste opdracht geeft de coach een eindoordeel. Het eindoordeel komt tot stand op basis van:
1. Eindscores: Over alle opdrachten heen worden eindscores (niveauscores) berekend voor alle vaardigheden. Bovendien wordt een 'leereffectscore' berekend, die gebaseerd is op de getoonde groei tussen opdracht 0 en de laatste opdracht, en op de reflectie. Op basis van het eindniveau én het leereffect wordt een voorlopig eindoordeel opgemaakt (op basis van beslisregels).
2. Beoordelingsgesprek: Er vindt een beoordelingsgesprek plaats tussen coach, cursist en een verantwoordelijke van het bedrijf. Het uiteindelijke oordeel wordt opgesteld en verwerkt door deze laatste.

Het bovenstaande voorbeeld omvat veel elementen van genoemde trends in assessment. Het is daarmee een goede illustratie van de beweging die gaande is. Opvallend is dat het voorbeeld plaats vindt in een bedrijfsmatige context. In het algemeen geldt dat de verschuivingen in de assessment cultuur in het onderwijs vaak zijn geïnspireerd door de ontwikkelingen op dit vlak in bedrijven en organisaties. Hierop wordt in de volgende paragraaf ingegaan.

1.3. Organisatieliteratuur: Trends in 'organizational assessment'

Hoewel de term assessment in het onderwijs een klank van vernieuwing heeft, is het een gangbare term in de organisatiekunde en industriële psychologie. Hier kennen de 'alternatieve' vormen van assessment al vaak een langere traditie en is er ervaring mee opgedaan. Performance assessment en assessment centers zijn hier vrij gangbaar (denk bijvoorbeeld aan selectie-assessment in militaire organisaties). Voor organisaties is de term 'alternatief' dus niet meer van toepassing.

De trend bij assessment in organisaties is om steeds minder te werken met duidelijk omschreven functie-eisen die beoordeeld worden in functie van selectie of beloning. In plaats hiervan worden flexibele en persoonsgebonden criteria ontwikkeld en staat de ontwikkeling en het groeitraject van personen centraal bij het beoordelen (Tillema, 1996). Hierbij nemen werknemers ook zelf de verantwoordelijkheid voor het behalen van bepaalde competentieniveaus. In deze trend komt de 'locus of control' steeds meer bij het individu te liggen zodat er sprake is van meer zelfsturing bij de eigen ontwikkeling.
De nadruk op competentieontwikkeling is verklaarbaar vanuit maatschappelijke ontwikkelingen. In de jaren negentig is de maatschappij turbulenter en meer geïndividualiseerd geworden. Organisaties zijn alleen nog concurrerend als ze flexibel en snel kunnen reageren op veranderingen. Hierdoor hebben ze medewerkers nodig die beschikken over een grote mate van zelfstandigheid en flexibiliteit. De zorg voor medewerkers verschuift dan ook van zorg voor zekerheid en salaris, naar zorg voor ontwikkeling en welbevinden.

De la Parra, Slotman, Tillema en Spannenburg (2000) benoemen ditzelfde verschijnsel, in het kader van competentiemanagement in organisaties. Op dit terrein wordt het zogenaamde beheersingsmodel verlaten voor het faciliteermodel. Een beheersingsmodel verwijst naar een managementvisie die is te typeren met termen als sturen, controleren, beheersen en instrumenteren. Bij het faciliteermodel passen de termen faciliteren, laten ontstaan, uitgaan van en mobiliseren. Het eerste model is meestal zeer systematisch, kent duidelijke procedures en is geïnstrumentaliseerd. Het onderliggende idee hierbij is dat de organisatie controle heeft over doelen, proces en instrumenten van competentiemanagement. Volgens Kessels (2000) werkt het echter niet helemaal zo, en zullen individuen hun competenties het best ontwikkelen in netwerken waarin initiatieven kunnen ontstaan die de organisatiestrategie mede kunnen gaan bepalen. De term 'lerende organisatie' past ook bij dit laatste model.

In de volgende twee paragrafen wordt dieper ingegaan op twee veelgebruikte assessmentvormen in organisaties, namelijk 360-graden feedback en assessment en development centers.

1.3.1. 360-Graden feedback

Bij 360-graden feedback krijgt een werknemer feedback van verschillende andere personen uit de werkomgeving, zoals de leidinggevende, een collega, klanten, medewerkers e.d. (Jellema, 2000). Daarom wordt het ook wel *multi-rater* feedback genoemd. Daarnaast beoordeelt de werknemer ook het eigen werkgedrag. In feite is er dus sprake van self- peer-, en co-assessment. Door het complete beeld dat ontstaat, is het aannemelijk dat deze vorm van assessment betrouwbaarder is dan een beoordeling door één persoon (traditioneel de leidinggevende).

De populariteit van 360-graden feedback is voortgekomen uit ontwikkelingen op de werkvloer zoals het vaker voorkomen van teamwork, het ontstaan van platte organisaties en snelle technologische vernieuwingen. Door dergelijke ontwikkelingen heeft de leidinggevende vaak een minder goed zicht op het functioneren van werknemers dan dat zij dat van zichzelf en van elkaar hebben.

Meestal wordt in de procedure een vragenlijst gebruikt die bestaat uit een aantal gedragscategorieën of competenties die weer opgebouwd zijn uit een aantal items. Verder gaat het meestal om gedragsbeschrijvingen en minder om persoonlijkheidskenmerken of objectieve gegevens zoals omzet en winst. Gedragsbeschrijvingen blijken de meest betrouwbare informatie op te leveren en zijn bovendien gemakkelijker te ontwikkelen (Jellema, 2000). 360-Graden feedback wordt meestal gebruikt als *ontwikkelings*instrument. De feedback van verschillende personen uit de werkomgeving dient dan ter stimulering van de ontwikkeling van de werknemer.

Naast 360-graden feedback is ook de term 90-graden feedback geïntroduceerd (Lipo-site). Hierbij gaat het om de terugkoppeling van de medewerker naar de chef over zijn of haar functioneren als leidinggevende. Bij 90-graden feedback wordt het oordeel van de medewerker over de ontvangen leiding en sturing bewust in het assessment (bijvoorbeeld in het functioneringsgesprek) betrokken. Door dit aspect te betrekken in assessment wordt communicatie tussen medewerker en leidinggevende bevorderd. Het maakt duidelijk hoe leidinggevende en medewerker tegenover elkaar staan en wat zij van elkaar begrijpen.

1.3.2. Assessment centers en Development centers

Een assessment center (AC) is een methode waarbij door praktijksimulaties relevant gedrag wordt opgeroepen dat vervolgens wordt beoordeeld door meerdere getrainde observatoren. Grondgedachte is dat gedrag wat tijdens de simulaties wordt opgeroepen een steekproef is uit gedrag dat voor de latere functie relevant is. Op deze manier hoopt men het gedrag van toekomstige medewerkers in de praktijk te voorspellen. Het AC wordt overigens niet alleen gebruikt voor selectiedoeleinden, maar steeds vaker ook voor het vaststellen van doorgroeimogelijkheden voor personeel (Kuiken, 2000). In dit laatste geval wordt gesproken van een Development Center (DC). Development centers verbinden assessment met ontwikkeling en training met de bedoeling om verandering van gedrag in een gewenste richting aan te brengen (Tillema, 1996).

In een assessment en development-center (ADC) wordt aan de hand van simulatie-opdrachten de medewerker getest op verschillende gedragsdimensies. Deze opdrachten zijn gebaseerd op de inhoud van de toekomstige functie en bootsen deze zo goed mogelijk na. Hierdoor kan systematisch informatie worden verzameld over iemands sterke en zwakke punten. Het gedrag wordt beoordeeld aan de hand van tevoren geformuleerde criteria. Deze criteria zijn gebaseerd op een functie-analyse. Bij Development Centers wordt geen functie-analyse maar een competentie-analyse gebruikt. Verder spreekt men in een DC niet van beoordelingsinstrumenten, maar van diagnose- of ontwikkelingsinstrumenten. De beoordelaars kunnen getrainde managers uit de eigen organisatie zijn, maar soms wordt een AC of DC ook uitbesteed aan een gespecialiseerd bureau.

Het soort praktijkopdrachten dat in een ADC gebruikt wordt, kent een grote verscheidenheid, maar er zijn toch enkele typische soorten opdrachten aan te geven. Kuiken (2000) vat deze als volgt samen:
- In-basket opdrachten (‘postvakje’). Dit is de bekendste praktijkopdracht binnen AC's. In een in-basket opdracht wordt het postvakje van een medewerker nagebootst. In het bakje liggen bijvoorbeeld memo's, beleidsnota's, klachtbrieven of bestellingen. Binnen een afgesproken tijdslimiet moet structuur in de chaos worden aangebracht en actie worden ondernomen. Er wordt daarmee een beroep gedaan op organiserend en probleemoplossend vermogen, op timemanagement en op stressbestendigheid. Soms vloeien uit deze opdracht andere simulatieopdrachten voort, zoals het houden van een presentatie, het voeren van een gesprek of het schrijven van een nota.
- Fact-finding opdracht.
Er wordt een probleem voorgelegd, maar er ontbreekt belangrijke informatie. Die informatie dient tijdens een gesprek met een rolspeler achterhaald te worden. Deze opdracht vergt analyserend vermogen, sociale vaardigheid en creativiteit.
- Tweegesprek.
Meestal wordt een tweegesprek uitgevoerd met een rolspeler. Afhankelijk van de functie waarop het AC is gericht, kan het gaan om een functioneringsgesprek, een verkoopgesprek of een gesprek met een klagende klant. Er wordt dan bijvoorbeeld gelet op sociale vaardigheden en wilskracht.

De Christelijke Hogeschool Noord-Nederland (CHN) heeft in haar onderwijs het gebruik van een ADC geïntegreerd. Hieronder wordt de implementatie van **een ADC bij de Faculteit Economie en Management (FEM)** beschreven als praktijkvoorbeeld van een ADC in het onderwijs.

De CHN ziet het AC met name als feedbackinstrument voor die vaardigheden waarop de student later, in de arbeidssituatie, geselecteerd en beoordeeld wordt. Men zou dus kunnen beargumenteren dat hier eerder sprake is van een development center. Het AC bij de FEM maakt gebruik van verschillende simulatieoefeningen (Snippe en Smit, 1997). Een

belangrijke eindterm voor alle opleidingen van de FEM is het verwerven van leidinggevende vaardigheden. Een van de oefeningen van de totale AC bestaat dan ook uit sociale vaardigheden en houdingsaspecten. In overige oefeningen komen vaardigheden als delegeren, plannen en analyseren aan bod.

De simulatieoefening voor sociale vaardigheden bestaat uit een tweegesprek op basis waarvan sociale vaardigheden kunnen worden beoordeeld. Gekozen is voor een situatie die voor de drie opleidingen binnen de FEM herkenbaar is: een situatie waarin de student, in de rol van manager, gaat praten met een medewerker (een geïnstrueerde acteur) op wie hij kritiek heeft. Het doel van het gesprek is te komen tot afspraken door het adequaat toepassen van gespreksvaardigheden. De setting is een restaurant, supermarkt of camping, afhankelijk van de opleiding.

De student krijgt 15 minuten om twee problemen te bespreken met de gesprekspartner en om te komen tot afspraken. De rol van de acteur is duidelijk omschreven in een script en de acteur is getraind in hoe hij moet reageren. Het script bestaat uit aanwijzingen die de acteur moet vertonen en uit instructies waarmee de acteur lastige gesprekssituaties moet creëren. Deze richtlijnen hebben tot doel om elk rollenspel zo gestandaardiseerd mogelijk te laten verlopen.

Studenten worden beoordeeld op gespreksvaardigheden en algemene houding. Hiertoe is een beoordelingslijst ontwikkeld die bestaat uit acht categorieën van vaardigheden (o.a. aandachtig luisteren, de ander in staat stellen zijn verhaal te doen, doorvragen, voor duidelijkheid in de situatie zorgen door herhalingen in eigen woorden, en de hoeveelheid verkregen informatie). Elke actie van de student wordt gecategoriseerd en gewaardeerd (met een ‘+’ voor een goede actie, een ‘±’ voor een matige actie en een ‘-‘ voor een slechte actie). Aan het eind kijkt de observator per vaardigheid naar het totaal aantal gescoorde plussen, minnen en plusminnen en geeft op grond hiervan een eindbeoordeling voor iedere vaardigheid op een schaal van 1 tot 5. Zowel de kwaliteit als de kwantiteit van de acties zijn van invloed op de beoordeling.

Gekozen is om criteria zoveel mogelijk op te splitsen naar zinvolle en herkenbare gedragseenheden. Hiermee wordt het geven van feedback vereenvoudigd en het leereffect vergroot. Wanneer een student na afloop bijvoorbeeld krijgt te horen dat hij niet goed heeft doorgevraagd, is dit eenvoudiger te veranderen dan wanneer wordt meegedeeld dat hij onjuist met het conflict omging.

1.4. Beoordeling van competenties als brug tussen organisaties en onderwijs

Een van de problemen die goede afstemming en integratie tussen onderwijs en werk in de weg staat is de *beoordeling van kwaliteiten*. De kwaliteiten waarop diplomering in het onderwijs gebaseerd is, en de standaarden waarop het bedrijfsleven selecteert en beoordeelt zijn verschillend van aard en systematiek.

In het (beroeps)onderwijs wordt gewerkt met een kwalificatiestructuur, terwijl men in het bedrijfsleven uitgaat van competentieprofielen. Het verschil tussen kwalificaties en competenties ligt in het perspectief van waaruit ze gedefinieerd zijn. Kwalificaties zijn beroepseisen die formeel zijn vastgesteld. Een competentie kan worden gedefinieerd als het in samenhang toepassen van individuele kennis, vaardigheden, en attitudes in het construeren van een oplossing of een product in een bepaalde context (Klarus, Tillema & Veenstra, 2000).

Kwalificaties worden kortom geformuleerd vanuit de vraagkant (beroepseisen), terwijl competenties betrekking hebben op de aanbodkant (persoonlijke vermogens). Het verschil in perspectief betekent ook dat in het onderwijs vooral wordt gestreefd naar het 'voldoen aan de kwalificaties', terwijl in organisaties het 'ontwikkelen van competenties' centraal staat. Een ander belangrijk punt is dat bij het competentiebegrip in organisaties de relatie met een bepaalde context of situatie veel meer benadrukt wordt (Bos, 1998). Opgemerkt moet worden dat het belang van gesitueerdheid en context bij het ontwikkelen van kennis en vaardigheden ook in het onderwijs sterke opgang vindt (denk bijvoorbeeld aan 'situated cognition' (Brown, Collins & Duguid, 1989).

Een betere aansluiting tussen de 'wereld van het werk' en 'de wereld van het onderwijs' (vergelijk van Merriënboer, 1999) is noodzakelijk. Het idee van life-long learning (zie bijvoorbeeld Onderwijsraad, 1998) betekent dat de grenzen tussen leren en werken vervagen. Een mogelijk verbindingspunt tussen leren en werken kan liggen in het competentiebegrip, met name in het beoordelen van competenties, omdat daar een scharnierpunt ligt.

De nota 'Een leven lang leren' heeft duidelijk gemaakt dat competenties niet alleen binnen de school worden verworven, maar ook vaak daarbuiten. De consequentie hiervan is dat 'elders verworven competenties' beoordeeld en erkend zouden moeten worden. Hierbij gaat het om een beoordeling van iemands verworven vermogens, die onafhankelijk zijn van de gevolgde leerweg (zij het via een de opleiding, werkervaring of anderszins). Men spreekt hier van leerwegonafhankelijk beoordelen. Hierbij wordt een expliciet onderscheid gemaakt tussen het leertraject enerzijds en verworven competenties die aan een bepaalde standaard moeten voldoen anderzijds. Het aansluitingspunt tussen opleiding en beroep zit in verworven competenties en niet in de gevolgde leerweg.

Het erkennen van elders of informeel verworven competenties (Klarus, 1998) zal moeten worden gebaseerd op een assessmentprocedure. De commissie Wijnen (1994) en Nieskens & Van Meteren (1998) geven de volgende procedure aan:
Inventarisatie van leer- en werkervaringen en vergelijking met kwalificatiecriteria
Uitvoeren van praktijkopdrachten, te beoordelen aan de hand van interviews gericht op planning, observatiechecklists gericht op uitvoering, en interviews gericht op reflectie en transfer.
Vaststellen aan welke vereisten iemand reeds voldoet. Op basis hiervan kunnen deelcertificaten worden verstrekt en vervolgopleidingen gepland.
In deze procedure is te zien dat assessment in onderwijs en organisaties steeds meer op elkaar gaan lijken. In eerste instantie is deze tendens zichtbaar in beroepsopleidingen, waar de link tussen leren en het toekomstig beroep vrij duidelijk is. Op dit moment krijgt de beoordeling van competenties ook in het academisch onderwijs veel aandacht.

Hoofdstuk 2: Een indeling in assessmentvormen

In deze paragraaf wordt een indeling gegeven waarmee verschillende assessmentvormen worden ingekaderd. De indeling vormt de rode draad voor het verdere rapport en voor het beslismodel.

De aandacht voor het toetsen van complexe vaardigheden en de sterkere integratie van assessment in het opleidingsproces heeft geleid tot een veelheid van assessmentinstrumenten en -methodes. Het beslismodel waar dit rapport uiteindelijk op gericht is, dient om beargumenteerde keuzes te kunnen maken voor bepaalde assessmentvormen in een gegeven onderwijssituatie. Hiervoor is het nodig om te kiezen voor een bepaalde indeling in verschillende assessmentvormen, die ieder een specifiek label krijgen, en op een zo helder mogelijke manier van elkaar onderscheiden worden.

Bij de bespreking van de veelheid aan assessmentvormen die ontstaan zijn, hanteren bepaalde auteurs welbepaalde indelingsprincipes, terwijl anderen zich beperken tot een opsomming of indeling in grove categorieën en beschrijving van een aantal aspecten van deze toetsvormen. Een erg voor de hand liggende, grove, indeling is in klassieke, meer op kennis gerichte toetsvormen versus alternatieve assessment (Eindrapportage ELO-project 1.1). Brookhart (1999) en Stiggins (1992) hanteren ook een vrij grove indeling waarbij vier assessmentmethodes worden onderscheiden: (1) paper and pencil tests (schriftelijke toetsen), (2) performance assessment van processen en producten, (3) mondelinge assessment en (4) portfolio assessment.

Andere auteurs kiezen voor een beschrijving van geschikte toetsvormen *in functie van* welbepaalde doelstellingen of een bepaalde onderwijscontext. Het indelingscriterium is daarbij het leerdoel en niet de verschijningsvorm van de toets. Zo bespreken van der Vleuten & Driessen (2000) een aantal vormen van alternatieve assessment (toegepast in de context van het Maastrichtse PGO-onderwijs) die zijn gekoppeld aan respectievelijk het toetsen van basiskennis, hogere cognitieve vaardigheden, beroepsspecifieke vaardigheden en algemene vaardigheden. Op dezelfde wijze maakt Moerkerke (1998) een opsomming van mogelijke toetsvormen gebaseerd op het type eindtermen van een academische opleiding en hun functionaliteit in het onderwijs.

Deze en andere gehanteerde indelingen bieden een kader om naar de veelheid van in de literatuur beschreven (alternatieve) toetsvormen te kijken. Indelingen waarbij het indelingscriterium is gelegen in de leerdoelen vormen echter een zekere beperking, doordat er geen andere aspecten dan de geschiktheid voor het leerdoel in ogenschouw worden genomen. In het beslismodel waar hier naar toe wordt gewerkt is de geschiktheid voor het leerdoel slechts één van de overwegingen om tot een keuze te komen. Daarom wordt de voorkeur gegeven aan de benadering die ook terug te vinden is bij auteurs als Brown, e.a. (1997) en Nedermeyer en Pilot (2000), namelijk het bespreken van een overzichtelijk aantal veelgebruikte assessmentvormen die herkenbaar zijn voor studenten en docenten. Op basis van een dergelijke indeling wordt verderop getracht om per categorie vanuit verschillende invalshoeken te kijken naar geschiktheid en overwegingen die verbonden zijn aan deze assessmentvormen.

Op basis van de literatuur, en de functionaliteit in het kader van een beslismodel dat een beargumenteerde keuze kan ondersteunen, wordt hier de volgende indeling in categorieën van assessmentvormen gehanteerd:

I (Klassieke) toetsen met gesloten vragen

Naar deze toetsvorm wordt ook wel verwezen met de term 'objectieve toetsing' of 'selected response items' (Hambleton, 1996). Hieronder vallen alle toetsvormen die een respons uitlokken die kan worden gescoord zonder dat het antwoord op dat moment hoeft te worden geëvalueerd. De scoring berust veelal op antwoordsleutels. Enkele voorbeelden zijn juist/onjuist vragen, multiplechoice vragen en matching vragen (waarbij zaken op de juiste manier gegroepeerd moeten worden).

II (Klassieke) toetsen met open vragen

Deze vorm van assessment bestaat uit een schriftelijke of mondelinge toets, waarbij studenten zelf hun antwoorden moeten construeren, zonder de hulp van keuzealternatieven. Omdat de door de student geconstrueerde antwoorden een open vorm hebben, zullen de antwoorden stuk voor stuk geëvalueerd moeten worden op hun kwaliteit. Dit wordt ook wel 'subjectieve' beoordeling genoemd (wat geen waardeoordeel inhoudt, noch de suggestie dat dit onbetrouwbaar zou zijn). De schriftelijke toets kan uiteenlopende formats hebben, zoals het invullen van open plaatsen in een tekst, geven van korte antwoorden of korte essay vragen (MEQs, ofwel 'Modified essay questions, Brown e.a., 1997). Onder een mondelinge toets wordt een tentamenvorm verstaan waarbij een docent vragen stelt en de student mondeling antwoord geeft.

III Performance assessment

Onder performance assessment wordt hier een vorm van assessment verstaan waarbij gebruik wordt gemaakt van complexe en relevante taken of opdrachten, die aansluiten bij leeractiviteiten en/of situaties uit het toekomstige beroepenveld. Studenten moeten hun voorkennis en hun recent geleerde kennis in samenhang met hun (cognitieve) vaardigheden aanwenden in het oplossen en uitvoeren van de taken. Aan het eind van de performance assessment wordt meestal een product van redelijke omvang opgeleverd, zoals een beargumenteerde oplossing voor een complex probleem, een rapport over een project, een presentatie, etc.. Bij de beoordeling worden meestal zowel product- als procesaspecten betrokken.

IV Assessment- en development centers

Een assessment center (AC) is een methode om het *gedrag* van studenten of toekomstige medewerkers te meten en/of te voorspellen. Aan de hand van opdrachten in een gesimuleerde praktijksituatie worden kandidaten getest op verschillende gedragsdimensies. De praktijkopdrachten zijn gebaseerd op de inhoud van de toekomstige functie, en bootsen deze zo goed mogelijk na. De beoordeling van het gedrag vindt meestal plaats door meerdere observatoren (ook vaak peer-assessment), die hier meestal voor getraind zijn. In een development center (DC) is het doel expliciet gelegen in het stimuleren van ontwikkeling en verandering van gedrag in een gewenste richting (Tillema, 1996).

V Portfolio assessment

Een portfolio is een door de student in de loop van de tijd zorgvuldig samengestelde verzameling van bewijsmateriaal die (steekproefsgewijs) aantoont wat hij of zij geleerd heeft (Brown e.a., 1997). Het 'bewijsmateriaal' van het portfolio kan bestaan uit producten, werkstukken, beoordelingsrapporten van docenten (en/of medestudenten), video-opnames, e.d.. Het portfolio omvat ook meestal reflectieverslagen over het eigen leerproces en het huidige ontwikkelingsniveau, en een uitleg van de student over wat hij of zij wil demonstreren met de verschillende onderdelen van het portfolio. Portfolio's worden veelal gebruikt in combinatie met self-assessment, in functie van het stimuleren van de eigen ontwikkeling, waarbij de feedback van een docent alleen een toetssteen is voor de student zelf.

VI Self-, peer-, en co-assessment

Bij self-, peer- en co-assessment gaat het om de specifieke wijze waarop een beoordeling tot stand komt, namelijk met of zonder een actieve inbreng van de student (deze assessmentvorm is strikt genomen van een andere orde dan de anderen, en kan in principe gecombineerd worden met ieder van de genoemde vormen). Bij self-assessment toetsen de studenten zichzelf. De nadruk ligt hierbij veelal op toetsing in functie van ontwikkeling door middel van zelfreflectie. Bij peer-assessment beoordelen studenten elkaar, zonder daarom de taak van de docent over te nemen. Ook hier ligt de nadruk meestal op formatieve beoordeling (het geven van feedback aan elkaar om zo het leerproces te bevorderen). Co-assessment tenslotte is een vorm waarbij docent en studenten gezamenlijk beslissen over de wijze van beoordelen en de geldende beoordelingscriteria.

Over de geldigheid van de indeling die is gekozen valt natuurlijk te twisten. Met name over de derde categorie die het label ‘performance assessment’ heeft gekregen, zullen de meningen uiteenlopen. Hambleton (1996) ziet performance assessment bijvoorbeeld als een verzamelnaam voor allerlei vormen van alternatieve assessment (waaronder bijvoorbeeld ook het portfolio en het ADC vallen). Zo breed is het hier niet bedoeld. Meyer (1992) definieert performance assessment meer specifiek als een systematische manier om de toepassingsvaardigheid van de lerende te meten. Het gaat er dan om of lerenden in staat zijn om kennis te gebruiken bij het oplossen van nieuwe problemen of complexe taken (zie ook Dochy & Segers, 1999). Deze definitie ligt dichter in de buurt bij wat hierboven bedoeld wordt. In vergelijking met een ADC (categorie IV) ligt de nadruk bij performance assessment zoals hier bedoeld op cognitieve aspecten, terwijl de nadruk bij ADC’s ligt op gedragsaspecten.

Zoals al eerder werd vermeld is de categorie self-, peer- en co-assessment strikt genomen geen assessmentvorm. Er is gekozen om het toch als categorie op te nemen vanuit de overweging dat de indeling functioneel moet zijn in het kader van een beslismodel. Bij een te maken keuze voor self-, peer- en co-assessment spelen soortgelijke argumenten en overwegingen een rol als bij de andere categorieën.

Hoofdstuk 3: Fundament voor beslismodel (1): Analyse met behulp van basisvragen

Om tot een beslismodel te komen worden de onderscheiden assessmentcategorieën eerst geanalyseerd op basis van twee 'instrumenten'. Beide instrumenten worden in de literatuur teruggevonden als een basis om de geschiktheid van een toetsvorm te beargumenteren (meestal afzonderlijk van elkaar). De twee 'instrumenten' verwijzen naar (1) vijf basisvragen, en (2) acht kwaliteitscriteria voor assessment. Tezamen vormen zij de fundamenten voor het beslismodel.

In de huidige paragraaf wordt het eerste analyse-instrument geïntroduceerd en toegepast op de onderscheiden assessmentvormen. Het gaat hierbij om een model met vijf basisvragen. In paragraaf 4 wordt het tweede analyse-instrument gepresenteerd en toegepast op de assessmentvormen.

3.1. Vijf basisvragen – Introductie 1e analyse-instrument

Harden (1979) onderscheidt vijf basisvragen die voor iedere assessmentmethode van toepassing zijn, en alle componenten systematisch in kaart brengen. Deze vragen zijn:
- Wat moet worden gemeten?
- Hoe moet worden gemeten?
- Waarom moet worden gemeten?
- Wanneer moet worden gemeten?
- Wie moet beoordelen?

De basisvragen zijn bedoeld om op eenvoudige maar doeltreffende wijze te werk te gaan bij de opzet van een toetsprogramma. De vragen moeten steeds allemaal beantwoord worden bij de keuze voor een adequate invulling van het assessment. De betekenis van deze vragen wordt hieronder uitgewerkt op basis van Harden (1979), Brown (1997) en van der Vleuten en Driesen (2000).

3.1.1. Wat moet worden gemeten?

Volgens Harden is dit de belangrijkste vraag die gesteld moet worden (vergelijk ook Brookhart, 1999; Stiggins, 1992). Veel problemen komen voort uit het feit dat aan deze vraag onvoldoende aandacht is besteed. Zo worden de leerdoelen vaak pas bepaald in de uitwerking van het assessment. Ook is het toetsen van kennis meestal oververtegenwoordigd in assessments en blijven vaardigheden en attitudes onderbelicht.

Bij het beantwoorden van de vraag wat moet worden beoordeeld zijn classificatiesystemen van leerdoelen bruikbaar (bijvoorbeeld Bloom, 1956). Een veelgebruikt classificatiesysteem is een indeling in kennis, vaardigheden en attitudes. Voor het vormgeven van het beslismodel wordt echter een wat verfijnder en 'moderner' classificatiesysteem voorgesteld. Dit is grotendeels gebaseerd op vijf soorten leerdoelen, of 'achievement targets' die worden onderscheiden door Brookhart (1999) en Stiggins (1992, 1997). Zij spreken over (1) kennis, (2) denken, (3) producten, (4) procedurele vaardigheden en (5) disposities.

In aansluiting op de huidige literatuur zou in plaats van over 'producten' wellicht beter kunnen worden gesproken over 'competenties'. Een competentie werd eerder in dit rapport al

gedefinieerd als het in samenhang toepassen van individuele kennis, vaardigheden, en attitudes in het construeren van een oplossing of een product in een bepaalde context (Klarus, Tillema & Veenstra, 2000). Verder wordt in afwijking van Brookharts' indeling het onderscheid tussen de categorieën 'denken' en 'procedurele vaardigheden' aangescherpt. Daarbij wordt gekozen om 'denken' te benoemen als 'cognitieve vaardigheden', en de categorie 'vaardigheden' op te vatten als 'gedragsvaardigheden' waarin uiteraard een dominante gedragscomponent aanwezig is. De categorie disposities wordt vervangen door de vaak gebruikte term 'attitudes'. De verschillende leerdoelen die worden onderscheiden in het kader van de 'wat-vraag' voor het beslismodel zijn dan:

1. Kennis.
Hiermee wordt bedoeld het herkennen en herinneren van feiten en concepten.

2. Cognitieve vaardigheden.
Hiermee wordt verwezen naar begrip en toepassing van kennis. Er is sprake van het toepassen van kennis om redeneringen op te zetten, te argumenteren, uit te leggen, te bediscussiëren en nieuwe problemen op te lossen.

3. Gedragsvaardigheden.
In deze categorie gaat het om vaardigheden waarin een dominante gedragscomponent aanwezig is. Enkele voorbeelden zijn het gebruik van de microscoop, software, de bibliotheek, maar ook gespreksvaardigheden, leiderschapsvaardigheden en dergelijke.

4. Attitudes.
Deze categorie verwijst naar de persoonlijke interesse, betrokkenheid en waardering van het vakgebied.

5. Competenties.
Hiermee wordt verwezen naar het in samenhang toepassen van kennis, cognitieve en gedragsmatige vaardigheden en attitudes in het oplossen van levensechte taken of problemen. Voorbeelden waarin dit tot uiting kan komen zijn een schriftelijk betoog, de uitvoering van een project, een rapport van een laboratorium experiment, etc.

De categorie competenties onderscheidt zich van de eerder genoemden doordat het hier juist gaat om het combineren van kennis, cognitieve- en gedragsvaardigheden, en attitudes in het oplossen van levensechte taken of problemen. Er zit aldus een logische opbouw in de systematiek (zonder daarmee te suggereren dat dit een strikte hiërarchie zou zijn).

3.1.2. Hoe moet worden gemeten?

De manier waarop wordt gemeten moet nauw aansluiten op wát er wordt gemeten. Niet ieder methode is geschikt voor ieder doel. Een assessment waarbij een bekend probleem moet worden opgelost is bijvoorbeeld niet geschikt om het toepassen van kennis te meten, daar het met behulp van gememoriseerde kennis kan worden opgelost.

De hoe-vraag kan worden uiteengelegd in twee vragen, namelijk hoe wordt gemeten en hoe wordt gescoord (Brown, 1997). Bij de vraag hoe wordt gemeten gaat het om de gebruikte methode. Enkele voorbeelden zijn: schriftelijke of mondelinge examens, essays, het observeren van probleemoplossingsgedrag, lab reports of andere producten, presentaties en het observeren van processen bij groepsprojecten. In feite zijn dit voorbeelden van de verschillende categorieën van assessmentvormen. Bij de vraag hoe er wordt gescoord gaat het er om welke instrumenten en criteria worden gebruikt. Voorbeelden hiervan zijn: checklists met expliciete criteria, scoringsschema's, scoringsdimensies, etc. (Brown, 1997).

3.1.3. Waarom moet worden gemeten?

Het doel van assessment kan verschillend zijn. Het belangrijkste verschil is of de resultaten van een assessment worden gebruikt voor ontwikkelingsdoeleinden (formatieve evaluatie) of voor beoordelingsdoeleinden (summatieve evaluatie). Formatieve evaluatie is gericht op het verbeteren van het leren. Binnen deze categorie vallen doelen als: feedback verschaffen aan studenten om hun leren te verbeteren; studenten motiveren; diagnosticeren van sterktes en zwaktes van studenten. Summatieve evaluatie is gericht op het beoordelen en certificeren van studenten; studenten moeten een bepaalde barrière passeren om door te kunnen naar een volgende fase.

Opgemerkt kan worden dat de summatieve evaluatiefunctie nog onderscheiden zou kunnen worden in selectie- of toelatingsdoeleinden versus certificeringsdoeleinden (Voeten, 2000; Wolters, 1998). Het huidige rapport en het beslismodel hebben echter als kader toetsing *binnen* een opleiding. Selectie en toelating zijn daarmee buiten beeld.

3.1.4. Wanneer moet worden gemeten?

De beoordeling kan vooraf, tijdens of na het onderwijsleerproces plaatsvinden. Vooraf meten kan de voorkennis in kaart brengen: wat hebben studenten al geleerd, en wat zou eigenlijk nodig zijn als voorkennis voor de cursus of opleiding? Meten tijdens het leerproces wordt ook wel 'continuous assessment' genoemd, of 'classroom assessment'. Deze vorm van assessment zal veelal een formatieve functie hebben, terwijl het meten aan het eind van een cursus of semester meestal een summatieve functie heeft. Het portfolio heeft het karakter van een 'continuous assessment' en heeft veelal een formatieve functie, waarbij de student op de eigen ontwikkeling reflecteert.

3.1.5. Wie moet beoordelen?

De docent, de student, de medestudenten (peers), experts of combinaties van deze personen kunnen de beoordeling uitvoeren. Met de opkomst van 'accountability' (de maatschappelijke verantwoording van onderwijsinstellingen ten aanzien van hun kwaliteitsniveau) worden ook externe examinatoren vaker betrokken in het beoordelen van studenten.

3.2. Analyse van de praktijkvoorbeelden uit paragraaf 1 aan de hand van de basisvragen

In paragraaf 1 werden in de tekst vijf praktijkvoorbeelden gegeven van implementaties van assessmentvormen. Het ging hierbij om (1) SYS, gebruikt bij de Open Universiteit Nederland (voorbeeld van 'klassieke' toetsing), (2) de Overall toets, gebruikt aan de Universiteit Maastricht (voorbeeld van performance assessment), (3) een ADC, gebruikt aan de Christelijke Hogeschool Noord-Nederland, (4) portfolio assessment, gebruikt bij de postdoctorale lerarenopleiding van het ICLON en (5) assessment in het kader van het virtueel bedrijf, gebruikt bij de opleiding tot basisontwerper bij TAS-opleidingen (voorbeeld van self-, peer-, en co-assessment).

Om de werking van de vijf basisvragen te demonstreren wordt op de volgende pagina in kaart gebracht hoe deze gerelateerd zijn aan de vijf praktijkvoorbeelden.

Bron	Voorbeeld	Waarom wordt gemeten?	Wat wordt gemeten	Hoe		Wanneer	Wie meet?
				Hoe wordt gemeten? Methode	Hoe wordt gescoord? Instrumenten		
Moerkerke (1996)	SYS	Beoordeling	Kennis en begrip	Door het systeem gegenereerde toets met gesloten en open vragen	Met behulp van een antwoordmodel en het tentamenprofiel (vormen samen itembank).	Na afloop, wanneer het de student uitkomt	Systeem
Segers (1998)	OverAll Toets	Beoordeling	Probleemoplossende vaardigheden m.b.t. authentieke problemen Cognitieve vaardigheden: definiëren; analyseren; synthetiseren; beargumenteren; evalueren.	(Open-boek)-toets met gesloten (juist/onjuist) en open vragen rondom onbekende probleemsituatie. Probleemsituatie is gebaseerd op set van artikelen, die student in de 2 vooraf-gaande weken heeft kunnen bestuderen	Per vraag is een antwoordmodel ontwikkeld	Na afloop van 2 blokken	Docent
Snippe en Smit (1997)	Assess-ment center	Bijsturing (beoordeling)	Leidinggevende vaardigheden (sociale vaardigheden, delegeren, plannen en analyseren)	Simulatie van tweegesprek	Beoordelingslijst voor observator waarop kwaliteit en kwantiteit van alle acties van student worden gescoord (3-puntsschaal)	Op meerdere momenten tijdens opleiding	2 getrainde observatoren
Beijaard, Longay-roux en Tanner (1997)	Portfolio assess-ment	Ontwikkeling	Ontwikkelingslijn Reflectievaardigheden m.b.t. het eigen leerproces en ontwikkeling als docent	Het portfolio bevat: 1. Doelen die een afspiegeling zijn van het leraarsvak 2. Bewijsmateriaal voor het behalen van die doelen 3. Reflecties op de eigen professionele ontwikkeling	Er vindt een eindgesprek plaats over het portfolio met de student, insti-tuutsdocent en stage-begeleider.	1x tijdens en 1x na afloop van de opleiding	Student, instituuts-docent en stage-begeleider
OUNL	Virtueel bedrijf	Beoordeling en ontwikkeling	Basisontwerpvaardigheden gecategoriseerd naar persoonlijke, vaktechnische, communicatieve en coördine-rende vaardigheden.	Performance assessment: groepsopdracht die bestaat uit meerdere authentieke taken Portfolio assessment (producten en reflecties)	Beoordelinglijsten voor de verschillende assessoren over de uitgevoerde taak Beoordelingslijst voor kwaliteit van het portfolio	Na afloop van iedere taak en na afloop van de groepsopdracht	Gespreks-partners; expert-, coach- en selfassess-ment.

Analyse van de assessmentvormen aan de hand van de basisvragen

	Waarom	**Wat**	**Hoe**	**Wanneer**	**Wie**
Klassieke toetsing met gesloten vragen	Primair summatief.	-Gememoriseerde kennis: gemakkelijk. -Cognitieve vaardigheden: kan maar is moeilijk.	*-Methode:* meestal (geautomatiseerde) MC-toetsen. *-Scoring:* objectief; antwoordsleutel.	*-Tijdsperspectief*: cursus. *-Tijdstip*: aan het eind, maar ook aan het begin als voorkennistoets.	Docent of geautomatiseerd.
Klassieke toetsing met open vragen	Primair summatief.	-Cognitieve vaardigheden. -Risico: onbedoeld meten van gememoriseerde kennis.	*-Methode:* meestal schriftelijke of mondelinge open vragen. *-Scoring:* subjectief, de docent interpreteert en waardeert.	*-Tijdsperspectief*: cursus. *-Tijdstip*: aan het eind.	Docent.
Performance assessment	Summatief of formatief.	Met name cognitieve vaardigheden, en competenties.	*-Methode:* complexe en realistische taken/opdrachten. *-Scoring*: wisselend, o.a. checklists, antwoordmodellen, (expliciete) criteria.	*-Tijdsperspectief*: cursus. *-Tijdstip*: aan het eind.	Docent en/of meerdere deskundigen, ook vaak co- en peer-assessment.
ADC's	Summatief of formatief.	Gedragsvaardigheden en attituden.	*-Methode:* opdrachten in gesimuleerde praktijk. *-Scoring*: meestal met gedragscriteria en beoordelingslijsten met antwoordschaaltjes.	*-Tijdsperspectief*: meestal curriculum. *-Tijdstip*: meestal op meerdere momenten.	Meerdere (getrainde) deskundige beoordelaars, ook vaak peer- en co-assessment.
Portfolio assessment	Primair formatief.	*Ontwikkelingen in*: (kennis), cognitieve- en gedragsvaardigheden, attituden en competenties.	*-Methode:* verzameling van 'bewijsmateriaal' met expliciete reflectie op ontwikkelingen. *-Scoring:* vaak brede criteria en subjectief oordeel.	*-Tijdsperspectief*: meestal curriculum. *-Tijdstip*: 'continuous'.	Self-assessment, vaak in combinatie met docent en/of andere deskundige.
Self-, Peer-, en Co-assessment	Primair formatief.	Meestal gedragsvaardigheden, competenties en attituden.	*Methode:* -peer- en co-assessment meestal bij performance assessment en ADC, -self-assessment vaak bij portfolio's. *Scoring:* meestal (expliciete) criteria en beoordelingslijsten.	*-Tijdsperspectief*: afhankelijk van toetsvorm. *-Tijdstip*: afhankelijk van toetsvorm.	Self-, peer-, en co-assessment.

Bovenstaande tabel is een vertaling van de analyse op het niveau van praktijkvoorbeelden naar een algemener niveau, en is bovendien gebaseerd op de literatuur. Deze analyse ligt direct ten grondslag aan het beslismodel. Hieronder volgt een toelichting op de tabel.

Ten eerste is de manier waarop de tabel is ingevuld gebaseerd op de *gebruikelijke* manier waarop de verschillende toetsvormen worden ingezet. Het is daarmee een afspiegeling van wat in literatuur en praktijk geschikt en voor de hand liggend wordt gevonden bij deze toetsvormen. Dit betekent echter niet dat andere invullingen uitgesloten zijn, maar ze hebben waarschijnlijk consequenties voor de vormgeving van de toets waardoor deze niet echt typerend meer zijn voor de beschreven categorie. Zo is het bijvoorbeeld best mogelijk om een portfolio te gebruiken binnen één cursus, maar dan worden bijvoorbeeld verschillende versies van een eindproduct opgenomen, waarbij een ontwikkelingslijn expliciet wordt aangegeven door de student. Een ander voorbeeld is dat een klassieke toets met gesloten vragen ook kan worden gebruikt op curriculumniveau voor continue toetsing. Hierbij krijgt de objectieve toets de vorm van een voortgangstoets bestaande uit enkele honderden items (zoals in het Maastrichtse model, zie van der Vleuten, Verwijnen & Wijnen, 1996). Een laatste voorbeeld van een afwijkende mogelijkheid is dat een portfolio of peer-assessment gebruikt wordt voor summatieve doeleinden. Omdat summatieve evaluatie hogere eisen stelt aan de kwaliteit van de beoordeling zal dit echter consequenties hebben voor de instrumenten die ontwikkeld worden voor de scoring, alsook voor de kwaliteit van de beoordelaars. Beide zullen objectiever van aard moeten worden. Bovenstaande tabel is niet uitgegaan van deze mogelijke toepassingen c.q. aanpassingen, maar dat betekent niet dat ze niet nuttig of bruikbaar zouden zijn.

In de ‘waarom’ kolom wordt de tweedeling gehanteerd die reeds in paragraaf 3.1.3 werd geïntroduceerd (summatieve en formatieve functie). Klassieke toetsing wordt meestal gehanteerd met een certificerende (summatieve) functie. Performance assessment en ADC’s kennen deze functie ook (een scriptie wordt bijvoorbeeld meestal beoordeeld met een cijfer), maar deze toetsvormen hebben ook vaak primair een formatieve functie, waarbij ze feedback verschaffen aan studenten zodat hun verdere ontwikkeling wordt bevorderd en gestuurd. Assessment centers zijn in oorsprong ontwikkeld voor personeelsselectie en kunnen dus goed een summatieve functie dienen. Voorwaarde is het gebruik van heldere criteria en getrainde observatoren. Development centers zijn juist expliciet bedoeld voor ontwikkelingsdoeleinden. Van portfolio’s en self-, peer-, en co-assessment wordt doorgaans eerder heil verwacht bij gebruik in formatieve zin. Dit heeft te maken met een moeilijk te verwerkelijken hoge mate van betrouwbaarheid in scoring, terwijl de ‘leerzaamheid’ van de twee categorieën juist in potentie erg groot is.

In de ‘wat’ kolom is gebruik gemaakt van de indeling in vijf soorten leerdoelen (of soorten inhoud) die eerder werd gegeven in paragraaf 3.1.1. Klassieke toetsing met gesloten vragen is geschikt voor het meten van (gememoriseerde) kennis. Het meten of studenten kennis ook kunnen toepassen en gebruiken in nieuwe situaties (cognitieve vaardigheden) is mogelijk, maar vereist veel investering in de constructie van de vragen. Bij klassieke toetsing met open vragen is het laatste eenvoudiger te realiseren, hoewel de valkuil om onbedoeld gememoriseerde kennis te meten aanwezig is. Cruciaal is dat de toetsing betrekking heeft op onbekende (probleem)situaties.

Performance assessment vereist van studenten dat zij kennis en vaardigheden in samenhang toepassen in het construeren van een oplossing of een product (met andere woorden competenties), en biedt dus potentieel goede kansen om ook daadwerkelijk deze leerdoelen te meten. In klassieke toetsing is het meten van competenties nauwelijks mogelijk. ADC’s zijn meestal gericht op het meten van gedrag in gesimuleerde situaties, waardoor ze geschikt zijn voor het meten van gedragsvaardigheden en attituden. Portfolio assessment geeft een

specifieke kijk op alle soorten leerinhouden, namelijk een ontwikkelingsperspectief. Meestal wordt het meten van ontwikkelingen in vaardigheden en attitudes beoogd. Self-, peer-, en co-assessment zijn in principe toe te passen bij alle andere toetsvormen, maar worden meestal ingezet bij ADC's, performance en portfolio assessment, waardoor meestal gedragsvaardigheden, competenties en attituden worden gemeten.

Zoals beschreven in paragraaf 3.1.2 valt de 'hoe' vraag uiteen in de methode van meting en de wijze van scoring. De methode is eigenlijk een verbijzondering van de toetsvorm. Het geeft een concreter beeld van de genoemde categorieën. Qua wijze van scoring is alleen de eerste assessmentvorm als objectief te typeren. Alle overige toetsvormen vallen onder de noemer van 'subjectieve scoring'. In het algemeen is subjectieve scoring betrouwbaar(der) te maken door te werken met expliciete criteria, antwoordmodellen en/of meerdere beoordelaars.

In de 'wanneer' kolom is een tweedeling gehanteerd in het tijdsperspectief waar de toetsvorm zich meestal op richt en het tijdstip waarop de toets meestal plaatsvindt. De meeste toetsvormen richten zich op een cursus, of studie-eenheid. Verder vinden de meeste toetsen plaats aan het einde van de cursus. ADC's en portfolio's zijn vaak op een wat groter deel van het curriculum gericht. In een ADC worden dan bijvoorbeeld bepaalde kernvaardigheden gemeten op verschillende momenten tijdens de studieloopbaan. Dit verschaft studenten en docenten feedback en kan ook ontwikkelingen aangeven. Het portfolio heeft uitdrukkelijk een langer tijdsperspectief gericht op het in kaart brengen van ontwikkelingen.

In de 'wie' kolom is tenslotte aangegeven welke personen doorgaans de rol van beoordelaar vervullen. In klassieke toetsing is dat de docent of het geautomatiseerde systeem. Bij performance assessment, ADC's en portfolio's worden vaak meer personen betrokken, zoals deskundigen uit het beroepenveld, peers, en de studenten zelf. Bij self-, peer-, en co-assessment gaat het vanzelfsprekend over deze groepen.

Hoofdstuk 4: Fundament voor beslismodel (2): analyse met behulp van kwaliteitscriteria

In deze paragraaf worden de onderscheiden assessmentcategorieën geanalyseerd op basis van een model met acht kwaliteitscriteria. Dit betreft het tweede analyse-instrument waarop het beslismodel is geënt. In 4.1 wordt het tweede model of instrument gepresenteerd, waarna in 4.2 de zes assessmentvormen worden besproken in relatie tot ieder van de kwaliteitscriteria.

4.1. Kwaliteitscriteria voor assessment – Introductie 2e analyse-instrument

Er zijn gevestigde psychometrische criteria voor het beoordelen van de technische kwaliteit van toetsen en metingen. De meeste zijn afgeleid van de fundamentele concepten validiteit en betrouwbaarheid. Meestal wordt het begrip validiteit omschreven als een match tussen hetgeen wat je bedoelt te meten en hetgeen je werkelijk meet (Brown, Bull & Pendlebury, 1997). In het specifieke geval van assessment gaat het dan vooral om de aansluiting tussen het gedrag of de prestaties die je toetst, en de leerdoelen die je met het onderwijs beoogt. Betrouwbaarheid verwijst naar de mate waarin een score (zij het een objectieve of subjectieve score) consistent is, bijvoorbeeld tussen beoordelaars, en binnen een beoordelaar (Brookhart, 1999).

Omdat performance assessment een directe vorm van prestatiemeting is, en niet zoals traditionele toetsing alleen indirecte indicatoren meet, lijkt performance assessment potentieel zeer valide te zijn. Dit zal echter moeten worden aangetoond door het verzamelen van bewijsmateriaal ofwel evidentie (Birenbaum, 1995). Messick (1983, 1994) plaatst twee zeer belangrijke bedreigingen van validiteit op de voorgrond, te weten 'construct underrepresentation' (zie ook hieronder bij criterium 6) en 'construct irrelevant variance' (zie ook bij criterium 2). Bij onderrepresentatie wordt een (te) klein deel van het construct in beeld gebracht en missen er essentiële onderdelen van het construct. Bij het meten van irrelevante variantie spelen er onterecht competenties een rol in de beoordelingen die geen onderdeel vormen van het te meten construct. Minimale construct onderrepresentatie en minimale construct irrelevante variatie zijn de twee basisprincipes van validiteit (Messick, 1994).

Het lijkt gerechtvaardigd om de traditionele betekenis van de begrippen validiteit en betrouwbaarheid uit te breiden, gezien de geclaimde waarde van nieuwe benaderingen van assessment. Volgens Linn, Baker & Dunbar (1991) is er altijd teveel nadruk gelegd op betrouwbaarheid ten koste van validiteit en is het begrip validiteit te eng gedefinieerd. Het onderstaande raamwerk is aangepast aan actuele theoretische inzichten met betrekking tot validiteit, en op de aard en het potentiële gebruik van nieuwe vormen van assessment (Linn e.a., 1991). Het raamwerk bestaat uit acht criteria waarin wordt beschreven welke evidentie van validiteit (validity evidence) nodig is gezien de huidige ontwikkelingen op het gebied van assessment (zie ook Hambleton, 1996).

1. Consequenties

Hierbij gaat het om de vraag wat de (on)gewenste consequenties van de assessment zijn, in termen van ontplooide instructieactiviteiten en ontplooide leeractiviteiten. Het gaat dus in

feite om de invloed die assessment heeft op het leren zelf. Een specifiek geval binnen 'consequential validity' is zogenaamde 'systemic validity' (Messick, 1994; Frederiksen en Collins, 1989). Systemic validity is gericht op de vraag of de toets veranderingen in curriculum, instructie en leerstrategieën van studenten teweegbrengt, en of het de ontwikkeling van de kennis/vaardigheden of competenties die het wil meten bevordert. In feite gaat het dan vaak om de validiteit van het hele onderwijssysteem, en niet alleen om de validiteit van de toetsing. Dit soort criteria worden vaak afgewogen tegen criteria als efficiëntie en betrouwbaarheid (Messick, 1994).

2. Rechtvaardigheid

Bij rechtvaardigheid wordt gedacht aan de vraag of er geen construct irrelevante variantie wordt gemeten. Dit is bijvoorbeeld het geval als niet ter zake doende competenties (zoals schriftelijke uitdrukkingsvaardigheid) een rol spelen en de uitkomsten beïnvloeden. Bepaalde culturele groepen kunnen hier specifiek voor- of nadelen van ondervinden. Dit criterium ligt het dichtst bij betrouwbaarheid.

3. Generaliseerbaarheid en transfer

De prestaties op assessmenttaken moeten generaliseerbaar zijn naar een bredere klasse van taken, en het moet aantoonbaar zijn dat de prestaties op de assessmenttaak ook vertaald kunnen worden naar levensechte taken. Men is natuurlijk niet alleen geïnteresseerd in de specifieke kennis en vaardigheden die de student in die concrete situatie toepast, maar in 'something beyond the specific task, … a larger domain of knowledge and/or skills' (Herman, 1992). De vraag is of de assessment een goede indicatie geeft van het gebruik van de kennis, vaardigheid, competentie, etc. in het toekomstige beroep.

4. Cognitieve complexiteit

Complexe taken garanderen nog geen beroep op complexe denkprocessen. Zelfs de moeilijkste taken kunnen soms goed uitgevoerd worden door dingen uit het hoofd te hebben geleerd. Duidelijk gestructureerde en gedecontextualiseerde opdrachten (zoals klassieke vraagstukken in de wetenschappen) kunnen door studenten veelal worden opgelost op basis van herkenning van het probleem en het uit het geheugen toepassen van vooraf geoefende procedures. Cognitieve complexiteit wordt vooral bepaald door de mate waarin de student zelf de probleemsituatie dient te structureren en eigen oplossingsmethodes te bedenken in functie van de context (als-dan redeneringen)

Om enig zicht te verwerven op de cognitieve complexiteit van opdrachten voor studenten kan men hen hardop laten reflecteren over de wijze waarop zij dit soort problemen aanpakken, en de bestede tijd registreren (Birenbaum, 1996).

5. Inhoudskwaliteit

Evidentie moet verzameld worden met betrekking tot het belang en de relevantie van de inhoud van de taak. De inhoud moet overeenkomstig zijn met de meest actuele inzichten in het veld, en experts op dit gebied moeten taken goedkeuren als zijnde contextgebonden, betekenisvol en de moeite waard van de tijdsinvestering.

6. Inhoudsrepresentatie

Evidentie is nodig voor het aantonen van de adequaatheid van de steekproef van taken ten opzichte van het totale curriculum. Een te smalle representatie kan invloed hebben op het onderwijs (dat zich ook kan gaan beperken) en op de effectiviteit van coaching.

7. Transparantie/betekenisvolheid

De assessment moet betekenis hebben voor studenten en docenten. Er wordt mee bedoeld dat de assessment zelf een zinvolle onderwijservaring moet zijn, die de motivatie van studenten bevordert en die sturing en richting geeft aan het verdere leren (Messick, 1994).

Studenten moeten daarnaast weten wat er gemeten wordt, criteria en normen voor goede prestaties moeten duidelijk zijn, inclusief aanwijzigen voor hoe zij hun prestaties op dat niveau kunnen krijgen. Frederiksen en Collins (1989) noemen dit laatste transparantie of doorzichtigheid.

8. Efficiëntie en kosten

Evidentie is nodig om aan te tonen dat de tijd en moeite die nodig is voor het maken en afnemen van de assessment zich loont in termen van positieve consequenties in relatie tot negatieve consequenties. Als de toets gebruikt wordt voor summatieve doeleinden zal de afweging anders liggen dan bij toetsing met formatieve doeleinden. Het belang van de gevolgen van de beslissing maakt dat bij summatieve toetsing meer aandacht moet worden besteed aan standaardisering. Bij formatieve toetsing zal de investering eerder op andere vlakken liggen.

4.2. Analyse van de assessmentvormen aan de hand van de kwaliteitscriteria

4.2.1. Klassieke toetsing met gesloten vragen

Consequenties

Met betrekking tot de 'consequential validity' valt te verwachten dat klassieke toetsing met gesloten vragen kan leiden tot leergedrag bij studenten dat gekenmerkt wordt door veel herhaling en gedachteloos memoriseren. Ook de docent wordt bij deze vorm van assessment gestimuleerd tot het volstaan met het overdragen van kennis en aanbieden van leerinhouden, los van vaardigheden, attitudes of toepassingen in concrete situaties. In de recente ideeën over onderwijs (bijvoorbeeld competentiegericht onderwijs) volstaat deze vorm van assessment niet, omdat het niet in overeenstemming is met de leerdoelen. Het SYS toetssysteem dat bij de OUNL gebruikt wordt bestaat voor een groot deel uit gesloten vragen. Studenten van de OUNL vinden de feedback die wordt verstrekt op summatieve beoordelingen met behulp van SYS niet informatief genoeg. Dit betekent dat er weinig sturing uitgaat van de toetsing en deze in die zin weinig consequenties voor het leerproces heeft.

Rechtvaardigheid

Een voordeel van gesloten vragen is dat de betrouwbaarheid achteraf (statistisch) kan worden gecontroleerd (Elsen, 1998). De psychometrische eigenschappen van de toetsitems worden bij SYS bijvoorbeeld regelmatig door toetsdeskundigen onderzocht.

Generaliseerbaarheid en transfer

Of goede prestaties op gesloten vragen ook te generaliseren zijn naar levensechte taken is maar zeer de vraag. Als dit wel wordt verondersteld kan niet volstaan worden met gesloten vragen.

Cognitieve complexiteit

Meestal is de cognitieve complexiteit van gesloten vragen niet erg groot. SYS blijkt bijvoorbeeld met name feitenkennis te toetsen en soms stemt de module-inhoud niet overeen met de moduletoets (Verreck en Weges, 1994). Het meten van cognitieve complexiteit met gesloten vragen is echter niet uitgesloten. Dit vereist echter veel vakmanschap en tijdsinvestering voor het maken van goede vragen (ook wel 'stimuli').

Inhoudskwaliteit

Hiervoor geldt hetzelfde als wat hierboven vermeld is. In principe kan deze toetsvorm goede inhoudskwaliteit bieden maar dit hangt af van de investeringen in de gemaakte vragen.

Inhoudsrepresentatie
Als het gaat om goede representatie van de te leren kennis biedt deze categorie van toetsing goede kansen. Volgens Brookhart (1999) is het (samen met open vragen) de beste manier om een breed domein van kennis te bestrijken.

Transparantie/betekenisvolheid
Transparantie en betekenisvolheid kunnen nogal eens tekort schieten bij objectieve toetsing met open vragen. Studenten van de OUNL constateerden bijvoorbeeld dat de gesloten vragen van SYS nauwelijks vaardigheden of attitudes meten. Zij hadden het gevoel dat hun studie-inspanningen niet beloond werden en slechts feitenkennis gevraagd werd in plaats van begrip en probleemoplossende vaardigheden (Moerkerke, 1996). Dit ervoeren zij als een minpunt van het toetssysteem.

Efficiëntie en kosten
Een belangrijk voordeel is dat het nakijken van gesloten vragen zeer snel kan gebeuren, en ook volledig geautomatiseerd kan worden. Daardoor kan deze toetsvorm gemakkelijk op grote schaal worden gebruikt. Het nakijken is bovendien zeer betrouwbaar. Ook vanuit administratief oogpunt kan deze vorm van toetsing voordelen bieden. De administratieve kwaliteit van SYS is bijvoorbeeld hoog, er worden in de organisatie van tentamens nauwelijks fouten gemaakt.
Een nadeel is dat het maken van vragen en het up-to-date houden van een databank met gesloten vragen wel tijdrovend is. De ervaring bij SYS is dat veel toetsitems worden afgekeurd in de controles, uitgevoerd door toetsspecialisten en dat itemconstructie onvoldoende efficiënt plaatsvindt (Verreck en Weges, 1994).

Samengevat
Gesloten vragen zijn vooral geschikt voor het meten van kennis. Dit kan zowel gememoriseerde kennis, als toegepaste kennis zijn (wat hogere eisen aan de vragen stelt). Een belangrijk voordeel van objectieve toetsing is dat dit de meest betrouwbare methode is voor het meten van kennis en begrip, en de beste manier om een brede reeks van kennis te bestrijken (Brookhart, 1999). Een valkuil is echter dat gesloten vragen met objectieve scoring snel leiden tot het toetsen van gememoriseerde kennis. Het maken van gesloten vragen, en met name van vragen die inzicht, begrip en toepassing van kennis meten, vergt veel vaardigheid en is tijdsintensief (Elsen, 1998). Nakijken van gesloten vragen is daarentegen een zeer gemakkelijk en snel proces.

4.2.2. Klassieke toetsing met open vragen

Consequenties
Goede open vragen en mondelinge tentamens kunnen tot gevolg hebben dat leren en instructie ook gericht worden op het leren gebruiken, begrijpen en toepassen van kennis. Een dergelijke toetsvorm kan er toe leiden dat studenten vragen om meer gelegenheid tot oefening en feedback, of dat docenten hier meer aandacht aan gaan geven. Open vragen die beantwoord kunnen worden op basis van gememoriseerde kennis zullen echter eerder leeractiviteiten stimuleren die gekenmerkt worden door (klakkeloos) van buiten leren. Klassieke toetsing wordt meestal gekenmerkt door normgeoriënteerde beoordeling (vergelijking met andere studenten of met een gemiddelde). Dit geeft weinig ondersteuning en sturing aan het leerproces, in vergelijking met criteriumgeoriënteerde toetsing.

Rechtvaardigheid
Een nadeel van open vragen kan zijn dat de schriftelijke uitdrukkingsvaardigheid van studenten meetelt in de beoordeling, terwijl de assessment geen schriftelijke uitdrukkingsvaardigheid bedoelt te meten. Bij mondelinge tentamens geldt dezelfde redenering, voor mondelinge uitdrukkingsvaardigheid. Beide vormen kunnen nadelig zijn voor studenten met Nederlands als tweede taal.

Generaliseerbaarheid en transfer
Of de prestaties op klassieke open vragen en mondelinge tentamens te generaliseren zijn naar levensechte taken is niet duidelijk. In zoverre ook in de levensechte situaties dergelijke problemen voorkomen valt te verwachten dat transfer optreedt.

Cognitieve complexiteit
Een hoge mate van cognitieve complexiteit kan gerealiseerd worden in deze categorie van assessment. Dit vergt echter goede vragen die deze complexiteit in zich hebben.

Inhoudskwaliteit
Kwaliteit van inhoud kan ook gerealiseerd worden in deze toetsvorm. Het zal afhangen van de inspanningen en inzichten van de maker van de toets.

Inhoudsrepresentatie
Volgens Brookhart (1999) is deze toetsvorm (samen met gesloten vragen) de beste manier om een breed domein van kennis te bestrijken.

Transparantie/betekenisvolheid
Het is van belang om studenten duidelijk te maken wat voor soort vragen zij kunnen verwachten, vooral omdat deze toetsen meestal summatief worden ingezet. De betekenisvolheid van deze categorie van toetsen voor studenten loopt in principe geen gevaar, tenzij er teveel op feitenkennis wordt getoetst.

Efficiëntie en kosten
De efficiëntie is vrij hoog, hoewel het maken en nakijken van een toets die cognitieve complexiteit meet de nodige investering kost. Het afnemen op grote schaal is eenvoudig (bij schriftelijke vragen), maar het nakijken van open vragen voor grote groepen is tijdrovend.

Samengevat
In vergelijking met de vorige categorie bieden open vragen en mondelinge toetsen in principe een grotere kans op het meten van inzicht, begrip en toepassing van kennis. Indien gevraagd wordt om bijvoorbeeld een onbekende situatie te analyseren, en het antwoord te beargumenteren, wordt daadwerkelijk inzicht en toegepaste kennis gemeten. Ook hier bestaat echter de valkuil om vragen te stellen die een student ook goed kan beantwoorden als hij of zij slechts gedachteloos zaken uit het hoofd heeft geleerd. De vragen zijn over het algemeen gemakkelijker te maken dan gesloten vragen, maar het nakijken daarentegen is weer minder betrouwbaar en meer tijdrovend.

4.2.3. Performance assessment

Consequenties
Performance assessment, inclusief de beoordeling en feedback achteraf, zijn vaak sterk geïntegreerd in het instructieproces en voornamelijk gericht op het bevorderen en sturen van de leeractiviteiten. Doordat meestal sprake is van criteriumgeoriënteerd toetsen waarbij de prestaties van een student met vooropgestelde doelstellingen en criteria worden vergeleken, geeft de assessment sturing en ondersteuning aan de leeractiviteiten van de student. De

toetsvorm draagt er dan toe bij dat zowel docenten als studenten zich gaan concentreren op de essentiële doelstellingen van de opleiding.

Zonder een grondige procedure bij het samenstellen en analyseren van de taken en beoordelingsprocedures, en zonder docenten en studenten te betrekken bij deze procedure, valt er echter weinig impact te verwachten op het leer- en instructieproces (Herman, 1992; McDowell & Sambell,1999). Bij ondoordacht gebruik van assessment kunnen studenten in de loop van het leerproces overbelast geraken, zodat de herhaalde aanwijzingen of zelfconfrontaties niet echt bijdragen tot een rustige ontwikkeling.

Rechtvaardigheid

Performance assessment bestaat vaak uit open opdrachten met open antwoorden en oplossingen, wat objectiviteit bij de scoring in de weg staat (Elsen, 1998). Fleming (1999) geeft een overzicht van mogelijke foutenbronnen bij het scoren van open vragen. Omdat er in het onderwijs weinig ervaring is met performance assessment zijn instructies bij de taken en vooral de scoringsregels en criteria niet altijd voldoende helder en ondubbelzinnig geformuleerd. Dit kan de rechtvaardigheid van de beoordeling van prestaties in het gedrang brengen.

Generaliseerbaarheid en transfer

Omdat authenticiteit van taken één van de belangrijkste kenmerken van performance assessment is, biedt deze toetsvorm een grote kans op generaliseerbaarheid van resultaten naar levensechte taken. Elsen (1998) geeft bijvoorbeeld aan dat de casustoets (bijvoorbeeld de OAT) zeer geschikt is voor het toetsen van wetenschappelijke werk- en denkvaardigheden, onderzoeks-(deel)vaardigheden, en voor het beoordelen van beroepsmatige competenties. Belangrijk voor het trekken van generaliserende conclusies is echter ook de representativiteit van de taken voor de beroepspraktijk. Van der Vleuten & Driessen (2000) en Herman (1992) komen tot de vaststelling dat om tot generaliseerbare uitspraken te komen over de competenties van een persoon, tussen de 10 a 20 metingen (opdrachten) zouden moeten plaatsvinden.

Cognitieve complexiteit

In het algemeen is performance assessment geschikt voor het meten van diepgaand begrip en denken, en toepassing van kennis. De student wordt gevraagd zelf een taak uit te voeren of een probleem op te lossen. Als de opdrachten of problemen zodanig geformuleerd zijn dat de student ze alleen op een actieve en creatieve manier kan oplossen met gebruikmaking van eerder verworven kennis (Birenbaum, 1996) meet de performance assessment een hoge mate van cognitieve complexiteit (constructie, bewerking en toepassen van kennis). Bij de beoordeling wordt meestal ook aandacht besteed aan het verloop van het antwoordproces en zelfs de (meta)cognitieve strategieën zoals planning en uitvoering en zelfcontrole. Hierdoor biedt performance assessment grote kans op het voldoen aan het criterium van cognitieve complexiteit.

Inhoudskwaliteit

In vergelijking met andere toetsvormen biedt performance assessment niet meer of minder kans op een goede inhoudelijke kwaliteit. De inhoudelijke kwaliteit van een toets is afhankelijk van de onderlegdheid en ervaring van de samenstellers van de toets.

Inhoudsrepresentatie

Omdat een performance assessment meer prioriteit geeft aan diepgang en aan het meten van gecombineerde toepassing van kennis, vaardigheid en attitudes in een levensechte (probleem)situatie kan dit ten koste gaan van het representeren van een volledig inhoudsdomein. Dit is dus eerder een zwak punt van performance assessment.

Door het voorkomen van nodeloze complexiteit, zowel wat betreft te beoordelen dimensies als qua taken of opdrachten, kan meer ruimte vrijkomen voor de representativiteit van de inhoud. Een mogelijke oplossing hier biedt de 'key features' methode (van der Vleuten, 1998; van der Vleuten & Driessen, 2000) waarbij de student gevraagd wordt alleen de essentiële stappen te doorlopen bij het oplossen van een probleem of het uitvoeren van een taak (en de beoordeling ook hier alleen op is gericht). Volgens Elsen (1998) kan de casustoets bijvoorbeeld voldoende representatief worden gemaakt.

Transparantie/betekenisvolheid

Doorgaans beleven studenten deze vorm van assessment als een meer zinvolle leerervaring zodat de motivatie voor leren versterkt wordt. Toch waarschuwen een aantal auteurs voor juist een mogelijk averechts effect doordat performance assessment vrij vlug teveel druk zet op studenten. Dit kan tot gevolg hebben dat hun studieactiviteit zich gaat beperken tot een zo economisch mogelijk uitvoeren van de opgelegde taken.

Voor een succesvolle introductie van performance assessment komt het er dan ook op aan dat de studenten weten wat er van hen verwacht wordt, wat er gemeten wordt, welke aspecten en welke criteria hierbij van toepassing zijn. Een juiste perceptie van de studenten kan bevorderd worden door studenten vanaf de aanvang van het onderwijs te betrekken bij de keuze van de taken en het opstellen van de scoringsrubrieken en criteria, oefeningen te voorzien van modelantwoorden of hen eventueel te betrekken bij de beoordeling van conceptversies van medestudenten (peer-assessment).

Efficiëntie en kosten

Performance assessment legt doorgaans heel wat beslag op de tijd van de docent, zowel bij constructie van de opdrachten, de analyse van de taak in relevante deelcompetenties, noodzakelijk voor het op voorhand vastleggen van de scoringsregels en criteria, en bij de afname, scoring en rapportering (feedback). Belangrijk is de experimenteerruimte die docenten en studenten elkaar geven om samen een voor beide partijen zinnige en bruikbare vorm van beoordeling uit te bouwen.

Het is mogelijk dat na een periode van investeren het gebruik van performance assessment minder tijdrovend wordt. Het construeren van de OAT vroeg met name de eerste drie jaren een aanzienlijke investering van de docenten. Na een aantal jaren liep dit met de helft terug. Dit geldt ook voor de correctie van de open vragen. Door het zo zorgvuldig mogelijk formuleren van de antwoordmodellen is de correctietijd erg teruggelopen (Segers, 1998).

Samengevat

De waarde en het gebruik van performance toetsen ligt voornamelijk in de sturing en activering van het leer- en instructieproces. Essentiële doelstellingen van de opleiding worden door performance assessment beter zichtbaar en krijgen dientengevolge meer aandacht in het leer- en instructieproces. Daarnaast is door het authentieke karakter van de toetsen de generaliseerbaarheid en transfer naar levensechte situaties een sterk punt. Om de toetsing ook representatief te laten zijn is echter een veelvoud van taken noodzakelijk. De risico's en kritische voorwaarden van performance assessment liggen in de transparantie en de communicatie tussen studenten en docenten. Om een voor beide partijen zinnige en bruikbare vorm van beoordeling uit te bouwen is experimenteerruimte nodig. Bij gebruik in het kader van selectie of certificering (summatieve evaluatie) is er meer behoefte aan formalisering van de toetsprocedures en standaardisatie.

4.2.4. Assessment- en Development Centers

Consequenties

Omdat een AC of DC beroepsrelevant gedrag meet, bestaat de kans dat studenten in hun opleiding zich hier ook meer op willen gaan richten. Snippe en Smit (1997) concludeerden bijvoorbeeld dat het ADC een goed instrument is voor het geven van feedback, omdat het inzicht in sterktes en zwaktes verschaft. Hierdoor kan het ook sturend zijn voor de verdere ontwikkeling van de student. De consequentie voor de instelling is dan wel dat een student die inzicht heeft verworven en verbeterpunten heeft geformuleerd, ook binnen de school de mogelijkheid dient te krijgen zich te verbeteren op deze punten. Verder geven de resultaten van een ADC inzicht in de effectiviteit van het onderwijs. Nagegaan kan worden of de eindtermen worden bereikt. De resultaten van een AC kunnen dus consequenties hebben voor het aanpassen van het curriculum van een opleiding.

Rechtvaardigheid

Er zijn bij ADC's een aantal bekende valkuilen, die de beoordeling minder betrouwbaar maken. Training van beoordelaars op deze zaken kan het resultaat verbeteren. Bekende beoordelingsfouten zijn:

- Het halo-effect: een positieve score op een dimensie kan ook de andere scores positief beïnvloeden; bijvoorbeeld een initiatiefrijke en enthousiaste student kan de beoordelaars gemakkelijk om de tuin leiden en tevens een positieve beoordeling krijgen op andere dimensies dan op initiatief.
- Het horn-effect: een negatieve score op een dimensie kan ook de andere scores negatief beïnvloeden.
- De neiging om een kandidaat die een sympathieke indruk maakt mild of juist heel streng te beoordelen, omdat beoordelaars bang zijn dat hun sympathie de scores kan vertekenen (Kuilen, 2000).

Generaliseerbaarheid en transfer

De kans dat prestaties op gesimuleerde situaties vertaalbaar zijn naar levensechte situaties is redelijk groot. In vergelijking met klassieke toetsen is de vergelijkbaarheid van de taaksituatie met de praktijk in ieder geval een stuk groter.

Cognitieve complexiteit

Bij ADC's ligt de nadruk meestal op het meten van gedrag via observaties. Deze kwaliteiten zijn vaak niet echt meetbaar via klassieke toetsen. Met name het meten van complex gedrag in een (gesimuleerde) authentieke situatie kan op valide wijze plaatsvinden in een AC.

Inhoudskwaliteit

Er kan in een ADC gezorgd worden voor voldoende inhoudskwaliteit. Dit betekent wel dat opdrachten regelmatig up-to-date moeten worden gemaakt. Daar het ontwikkelen van goede taken behoorlijk wat tijd vergt zal dit misschien niet al te vaak gebeuren.

Inhoudsrepresentatie

Goede inhoudsrepresentatie is moeilijk te bewerkstelligen met ADC's. Omdat er op een intensieve en diepgaande wijze naar de uitvoering van processen en gedrag wordt gekeken is een goede afspiegeling van de inhoud van een vakgebied hier van ondergeschoven belang. Qua inhoud is er in een ADC meestal maar plaats voor een kleine steekproef.

Transparantie/betekenisvolheid

Er dienen duidelijke verwachtingen geëxpliciteerd te worden naar de studenten met betrekking tot de taken en praktijkopdrachten, anders heeft de assessment niet die betekenis die gewenst is. Voor het beoordelen van kandidaten zijn goede criteria nodig. Duidelijkheid

hierover heeft een positieve invloed op de betrouwbaarheid van de beoordeling. Het belang van ADC's zal met name voor beroepsgerichte studenten snel duidelijk zijn.

Efficiëntie en kosten
Als nadeel van ADC's kan worden beschouwd dat de invoering ervan gepaard gaat met verhoogde kosten: docenten moeten getraind worden in het afnemen van assessmenttaken en het scoren ervan (Sluijsmans, 1998). Afname van performance assessment duurt langer dan bij klassieke toetsen: de taken moeten zorgvuldig worden ontwikkeld. Hetzelfde geldt voor de scoring. Geconcludeerd wordt dat deze nadelen met name spelen wanneer sprake is van 'accountability' doeleinden (Mello, 1993).

Samengevat
ADC's zijn met name geschikt voor het meten van gedrag. Ze zijn minder geschikt voor het meten van kennis van vakinhoud, omdat slechts een kleine representatie van vakinhoud een plaats kan krijgen. ADC's worden of werden in (werk)organisaties veel gebruikt voor selectiedoeleinden, maar de huidige trend in organisaties en onderwijs is om ADC's vooral als Development Center te zien, waarbij het doel is om feedback aan deelnemers te verschaffen zodat verdere ontwikkeling van de juiste gedragsvaardigheden kan worden gestimuleerd. Betrouwbaarheid van de scoring behoeft veel aandacht bij deze categorie van assessment. Training van observatoren en expliciete criteria kunnen de betrouwbaarheid vergroten.

4.2.5. Portfolio assessment

Consequenties
Portfolio assessment heeft vooral een waarde als 'tool for learning' omdat door het regelmatig zelf beoordelen van de eigen ontwikkelingen en het verkrijgen van informatieve feedback veel bijsturing van het leerproces plaatsvindt. Zelfreflecties zoals eigen sterke en zwakke punten en mogelijkheden of problemen die men ziet om zich verder te ontwikkelen vormen een ideaal kader voor feedback en bijsturing van het leerproces. Met name als er duidelijke criteria en standaarden zijn opgesteld voor het gewenste prestatieniveau ('criterion-referenced' assessment) zal de sturende werking op het leerproces groot kunnen zijn. Een grote rol hierbij speelt dat door de nadruk op self-assessment er meer controle en 'ownership' over het leer- en instructieproces bij de student zelf wordt gelegd (Brookhart, 1999). De ontwikkeling van metacognitieve vaardigheden en zelfverantwoordelijkheid wordt gestimuleerd, wat een belangrijke bijdrage levert aan de ontwikkeling tot 'reflective practitioners' (Dochy & Segers, 1999). Er moet hier wel opgemerkt worden dat nog weinig onderzoek is verricht naar deze verwachte consequenties, en dat een en ander natuurlijk zal afhangen van de kwaliteit en zorgvuldigheid waarmee de toetsvorm wordt gerealiseerd.

Rechtvaardigheid
Bij inzet voor summatieve evaluatie dient de inhoud (welk documentatiemateriaal?) sterk gestandaardiseerd te zijn, alsook de beoordelingsprocedures. Hiermee wordt echter het eerder genoemde 'ownership' en eigenlijke kracht van portfolio's om het leerproces bij te sturen aan banden gelegd (Wolf, 1999). Bovendien zijn slechts weinig werkwijzen op dit terrein voorhanden die voldoen aan de criteria zoals die in de literatuur worden gesteld aan beoordelingsinstrumenten voor personeelsbeslissingen (Wubbels, van Tartwijk en Brekelmans, 1996).

Generaliseerbaarheid en transfer
De inhoud van een portfolio bevat meestal documentatie met een grote mate van authenticiteit. Het prestatieniveau dat hieruit blijkt is daardoor waarschijnlijk ook generaliseerbaar naar levensechte situaties.

Cognitieve complexiteit
Een portfolio is vooral geschikt om progressie in vaardigheden en competenties te demonstreren. Het al of niet meten van cognitieve complexiteit is hier dus niet echt van toepassing. Een voorbeeld waarbij een portfolio wel cognitieve complexiteit meet is wanneer verschillende gereviseerde versies van een paper worden verzameld, met daarbij een reflectie op voortschrijdend inzicht (Brookhart, 1999).

Inhoudskwaliteit en Inhoudsrepresentatie
Het primaire doel van portfolio assessment is duidelijk niet gelegen in het bestrijken van een volledig inhoudsdomein, en de nadruk ligt ook niet op het meten van kwalitatief hoogwaardige inhoudelijke kennis. Over deze punten valt dan ook weinig te vermelden in het kader van portfolio's.

Transparantie/betekenisvolheid
Studenten moeten goed op de hoogte zijn van de criteria en standaarden waaraan de inhoud moet voldoen (Dochy & Sluijsmans, 1998). Het hebben van duidelijke en complete performance criteria is een kritische voorwaarde voor studenten om te kunnen selecteren welk materiaal in het portfolio gaat, alsook voor het aansturen van de reflecties (Brookhart, 1999). De criteria en standaarden kunnen volledig onafhankelijk door de docent worden vastgelegd, maar evenzeer het resultaat zijn van gezamenlijk overleg tussen docenten en studenten. In ieder geval dienen zij zo duidelijk mogelijk te worden gecommuniceerd naar de studenten. Hetzelfde geldt voor de *functie* van de portfolio assessment. Het moet volkomen helder zijn voor studenten en docenten of het gaat om een beoordelende functie of een formatieve functie. Van groot belang is het om portfolio-functies duidelijk van elkaar te onderscheiden (Evelein en Van Tartwijk, 2000) Een portfolio waarop beoordeeld zal worden, ziet er anders uit dan een die bedoeld is voor de eigen professionele ontwikkeling. Verschillende doelen nastreven met één portfolio levert spanning op (het is voor de student riskant om zich kritisch over zichzelf uit te laten als men daarop wordt beoordeeld).

Efficiëntie en kosten
Een risico bij gebruik van portfolio assessment ontstaat doordat studenten soms heel wat tijd kwijt zijn aan het produceren, verzamelen en ordenen van de samenstellende items, evenals de herhaalde schriftelijke zelfreflecties of verklarende rapporten. Zo kan over-assessment ontstaan waarbij studenten vooral bezig zijn met productieactiviteiten ten koste van leeractiviteiten. Tegelijkertijd ligt bij de docent/begeleider de verplichting om de items in de portfolio herhaald en tijdig te beoordelen en de noodzakelijke informatieve feedback te verschaffen. Zonder regelmatige en gedifferentieerde feedback biedt het samenstellen van portfolio's voor de student weinig ondersteuning van hun leerproces. De doeltreffendheid en efficiëntie van zowel het samenstellen als het bespreken van de portfolio kan aanzienlijk verhoogd worden door zich bewust te beperken tot een reeks kerncompetenties of essentiële elementen van het gerealiseerde product.

Samengevat
Bij doordachte integratie van portfolio assessment in het instructieproces, is sprake van een prima instrument voor het opvolgen en sturen van de ontwikkeling van het leerproces van de individuele student. Hieraan zijn een aantal voorwaarden verbonden zoals heldere criteria, een beperkt aantal essentiële elementen die in het portfolio naar voren moeten komen, een eenduidig en helder doel, en open communicatie hierover tussen studenten en docenten. Bij gebruik voor summatieve doeleinden zijn deze voorwaarden nog belangrijker, en zal ook de exacte inhoud van het portfolio expliciet gemaakt moeten worden. De beoordelingsprocedures moeten dan bovendien verregaand gestandaardiseerd worden. Een eerlijke zelfreflectie zal door summatieve doeleinden tegengewerkt worden. Bij formatieve doeleinden wordt zelfreflectie en zelfverantwoordelijkheid gestimuleerd. Een mogelijk nadeel

van het werken met portfolio's is dat er nog weinig onderzoek, richtlijnen en ervaringen voorhanden zijn, waardoor de nodige experimenteerruimte gecreëerd moet worden.

4.2.6. Self-, Peer-, en Co-assessment

Consequenties

Door self-, peer- en co-assessment wordt de betrokkenheid en zelfverantwoordelijkheid van de student ten aanzien van het eigen leerproces verhoogd. Dit is met name het geval als de studenten ook betrokken zijn geweest bij het vaststellen van de doelstellingen en beoordelingscriteria. De mate van zelfsturing en motivatie van studenten kan verhoogd worden door deze betrokkenheid (Nedermeyer & Pilot, 2000). Een andere, essentiële, consequentie is dat studenten meer inzicht krijgen in de gewenste leerresultaten en hun eigen prestaties waardoor ze hun eigen leerproces en ontwikkeling beter bij kunnen sturen.

Verder vormt het involveren van studenten bij het beoordelingsproces op zichzelf een belangrijke leerervaring, omdat het bijdraagt aan het verwerven van de complexe vaardigheden zelfreflectie, beoordelen van de prestaties van anderen en het geven van feedback. Om deze effecten daadwerkelijk te bereiken is wel expliciete training en ondersteuning op deze vaardigheden nodig.

Rechtvaardigheid

Bij self-assessment blijkt dat goede studenten zichzelf lager beoordelen dan de docenten zouden doen, terwijl de minder goede studenten juist optimistischer over zichzelf zijn (Nedermeyer & Pilot, 2000). Verder hebben studenten de neiging om zichzelf in het begin van het studiejaar hoger te beoordelen dan later in het studiejaar. Indien de docent feedback geeft op de beoordelingen, verbetert de self-assessment doordat de studenten meer inzicht krijgen in de eisen die aan hen gesteld worden. Bij peer-assessment bestaat het risico dat men vrienden gaat bevoordelen, dat dominante groepsleden hogere cijfers krijgen of dat 'meelopers' toch een goed cijfer krijgen.

Een kritische voorwaarde voor het verkrijgen van betrouwbare informatie is het gebruik maken van een beoordelingsinstrument, zodat de feedback systematisch kan worden gegeven en besproken. De items moeten daarbij duidelijk en begrijpelijk zijn geformuleerd (Jellema, 2000).

Generaliseerbaarheid en transfer

Dit punt is niet van toepassing hier, omdat generaliseerbaarheid en transfer afhangen van de methode (bijvoorbeeld de vragen, taken, simulaties of oefeningen), en niet van de personen die bij de beoordeling betrokken zijn.

Cognitieve complexiteit, Inhoudskwaliteit en Inhoudsrepresentatie

Voor deze drie punten geldt dezelfde opmerking: deze kwaliteitsaspecten hangen samen met de methode van de assessment, en niet met de personen die bij de beoordeling betrokken zijn. Wel kan het lastig zijn om studenten cognitief complexe zaken met hoge inhoudelijke kwaliteit te laten beoordelen wanneer zij zelf nog in het leerproces verkeren. Vaardigheden en gedragsaspecten zijn een meer voor de hand liggend en veelvoorkomend object van self-, peer en co-assessment.

Transparantie/betekenisvolheid

Het succes van self-, peer- en co-assessment is in belangrijke mate afhankelijk van de duidelijkheid van en aanvaarding door docent en student van de gezamenlijk bepaalde criteria. Een open dialoog tussen docent en student of studenten onderling is daarbij

noodzakelijk. Er dienen duidelijke afspraken te zijn over hoe de resultaten gebruikt zullen worden, wat de waarde ervan is.

Het zichzelf of elkaar beoordelen wordt gemakkelijk als bedreigend ervaren. Studenten zijn niet gewend zichzelf of hun vrienden te beoordelen en vrezen vooral in de beginsituatie unfair te worden beoordeeld of te beoordelen. Door adequate training kan dit worden verbeterd (Williams, 1992, in Sluijsmans, e.a., 2000). Vanwege de genoemde redenen wordt het in het algemeen afgeraden om self-, peer- en co-assessment te gebruiken in het kader van een afsluitende (summatieve) beoordeling. Hetzelfde geldt voor het gebruiken van 360-graden feedback in het kader van een formele personeelsbeoordeling. Dit is wel mogelijk, maar wordt niet aanbevolen omdat dan meestal een acceptatieprobleem speelt (Jellema, 2000).

Efficiëntie en kosten

Het ontwikkelen van goede beoordelingsinstrumenten, heldere criteria, open communicatie, alsook het trainen van de (onervaren) studenten zal een behoorlijke tijdsinvestering zijn. Als zij hierbij betrokken worden kunnen studenten echter essentiële dingen leren tijdens deze voorbereidende processen.

Samengevat

Self-, peer- en co-assessment zijn geëigende methodes om studenten te ondersteunen bij het verwerven van zelfreflectie- en beoordelingsvaardigheden. Hier is echter wel adequate training bij nodig. Er is een duidelijke voorkeur voor het inzetten van self-, peer- en co-assessment voor formatieve doeleinden. Ze geven (evenals 360-graden feedback) een goed diagnostisch inzicht in het eigen functioneren, en dit biedt een goed kader voor het zelf verder ontwikkelen en aansturen van het eigen leerproces. De betrokkenheid bij assessment bevordert bovendien de motivatie en zelfverantwoordelijkheid voor het leren. Potentiële nadelen zijn dat bij pure formatieve evaluatie het belang afneemt waardoor studenten oppervlakkiger kunnen worden, dat er bedreigende situaties kunnen ontstaan voor studenten, en dat geslaagde self-, peer- en co-assessment afhangen van de voorbereiding en investering die de nodige tijd en inspanning kosten.

Hoofdstuk 5: Beslissingsschema's voor een beargumenteerde keuze van een geschikte toetsvorm.

5.1. Toelichting bij het beslissingsschema

De keuze van een geschikte toetsvorm evenals het concipiëren van een geschikt toetsplan vormt een belangrijk aspect van een onderwijsvernieuwingsproces, in die zin dat er voldoende rekening moet worden gehouden met de impact van het toetsgebeuren op het leerproces van de studenten.
Verschillende auteurs (Brown, 1999; Nedermeyer & Pilot, 2000) hebben in dit verband een aantal richtlijnen of adviezen geformuleerd waarbij wordt uitgegaan van de vijf vragen van Harden (1987) die in dit rapport reeds aan de orde waren in paragraaf 3, evenals de kwaliteitscriteria validiteit en betrouwbaarheid.

'Waarom toets men' is de eerste vraag
In de hierboven vermelde adviezen wordt echter hoofdzakelijk uitgegaan van de wat- en de hoe-vraag, om vervolgens de merites qua validiteit, betrouwbaarheid en bruikbaarheid van de gekozen of geadviseerde toetsvorm de bespreken.
Hier ontbreekt naar ons oordeel een belangrijke voorafgaande vraag, namelijk. de waarom-vraag. Zoals in paragraaf 3 werd vermeld gaat het hierbij om de vraag of men toetst voor 'formatieve' of 'summatieve' doeleinden.
Formatieve toetsing heeft als doel het ondersteunen van het leerproces via informatieve feedback. Essentieel hierbij is de integratie van de toetsing in het onderwijsleerproces.
Summatieve toetsing is gericht op beoordeling (het geven van cijfers en certificaten) en heeft doorgaans verstrekkende gevolgen voor de student. Hier gelden dan ook strenge criteria qua validiteit en generaliseerbaarheid van de resultaten.
Samenvattend, de criteria of specifieke argumenten voor de keuze van een bepaalde toetsvorm verschillen in functie van de rol of functie die men aan de toetsing toebedeeld in het onderwijsleerproces. Het is dan ook beslist zaak om van tevoren scherp te stellen met welk doel men gaat toetsen. Brown e.a. (1997) merken hierover op: 'A common error is to use an assessment task for one set of purposes and then assume that the results from it are appropriate for other purposes'. Dezelfde toets moet dus nooit twee doelen dienen.
Op basis van deze argumentatie menen wij dan ook de 'waarom-vraag' als eerste stap te moeten definiëren in het hier voorgestelde beslissingsmodel en een afzonderlijke beslissingstabel te moeten opstellen voor formatieve en summatieve toetsing. Binnen ieder van deze tabellen vindt u dan de nodige aanwijzingen om de eigen keuze binnen een summatieve of formatieve context te beargumenteren.

'Wat en Hoe wil men toetsen' vormen de tweede en derde vraag
Als eenmaal de waarom-vraag is beantwoord, dient expliciet te worden gemaakt welke leerdoelen en leerinhouden men wil meten, de wat- en de hoe-vraag. Het beantwoorden van de wat-vraag betekent een keuze voor een classificatiesysteem van leerdoelen. Voor een toelichting op de hierna gehanteerde doelstellingencategorieën wordt terug verwezen naar paragraaf 3.1.1.
Op dezelfde wijze wordt voor het antwoord op de hoe-vraag, - welke toetsvorm(en) zijn voor het toetsen van een of meerdere leerdoelen geschikt – verwezen naar paragraaf 1.4. waar de verschillende toetsvormen zijn besproken.

Tenslotte, in de hierna volgende beslissingstabellen wordt de geschiktheid van een bepaalde toetsvorm (kolommen) voor het meten van een of meerdere leerdoelen (rijen) geëvalueerd of beargumenteerd op basis van de criteria van Linn, Baker & Dunbar (1991).

Relatie tussen de 'waarom-vraag' en 'wat- en hoe-vraag'
De betekenis en het belang van de criteria van Linn, Baker & Dunbar (1991) zijn verschillend naar gelang de doelstellingen of functie van de toetsing in het onderwijsleerproces (formatief of summatief).
Zo zullen wij ons in het kader van formatieve toetsing, van de 8 vermelde criteria van Linn, e.a., (1991), alleen de criteria: (1) consequenties, (2) betekenisvolheid en (3) efficiëntie hanteren.
In het kader van summatieve toetsing zijn dan weer de criteria: (1) rechtvaardigheid, (2) generaliseerbaarheid/transfer, (3) inhoudsrepresentativiteit, (4) transparantie en (5) efficiëntie en kosten van belang.
De mate waarin elk van deze criteria een rol speelt in de overwegingen, leerdoel-toetsvorm en valable argument aandraagt voor of tegen de keuze van deze combinatie wordt in de afzonderlijke cellen van de beslissingstabellen met een, twee of drie plus- of mintekens aangegeven.

Argumentatie voor de toewijzing van kwaliteitscriteria aan summatieve of formatieve doeleinden
Er is gekozen om het criterium van transparantie/betekenisvolheid in tweeën te splitsen. Bij summatieve toetsing is vooral transparantie van belang omdat studenten zich maximaal dienen te kunnen voorbereiden, wat van belang is omwille van de belangrijke formele consequenties. Betekenisvolheid van de toetsopdrachten is vooral van belang in het kader van formatieve toetsing omdat hier de 'leerzaamheid' een grote rol speelt.

De criteria inhoudskwaliteit en cognitieve complexiteit worden buiten beschouwing gelaten omdat zij niet echt samenhangen met de keuze van de toetsvorm, maar eerder met de zorgvuldigheid die de ontwerpers en inhoudsdeskundigen aan de dag leggen bij het creëren van toetsopdrachten (inhoudskwaliteit), dan wel met de keuze om al of geen processen te meten naast producten.

Consequenties vormen een belangrijk criterium als er formatief getoetst wordt. In formatieve toetsing gaat het om het ondersteunen van het onderwijsleerproces. In die zin is de mate waarin de toetsing en interpretatie van de toetsresultaten bijdraagt aan het ondersteunen en sturen van dit onderwijsleerproces (consequential validity) van essentieel belang.

In het kader van summatieve toetsing speelt rechtvaardigheid een belangrijke rol. Omwille van de formele consequenties is het extra belangrijk om te vermijden dat studenten ten onrechte worden afgewezen of vertraagd.

Ook generaliseerbaarheid en transfer zijn belangrijke criteria in het kader van summatieve toetsing. Een opleiding wil tenslotte kunnen verantwoorden dat zijn studenten zijn voorbereid op de (brede) beroepspraktijk. Generaliseerbaarheid en transfer van de gemeten prestaties naar volgende taken is minder belangrijk in het kader van het opvolgen en sturen van een ontwikkelingsproces (formatieve toetsing) waarbij de toetsing veelal gericht is op specifieke dimensies of deelvaardigheden van dit leer- en ontwikkelingsproces.

Inhoudsrepresentativiteit speelt vooral een rol bij summatief toetsen. Bij formatief toetsen vormt dit meestal geen probleem aangezien de toetsing zich veelal zal beperken tot het aftoetsen van een beperkt inhouds- of gedragsdomein. Hierbij moet echter onmiddellijk

worden gewaarschuwd voor een te modulaire benadering van formatieve toetsing. Dit leidt na bepaalde tijd tot verkokering en gebrekkige integratie tussen de deelvaardigheden.

Tenslotte speelt het criterium van efficiëntie en kosten een rol bij summatieve toetsing, net zo goed als bij formatieve toetsing. De afweging gaat hier in de richting van de vraag of de gehanteerde toetsvorm ook de nodige winst oplevert op het gebied van goed en rechtvaardig kunnen beoordelen van studenten.

Samengevat: Wat zit er nu allemaal in het beslismodel?
Waar in het algemeen de 'wat-' en 'hoe-vraag' de meest voor de hand liggende vragen zijn bij de keuze van een geschikte toetsvorm, zijn wij van oordeel dat de te stellen criteria ten aanzien van de toetspraktijk in belangrijke mate mede bepaald wordt door de 'waarom-vraag', het al of niet formatief of summatief gebruik van de toetsresultaten. Met andere woorden de geschiktheid van een toetsvorm wordt mede bepaald door het gebruik van de toetsresultaten.
Bekijken we volgend voorbeeld eens: een mc-toets (objectieve toetsing) is geschikt voor het meten van gememoriseerde kennis (dit is de aansluiting tussen 'hoe' en 'wat'). Het doel waarmee je toetst is echter zeer relevant: als het om een summatieve toets gaat is het criterium 'rechtvaardigheid' zeer belangrijk. De vraag wordt dan tweeledig, bijvoorbeeld: Is de mc-toets geschikt voor het meten van gememoriseerde kennis, en biedt de mc-toets grote kans om dit op een rechtvaardige wijze te meten? Bij formatieve toetsing zou de vraag bijvoorbeeld zijn: is de mc-toets geschikt voor het meten van gememoriseerde kennis en biedt de mc-toets grote kans om betekenisvol en sturend te zijn voor studenten ?

Beslissingsschema's

In de hierna volgende beslissingstabellen worden de verschillende toetsvormen (rijen) beoordeeld ten aanzien van hun geschiktheid voor het meten van bepaalde leerdoelen (kolommen). Die geschiktheid wordt uitgedrukt met een, twee of drie plus- of mintekens, in de overeenkomstige cellen, voor de hoger vermelde relevante criteria van Linn e.a., (1991). Het aantal plus- of mintekens kan dan gezien worden als een graadmeter voor het al of niet geschikt zijn van deze toetsvorm voor het meten van bepaalde leerdoelen.

Nadere inspectie van de aldus bekomen resultaten wijzen er ons onmiddellijk op dat de ideale toetsvorm voor een welbepaalde doelstelling niet bestaat. Meestal zijn meerdere toetsvormen toepasselijk. Zo constateren we dat performance assessment, ADC's en portfolio assessment nagenoeg gelijkwaardig scoren voor het meten van cognitieve vaardigheden, gedragsvaardigheden en competenties, terwijl zij anderzijds niet van toepassing zijn voor het louter meten van kennis. De keuze voor een welbepaalde toetsvorm of een mix van toetsvormen zal in die omstandigheid dan ook mede bepaald worden door andere dan de hier vermelde factoren. Een belangrijke factor vormt hier bijvoorbeeld de overeenkomst tussen toetsvorm en leeractiviteiten (authentic assessment).

Een andere opvallende vaststelling is de verschillende beoordeling van klassieke objectieve toetsen voor het toetsen van kennis en cognitieve vaardigheden in het kader van formatieve versus summatieve toetsing. Hier speelt vooral de 'consequential validity' de invloed van deze toetsvorm op het leergedrag een belangrijke rol. Deze vaststelling strookt niet met de opvatting dat deze toetsvorm vooral geschikt is voor continue toetsing (objectieve toetsen met beperkte afname en verwerkingstijd) in de loop van het leerproces.

Anderzijds is het gebruik van klassieke mc-toetsen niet van toepassing voor het toetsen van gedragsvaardigheden, attitudes en competenties.

Verder willen we opmerken dat self-, peer- en co-assessment, die niet direct een vorm van toetsing vertegenwoordigen maar een wijze van beoordelen, voornamelijk toepasselijk zijn bij het beoordelen van gedragsvaardigheden, attitudes en competenties. De bijdrage die deze vorm van beoordelen levert aan het toetsgebeuren is voornamelijk afhankelijk van de toepasbaarheid van deze vorm van beoordelen bij de voor deze doelstellingen geschikte toetsvormen.
Tenslotte, een belangrijke vraag ten aanzien van de bruikbaarheid van deze beslissingstabellen is: in welke mate wordt door een degelijk gedifferentieerde beoordeling (gebruik van verschillende criteria) van de combinatie doelstellingen-toetsvorm, de intuïtieve keuze van een docent voor een bepaalde combinatie al of niet bevestigd of tegengesproken?

5.2.1. Beslissingsschema voor formatieve toetsing

Relevante criteria:
1. Consequenties
2. Betekenisvolheid
3. Efficiëntie en kosten

Doelstellingen → Toetsvormen ↓	Kennis	Cognitieve vaardigheden	Gedragsvaar-digheden	Attitudes	Competen-ties
Klassieke toetsen met gesloten vragen	1 --- 2 +/- 3 +++	1 - 2 - 3 +	1 nvt 2 nvt 3 nvt	1 nvt 2 nvt 3 nvt	1 nvt 2 nvt 3 nvt
Klassieke toetsen met open vragen	1 ++ 2 ++ 3 ++	1 ++ 2 + 3 ++	1 + 2 - 3 +	1 ++ 2 ++ 3 ++	1 - 2 +/- 3 +/-
Performance assessment	1 nvt 2 nvt 3 nvt	1 +++ 2 +++ 3 ++	1 +++ 2 +++ 3 ++	1 +/- 2 + 3 -	1 +++ 2 +++ 3 ++
Assessment en Development Centers	1 nvt 2 nvt 3 nvt	1 +++ 2 +++ 3 +	1 +++ 2 +++ 3 +	1 ++ 2 + 3 +	1 +++ 2 +++ 3 ++
Portfolio assessment	1 - 2 -- 3 --	1 ++ 2 ++ 3 +	1 +++ 2 ++ 3 ++	1 ++ 2 ++ 3 ++	1 +++ 2 +++ 3 ++
Self-, Peer-, en Co-assessment	1 nvt 2 nvt 3 nvt	1 ++ 2 + 3 ++	1 ++ 2 ++ 3 ++	1 +++ 2 ++ 3 ++	1 ++ 2 ++ 3 ++

5.2.2. Beslissingsschema voor summatieve toetsing

Relevante criteria:

1. Rechtvaardigheid
2. Generaliseerbaarheid en transfer
3. Transparantie
4. Inhoudsrepresentativiteit
5. Efficiëntie en kosten

Doelstellingen → Toetsvormen ↓	Kennis	Cognitieve vaardigheden	Gedragsvaar-digheden	Attitudes	Competen-ties
Klassieke toetsen met gesloten vragen	1 +++ 2 +++ 3 +++ 4 +++ 5 +++	1 ++ 2 + 3 ++ 4 + 5 ++	1 nvt 2 nvt 3 nvt 4 nvt 5 nvt	1 nvt 2 nvt 3 nvt 4 nvt 5 nvt	1 nvt 2 nvt 3 nvt 4 nvt 5 nvyt
Klassieke toetsen met open vragen	1 ++ 2 ++ 3 + 4 + 5 +	1 ++ 2 ++ 3 + 4 + 5 ++	1 -- 2 -- 3 +/- 4 --- 5 ---	1 ++ 2 ++ 3 ++ 4 + 5 +	1 nvt 2 nvt 3 nvt 4 nvt 5 nvt
Performance assessment	1 nvt 2 nvt 3 nvt 4 nvt 5 nvt	1 ++ 2 ++ 3 ++ 4 + 5 +	1 +++ 2 +++ 3 +++ 4 +++ 5 +	1 + 2 + 3 + 4 + 5 +	1 +++ 2 +++ 3 +++ 4 + 5 ++
Assessment en Development Centers	1 nvt 2 nvt 3 nvt 4 nvt 5 nvt	1 ++ 2 ++ 3 + 4 + 5 +	1 + 2 +++ 3 +++ 4 ++ 5 +	1 ++ 2 ++ 3 + 4 + 5 +	1 +++ 2 +++ 3 +++ 4 ++ 5 ++
Portfolio assessment	1 nvt 2 nvt 3 nvt 4 nvt 5 nvt	1 ++ 2 ++ 3 + 4 +/- 5 -	1 +++ 2 +++ 3 ++ 4 ++ 5 +	1 +++ 2 +++ 3 ++ 4 +++ 5 +	1 +++ 2 +++ 3 +++ 4 ++ 5 +
Self-, Peer-, en Co-assessment	1 nvt 2 nvt 3 nvt 4 nvt 5 nvt	1 + 2 + 3 +/- 4 + 5 +	1 ++ 2 ++ 3 ++ 4 ++ 5 ++	1 ++ 2 ++ 3 ++ 4 + 5 +	1 ++ 2 ++ 3 ++ 4 ++ 5 ++

5.3. Kanttekening bij het beslissingsmodel

Deze tabellen vormen een hulpmiddel voor het beantwoorden van de ‘hoe-vraag’ of het bepalen van wat een geschikte toetsvorm is. Deze keuze wordt uiteindelijk bepaald op basis van het antwoord op achtereenvolgens de waarom- en de wat-vraag evenals de wezenlijk geachte kwaliteitscriteria of argumenten. ‘Wanneer’ er getoetst wordt is niet expliciet opgenomen in het beslissingsmodel In de praktijk zien we echter een nauwe samenhang tussen het antwoord op de wanneer-vraag en de waarom-vraag. Continue toetsing speelt bijvoorbeeld voornamelijk een rol bij formatieve toetsing, in het kader van continue bijsturing en ontwikkeling van het leerproces. Summatieve toetsing vindt voornamelijk plaats als afsluiting van het leerproces, zij het dat ook daarbij veelal gebruik gemaakt wordt van tussentijdse resultaten.

De vraag ‘wie’ beoordeelt kan bepaald worden op basis van overwegingen vanuit zowel de waarom-vraag als de hoe-vraag. Zo is bij formatieve toetsing de beoordeling eerder een gemeenschappelijke zaak van docent en student, terwijl bij summatieve toetsing het toch eerder de docent is die het eindoordeel vastlegt. Anderzijds sluit de wie-vraag ook aan bij de hoe-vraag, meer bepaald de deelvraag naar de wijze van scoring bij open vragen. Zo zullen, in het kader van alternatieve toetsing, docenten en studenten samen de scoringsrubrieken en beoordelingscriteria bepalen.

Een laatste opmerking betreft de verankering van de toetsvorm voor één onderdeel in het totale toetssysteem en het overkoepelende onderwijsmodel. Een concreet voorbeeld hiervan vinden we beschreven in ‘Toetsing in probleemgestuurd Onderwijs’ van van der Vleuten & Driessen (2000). De auteurs beschrijven in deze publicatie het ‘geïntegreerd toetsplan’ voor de eigen vorm van probleemgestuurd onderwijs zoals die in het Maastrichtse model is vorm gegeven. De mix van gehanteerde toetsvormen wordt daarbij uitdrukkelijk beargumenteerd zowel op basis van de ‘wat-vraag’ als op basis van de ‘alignment’ met het eigen onderwijsmodel.
Dit betreft een curriculair toetsplan, waarbij verschillende competenties over de afzonderlijke leereenheden heen worden getoetst. Deze verschillende competenties vragen verschillende toetsvormen, geïntegreerd in een weloverwogen toetsplan. Ofschoon dit laatste aspect, namelijk het curriculair opzetten van een geïntegreerd toetsplan zeker een belangrijk aspect vormt en de keuze van de toetsvorm mede bepaald, hebben wij in ons voorstel van beslissingschema ons bewust beperkt tot de keuze van toetsvormen in het kader van afzonderlijke leereenheden.

Naschrift

Voorafgaand aan de uitgifte van dit rapport werd de voorliggende tekst ter beoordeling voorgelegd aan enkele toetsdeskundigen, met name mevrouw Monique Doorten (OTEC/OUNL), Prof. Dr. C. van der Vleuten (UNIMAAS, Maastricht) en Drs. Klaas Eringa (CHN, Leeuwarden). Daarbij werden hen een viertal vragen voorgelegd. Van mevrouw Doorten ontvingen wij een uitgebreide schriftelijke reactie op de door ons gestelde vragen; door Prof. van der Vleuten en Drs. Klaas Eringa werden we ontvangen voor een boeiende bespreking. Aan de bespreking met Drs. Klaas Eringa werd bovendien deelgenomen door de collega's Richard Zwaal, Lonneke Hofstede en Frans Winters, allen verbonden aan de CHN en begaan met het toetsgebeuren aan de CHN.
Hun commentaar op het rapport en hun antwoorden op de gestelde vragen worden hierna samengevat. Bij deze willen we deze deskundige oprecht bedanken voor hun gewaardeerde opmerkingen en aanwijzingen voor bijsturing van deze poging tot het opstellen van een beslissingschema ter ondersteuning van docenten en ontwerpers van onderwijs bij de keuze van een geschikte toetsvorm.

Onderscheid formatieve en summatieve toetsing

Vraag 1:

In het rapport wordt bijzonder veel belang gehecht aan het onderscheid tussen formatieve en summatieve toetsing. Voor ons is dit de eerste vraag die beantwoord moet worden bij de keuze voor een geschikte toetsvorm. Wat is uw mening hierover?

Alle beoordelaars erkennen het onderscheid tussen formatieve en summatieve toetsing. De mate waarin dit de te stellen kwaliteitscriteria bepaald wordt, wordt echter verschillend ingeschat.

Zo wijst **Doorten** er op dat dit onderscheid wel van invloed is op de vormgeving van de toets maar is niet bepalend voor de vorm van de toets; de vorm van de toets wordt op de eerste plaats bepaald door wat er wordt getoetst, de beoogde doelstellingen. De vraag 'waarom' er wordt getoetst moet zeker beantwoord worden, maar niet noodzakelijk als eerste. Verder is mevrouw Doorten het niet eens met de stelling (rapport p. 38) dat een toets maar één doel zou kunnen hebben. De geschiktheid van een en dezelfde toets voor zowel formatieve als summatieve toetsing heeft wel consequenties naar de vormgeving toe. Bovendien heeft het formatief of summatief gebruik van een toets meer te maken met het gebruik van de toetsinformatie in het verdere leerproces, de mate van feedback.

Een zelfde visie vinden we terug in de commentaar van **Prof. van der Vleuten**. Ook hij benadrukt het onderscheid tussen formatieve en summatieve toetsing en het daarmee samenhangend verschillend wegen van het belang van de onderscheiden kwaliteitscriteria. Maar tegelijk wijst hij er ook op dat het de kunst is deze twee vormen van toetsen in de concrete onderwijscontext zo goed mogelijk bij elkaar te brengen en te integreren. De praktijk leert namelijk dat puur formatieve evaluaties door de studenten vaak niet serieus worden genomen, terwijl puur summatieve evaluaties te weinig leerwaarden hebben.

Deze verschillende doelstelling (formatief/summatief) bij toetsing leidt er aldus onvermijdelijk toe dat verschillende kwaliteitseisen worden gesteld of andere accenten worden gelegd bij de keuze van een toetsvorm. Feit is echter dat men niet alle criteria in iedere situatie kan realiseren; en bovendien zijn bepaalde van de in het rapport vermelde criteria in bepaalde contexten met elkaar in concurrentie. Welke criteria men dan dient na te streven is niet zo belangrijk.

- Belangrijk is dat er duidelijke criteria aanwezig zijn en dat er een afweging gebeurt op het moment dat er een concrete keuze wordt gemaakt. Daarbij worden die criteria niet zozeer gezien als kenmerken van een toetsvorm maar eerder als na te streven kwaliteitseisen van iedere toetsvorm in iedere context.
- Belangrijk is ook te beseffen dat men niet steeds aan alle belangrijk geachte criteria kan voldoen, en bijgevolg naargelang de context compromissen dient te sluiten. (Welke hier de belangrijke criteria zijn komt verder aan bod).
- Het realiseren van de gestelde criteria wordt bevorderd door een mix van toetsvormen samen te stellen in functie van de beoogde competenties.

Conclusie aldus van der Vleuten: het onderscheid tussen formatieve en summatieve toetsing is belangrijk, al is het, in een onderwijscontext, de kunst beide vormen van toetsing zo goed mogelijk met elkaar te integreren. Op dit ogenblik geven we nog te veel aandacht aan summatieve toetsing en onvoldoende aan de formatieve functie van toetsing.

Bij de **CHN** wenst men zo geen strikt onderscheid te maken tussen formatieve en summatieve toetsing. Het gaat hier om een theoretisch onderscheid, terwijl in de praktijk iedere summatieve toetsing een formatief aspect heeft en omgekeerd. Opdrachten tijdens het jaar tellen steeds mee in de afsluitende beoordeling, en waar dit niet het geval is worden deze toetsen niet ernstig genomen door de studenten. Maar ook docenten nemen dit soort opdrachten of toetsen niet ernstig, omdat ze teveel tijd vragen. Formatieve toetsing heeft nog niet echt een plaats verworven in het onderwijs. De toets heeft maar invloed op het gedrag van de studenten als men ze laat meetellen in de eindbeoordeling. Summatieve toetsen hebben in die zin een sterker sturend karakter dan formatieve toetsen.

Ervaring is wel dat opdrachten, cases, aan de praktijk ontleende kernproblemen in de loop van de studie een sterk sturend karakter hebben. Hier hebben we een goede integratie tussen onderwijsactiviteiten en toetsing en hier maakt men geen onderscheid tussen formatieve en summatieve toetsing.

Het onderscheid tussen formatieve en summatieve toetsing is dus te theoretisch, terwijl beide vormen of functies van toetsing in de praktijk sterk met elkaar geïntegreerd zijn. Bovendien maken docenten in de praktijk dit onderscheid niet. Normaliter zullen docenten met een dergelijk instrument, bestaande uit twee tabellen dan ook moeilijk uit de voeten kunnen. Zij vragen zich niet af of ze nu formatief of summatief toetsen. Zij gebruiken ook die termen niet; zij spreken eerder van oefentoetsen en eindtoetsen.

Vraag 2:

In het rapport worden acht kwaliteitscriteria besproken als basis voor het maken van een beargumenteerde keuze. Wij vinden dat afhankelijk van de functie van de toets (formatief versus summatief) andere kwaliteitscriteria van toepassing zijn.
Vind u de indeling in acht criteria voldoende ruim, helder en specifiek? Vindt u de wijze waarop de criteria zijn toebedeeld aan resp. formatieve en summatieve evaluatie verantwoord? Ontbreken er bepaalde criteria naar uw mening?

Alle beoordelaars formuleren heel wat bedenkingen bij enerzijds de in het rapport gehanteerde inhoudelijke omschrijving van de acht criteria van Linn en anderzijds de wijze waarop deze criteria verschillend gebruikt worden als basis voor de keuze van een geschikte toetsvorm voor formatieve of summatieve toetsing.

Voor **Prof. van der Vleuten** zijn de gekozen criteria niet zo bepalend. Belangrijker is dat er eenduidige criteria zijn, en dat er bij de keuze voor een bepaalde toetsvorm, een afweging gebeurt ten aanzien van de mate waarin men deze kwaliteitscriteria in die context en bij deze toetsvorm wil realiseren, of hoe aan deze criteria kan worden voldaan. Men zal in de praktijk nooit aan alle vooropgestelde criteria kunnen voldoen. Daarom zal men in functie van de context (formatieve of summatieve toetsing) steeds een compromis moeten maken over welk belang men hecht aan bepaalde criteria. Waar het om gaat is dat er criteria zijn en dat er een afweging gebeurt op het moment dat er een concrete keuze wordt gemaakt.

De keuze voor een bepaalde toetsvorm gebeurt niet alleen op basis van deze kwaliteitscriteria. Hier spelen tevens een heleboel contextuele gegevens een rol. Een belangrijk criterium is hier de ‘acceptabiliteit’, een criterium dat niet in het rapport vermeld wordt. Acceptabiliteit, omdat de keuze van een toetsvorm geen rationeel gebeuren is, maar veeleer een praktische beslissing, gebaseerd op enerzijds de ‘visie en ervaringen van docenten’ over wat goede toetsvormen zijn voor het toetsen van hun doelstellingen en anderzijds kosten en uitvoerbaarheid. In een opleiding brengt iedere docent een zekere ervaring en deskundigheid mee en zal in functie daarvan de voorkeur geven aan bepaalde toetsvormen. De vraag is dan niet of dit een geschikte toetsvorm is, maar wel ‘wat kunnen we doen om met die toetsvorm die criteria zo optimaal mogelijk te realiseren. Toetsen is dus geen psychometrisch probleem maar een onderwijskundig. En als onderwijskundige hecht Prof. van der Vleuten bijzonder veel aandacht aan een tweede criterium, ‘consequential validity’, de invloed van toetsen op het leer- en doceergedrag. Op dit punt wenst hij zo weinig mogelijk toegevingen te doen.

Wat nu die acht kwaliteitscriteria van Linn betreft, deze zijn niet zo overtuigend en bovendien in het rapport niet altijd even helder omschreven. De meeste van die criteria zijn in feite aspecten van het begrip validiteit of betrouwbaarheid. Verder blijkt uit onderzoek dat bepaalde van deze criteria in de praktijk met elkaar in concurrentie zijn. Bijvoorbeeld volledige transparantie qua te toetsen gedrag en scoringsregels kan ook een negatieve invloed hebben op het studeergedrag. Zo bleek in de geneeskunde dat bij een te ver doorgedreven beschrijving van gewenst of te beoordelen gedrag de studenten zich sterk gingen richten op het leren van dat toetsgedrag. Een ander voorbeeld houdt verband met ‘cognitieve complexiteit’. Te complexe toetstaken kunnen bij beginnende studenten juist aanleiding zijn tot gebrekkige validiteit, wegens gebrekkige afstemming op het niveau van de student. Bovendien is uit onderzoek gebleken dat die complexiteit van de toetssituatie soms weinig informatie toevoegt aan de meting, terwijl de meting zelf en de scoring heel wat extra werk vragen. Met andere woorden houd het zo simpel mogelijk.

Samengevat, de vijf criteria die door Prof. van der Vleuten naar voren worden geschoven zijn validiteit, betrouwbaarheid, efficiëntie en kosten, en vooral ‘consequential validity’ en ‘acceptability’; maar nogmaals het gaat niet zozeer om het feit of dit ‘de’ criteria zijn die men poogt de realiseren, maar wel dat er eenduidige criteria zijn, en dat er bij de keuze voor een bepaalde toetsvorm, een afweging gebeurt ten aanzien van de mate waarin men deze kwaliteitscriteria in die context en bij deze toetsvorm wil realiseren, of hoe aan deze criteria kan worden voldaan. Anderzijds, het wegen van bepaalde criteria in functie van de context ‘formatieve of summatieve toetsing’ is aanvaardbaar, maar mag ook weer niet worden overdreven omwille van het wenselijk samengaan van beide toetsfuncties. Bovendien het te aanvaarden compromis qua criteria voor een bepaalde toetsvorm wordt ook nog eens

beïnvloed door het integrale toetsprogramma binnen een opleiding. De mix van toetsvormen zal bepaalde tekorten compenseren. Belangrijker dan het handhaven van kwaliteitscriteria voor afzonderlijke toetsvormen is het optimaliseren van het hele toetsgebeuren; het opstellen van een toetsplan waarbij de mix van toetsvormen voldoen aan de vooropgestelde kwaliteitscriteria.

Ook mevrouw **Doorten** is het niet eens met de stelling dat het onderscheid tussen formatieve en summatieve toetsing noodzakelijk leidt tot het hanteren van verschillende kwaliteitseisen. Voor elke toets gelden kwaliteitseisen, ook voor formatieve toetsen. Natuurlijk zullen die eisen hoger liggen bij summatieve toetsing, als er voor de studenten belangrijke consequenties aan verbonden zijn. Maar formatieve toetsen, waarbij alleen de criteria consequenties, betekenisvolheid en efficiëntie en kosten (zie rapport) een rol spelen, lijken voor de kwaliteit van die toetsen erg gevaarlijk. Hoe kan men het verdere leerproces goed inrichten en sturen als de informatie waarop men steunt niet voldoet qua betrouwbaarheid en validiteit.

De detail opmerkingen van mevrouw Doorten bij ieder van de onderscheiden criteria komen in grote lijnen overeen met de opmerkingen van Prof. van der Vleuten, namelijk onduidelijke omschrijving en verschillende inhoud naargelang de toetsvorm. Daarnaast wijst zij er ook op dat de hier vermelde acht criteria niet allen fungeren als basis voor het maken van een bewuste keuze. Cognitieve complexiteit, inhoudskwaliteit en inhoudsrepresentatie zijn geen eigenschappen van een *toetsvorm* maar van de *toetstaak*, en afhankelijk van de zorgvuldigheid van de toetsconstructeur. Transparantie zoals hier gedefinieerd is dan weer afhankelijk van de zorgvuldigheid in communicatie van de instelling naar de student. Bovendien zijn deze acht criteria te ruim voor de beoogde doelgroep, docenten met een beperkte kennis van toetsconstructie. Vraag is of zij deze criteria wel begrijpt en ook daadwerkelijk wenst te hanteren bij het samenstellen van toetsen?

Dezelfde opmerkingen werden in grote lijnen geformuleerd door de **CHN.** Ook hier wijst men zeer uitdrukkelijk op onduidelijkheid in de formulering en het feit dat deze criteria aan zich weinig toevoegen aan de klassieke criteria validiteit en betrouwbaarheid. Verder wijst men op het ontbreken van het criterium ‘acceptabiliteit’, zowel voor de docent als de student.
Qua bruikbaarheid van deze criteria voor gewone docenten deelt men volledig de mening van mevrouw Doorten. Deze criteria zijn misschien nuttige nuanceringen van de klassieke kwaliteitscriteria, maar dan voor onderwijsdeskundigen die docenten moeten adviseren.

Vraag 3:

In het rapport wordt voor een eigen indeling qua assessmentvormen en doelstellingen gekozen, waarbij een zekere onderlinge afstemming is nagestreefd. Zijn naar uw mening deze indelingen voldoende ruim en voldoende specifiek? Sluiten ze naar uw mening voldoende aan op bestaande opvattingen en literatuur?

Waar mevrouw **Doorten** de voorgestelde indeling qua doelstellingen positief beoordeelt, hebben **Prof. van der Vleuten** en de **CHN** wel een aantal bedenkingen bij de inhoudelijke omschrijving van de voorgestelde categorieën en formuleren bovendien een ander voorstel. Zo oordelen zij dat in het kader van competentiegericht onderwijs het zinniger is van deze competenties uit te gaan en hierin bepaalde aspecten te onderscheiden. Vervolgens kan men dan voor deze onderscheiden aspecten aangeven welke toetsvormen hier het best bij passen.

Wat de indeling qua toetsvormen betreft worden door de drie beoordelaars vooral opmerkingen geformuleerd ten aanzien van de indeling in klassieke open en gesloten vragen en de daarbij vermelde eigenschappen. De hier geformuleerde opmerking betreft vooral het

feit dat we ons te zeer concentreren op de antwoordmodus en niet voldoende aandacht hebben voor de toetstaak of stimulus. Het is namelijk vooral de toetstaak die bepaald wat er gemeten wordt. De complexiteit van de responsvorm is niet bepalend voor de representativiteit of relevantie van een toets.

Ook is men niet akkoord met de in het rapport gehanteerde omschrijving van performance assessment. Performance assessment wordt namelijk te zeer opgevat als het meten van complexe cognitieve vaardigheden terwijl Doorten en van der Vleuten performance veel eerder zien als concreet gedrag dat via observatie wordt gemeten. Het gaat om complexe opdrachten waarbij zowel het product als het proces doorgaans in de beoordeling worden meegenomen.

Verder wordt opgemerkt dat portfolio maar zeker ADC niet echt als toetsvormen kunnen worden beschouwd daar zij in de praktijk meerdere toetsvormen of een mix van toetsvormen omvatten. Beide vormen, net zoals self-, peer- en co-assessment, passen niet helemaal binnen deze indeling.

In feite betekent dit dat we maar drie belangrijke toetsvormen hebben, gesloten vragen, open vragen met beperkt antwoord en open vragen van complexe aard (performance assessment), maar met een grote verscheidenheid aan vormen. Van der Vleuten vindt een dergelijke indeling in toetsvormen dan ook niet zo zinnig. Een indeling van toetsvormen is een indeling van een bijna oneindige verzameling. Het nut daarvan is niet zo groot voor docenten. Men kan beter een paar voorbeelden geven van representanten; van toetsvormen voor het toetsen van bepaalde doelstellingen. Men dient dan niet zo bekommerd te zijn om alle toetsvormen te bespreken. Iedere nieuwe toetsvorm kan dan onder een van deze categorieën worden geordend. Met de indeling in het rapport kan dit niet; waar hoort bijvoorbeeld een open-boek-examen ? Maar dit betekent dat men vooraf akkoord moet zijn over een goede indeling van doelstellingen. En nogmaals, waarom hier niet vertrekken van een opdeling of aspecten van competenties.

Vraag 4: Validering van de beslissingsschema's

Expliciet vragen wij uw medewerking bij de beoordeling van de scores zoals weergegeven in de twee beslissingstabellen. Op welke punten zou u zelf anders scoren, en kunt u dat toelichten?
Hebt u suggesties om de bruikbaarheid van de beslissingsschema's te verbeteren? Onze intentie is om docenten en onderwijsontwerpers, die niet al te veel achtergrond hebben met betrekking tot toetsing, in die mate te ondersteunen dat zij verantwoorde en beargumenteerde keuzes kunnen maken.

Alleen mevrouw **Doorten** heeft een poging ondernomen om de beide beslissingstabellen opnieuw te scoren, maar komt echter tot de vaststelling dat dergelijke beslissingsschema's niet zo nuttig zijn. Hierna volgt haar commentaar:
'De beslisschema's met de genoemde criteria zijn wel geschikt voor het analyseren van een eenmaal gekozen toetsvorm, om te controleren of een reeds gemaakte keuze voldoet c.q. ongewenste nadelen heeft. Het schema helpt mij echter niet goed om de initiële keuze voor een toetsvorm te maken. Daarvoor vind ik meer informatie in de tekst van het rapport. In zou liever die informatie in een overzichtelijke tabel willen, met daarbij de sterke en zwakke kanten van de toetsvorm. Iets dergelijks als een schema uit Stiggins (1992). Een dergelijk schema geeft mij meer informatie (die nu door het hele artikel verspreid staat) wijst mij op de gevaren van de toetsvorm, en vertelt mij in welke gevallen het vooral voor de hand ligt om voor een bepaalde toetsvorm te kiezen. Als docent met weinig achtergrond in toetsing zou ik in één oogopslag willen zien wat de meer en minder geschikte toetsvorm bij een

bepaalde doelstelling zou kunnen zijn……. Tabel 3.3 op pagina 24 geeft voor docenten meer richting dan de beslisschema's met de kwaliteitscriteria.
Misschien is de naam beslisschema verkeerd gekozen; eigenlijk is het een legitimeringstabel'.

Een gelijkaardig antwoord ontvingen wij van de **CHN:**
'Na het beoordelen van de beslissingstabel kom ik tot een zeer radicale conclusie, namelijk dat in het kader van het ondersteunen van docenten bij het organiseren van een adequate toetsing in het kader van competentiegericht onderwijs, er niet zozeer behoefte is aan een dergelijk beslissingsinstrument maar eerder aan een praktische handleiding voor toetsconstructie. In zo een handleiding voor toetsconstructie komt automatisch aan bod dat om de verschillende aspecten van een competentie te meten je een keuze moet maken uit een mix van toetsvormen. Je zou daarbij kunnen uitgaan van een toetsmatrijs waarbij je aangeeft hoe belangrijk de verschillende aspecten (kennis, vaardigheden en attitude) voor die competentie zijn en met welke toetsvormen je die aspecten wil meten.
Zo een handleiding voor toetsconstructie is bovendien belangrijker dan een beslissingstabel omdat een adequate keuze van toetsvormen nog niet leidt tot een goede toetsing. Je moet die toetsen ook nog samenstellen en toepassen. Welke criteria gelden daarbij en hoe kun je die realiseren?
Het probleem van een adequate toetsing in samenhang met de huidige onderwijsinnovatie is geen probleem van keuze van een geschikte toetsvorm, als zijn bepaalde toetsvormen minder of meer geschikt, maar op de eerste plaats een kwestie van toetsconstructie en toetsafname.
Welke toetsvorm je ook inzet in een bepaalde situatie, je hebt scholing nodig om van die toetsvorm een bruikbare en acceptabele toetsvorm te maken, die beantwoord aan de kwaliteitscriteria validiteit en betrouwbaarheid. Dit geldt zowel voor ja/neen vragen als voor portfolio's. Wil je in het kader van een dergelijke handleiding toch een dergelijke tabel opstellen dan kun je best per doelstellingencategorie een aantal min of meer geschikte toetsvormen vermelden en tevens een aantal tips of vuistregel waarop men moet letten bij de toetsconstructie en toetsafname.'

Ook **Prof. van der Vleuten** heeft heel veel moeite met het scoren van de beslissingsschema's. 'Ik kan dit niet scoren, omdat ik het niet contextloos kan scoren. Het hangt er maar van af of wij praten over een derde jaar van een EHO die meer competentiegericht onderwijs aan het uitproberen is, dan wel een middelbare school die herexamens moet afnemen of een propedeuse die selectief moet zijn en/of een bindend advies moet geven. Dan gelden volstrekt andere criteria. Dit zijn context gebonden gegeven die het nagenoeg onmogelijk maken om een generiek model op te stellen of generieke gewichten te geven'.

Bovendien vindt van der Vleuten dat men zich ten aanzien van toetsing niet mag beperken tot het beoordelen van individuele toetsmethodes. Toetsing is op de eerste plaats het maken van een toetsprogramma dat steeds uit meerdere toetsinstrumenten bestaat. En gegeven de constellatie van gekozen toetsinstrumenten in dat toetsprogramma, kan ik andere wegingen toepassen qua kwaliteitscriteria. Maakt men gebruik van portfolio's of performance assessment om meer generieke en metacognitieve vaardigheden of professionele vaardigheden te meten dan bekomt men meer subjectieve informatie en dit is erg onbetrouwbaar. Maar ik kan die in mijn toetsprogramma opnemen als ik daarnaast ook een aantal meer harde, geobjectiveerde en betrouwbare metingen opneem. Deze afweging van kwaliteitscriteria is dus niet alleen een punt van individuele toetsmethode, maar ook die van het programma in zijn geheel. Als we toetsing meer en meer integreren in het onderwijs en naar toetsing kijken vanuit een standpunt van curriculumconstructie, dan pakken we ook liefst de toetsing als een geheel aan.

We dienen dus niet alleen aandacht te besteden aan afzonderlijke toetsvormen, maar ook aan de integratie van die toetsvormen in een toetsplan. In dat kader zou men een min of meer generiek model kunnen voorstellen, maar men zal daarnaast in exemplarische zin een aantal voorbeelden moeten geven en eventueel een tabel die mensen dwingt om over hun model na te denken. Het gaat er ook niet om een generiek advies te geven, de stand van zaken is niet zo dat men een dergelijk generiek advies kan geven. Zo werk dat niet; mensen moeten en maken zelf hun keuzes. Maar mensen moeten worden gestimuleerd om na te denken over wat ze willen en hoe ze toetsing daartoe het best kunnen opzetten.
Daarbij is toetsing op de eerste plaats een onderwijskundig probleem en geen louter psychometrisch probleem.

Literatuur

Beijaard, D., Longayroux, D.D. & Tanner, R., (1997), Gebruik van portfolio door docenten in opleiding. *Onderzoek van Onderwijs*, (26), pp. 35 – 37.

Biggs, J., (1996), Enhancing teaching through constructive alignment. *Higher Education*, 32, pp. 347-364.

Birenbaum, M., (1996), Assessment 2000: Towards a Pluralistic Approach to Assessment. In: Birenbaum, M. & Dochy, F., *Alternatives Assessment of Achievements, Learning Processes and Prior Knowledge*. Boston, Dordrecht, London, Kluwer, pp. 3-29.

Birenbaum, M. & Dochy, F., (1996), *Alternatives Assessment of Achievements, Learning Processes and Prior Knowledge*. Boston, Dordrecht, London, Kluwer.

Brookhart, S.M., (1999), *The art and science of classroom assessment: the missing part of pedagogy*. Washington, G. Washington Univ., Graduate School of Educ. And Hum. Dev.

Brown, G., Bull, J. & Pendlebury, M., (1997), *Assessing students Learning in Higher Education*, London, New York, Routledge.

Brown, S. & Glasner A., (1999), *Assessment Matters in Higher Education*. Suffolk, St Edmindsbury Press, SRHE & OU.

Bos, E.S., (1998), *Competentie. Verheldering van een begrip*. Heerlen, OUNL, OTEC, 1998.

De la Parra, B., Slotman, R.H., Tillema, H.H., & Spannenburg, T., (2000), *Competentiegerichte leeromgevingen*, Utrecht, Lemma.

Dochy, F., (1999), *Een structureel beleid voor toetsing en assessment in constructiegericht onderwijs (CGO) : Kern van het slagen van universitaire onderwijsvernieuwing*. Diepenbeek, LUC.

Dochy, F., & Segers, M, (1999), Innovatieve toetstvormen als gevolg van constructiegericht onderwijs: op weg naar een assessment-cultuur. In: M. Lacante & P. De Boeck (red.). *Meer kansen creëren voor het Hoger Onderwijs. Handboek Leerlingenbegeleiding*, Dordrecht: Kluwer, pp. 181-206.

Dochy, F. & Sluijsmans, D., (1998), Het gebruik van self-, peer-, en co-assessment in studentgericht onderwijs. *Tijdschrift voor Hoger Onderwijs*, 16(4), pp. 298-314.

Doorten, M. & Moerkerke, G., (1997), *Performance Assessment, meten en beoordelen van complexe vaardigheden*. Heerlen, OUNL, Workshop.

Elsen, M. & Wolters, L., (2000), *De selectie van een geschikte toetsvorm voor het meten van competenties*. Leiden, Paper gepresenteerd op de Onderwijs Research Dagen, mei 2000.

Evelein, F. & van Tarwijk, J. (2000), Overwegingen bij het gebruik van portfolio's binnen een universitaire lerarenopleiding. *Velon: tijdschrift. voor lerarenopleiding*, 21(1), pp. 46-55.

Fleming, N.D., (1999), Biases in Marking Students Written Work: Quality. In : Brown, S. & Glasner A., *Assessment Matters in Higher Education*. Suffolk, St Edmindsbury Press, SRHE & OU, pp. 83-92.

Frederiksen, J.R. & Collins, A., (1989), A systems approach to educational testing. *Educational Researcher*, 18(9), pp. 27-32.

Hambleton, R.K., (1996), Advances in Assessment Models, Methods, and Practices. In : Berliner, D,C. & Calfee, R.C, (eds), *Handbook of Educational Psychology*. New York, McMillan, 1996.

Harden, R.M., (1979), How to assess students: an overview. *Medical Teacher*, 1(2), pp. 65-69.

Herman, J.H., (1992), Accountability and Alternative Assessment: Research and Development Issues. *Educational Leadership*, 49(8).

Hinett, K. & Thomas, J. (eds), (sd), *Staff Guide to Self and Peer Assessment*, Oxford, Center for Staff an Learning Development.

Jellema, F., (2000), De kwaliteiet van 360-graden feedback instrumenten. *M&O*, 3, pp. 23 – 35.

Joosten, G. & Boon, J., (1999), *Evaluatie van de beta-run van het Virtueel Bedrijf.* Heerlen, OUNL, OTEC.

Kessels, J.W.M., (2000), Duaal wetenschappelijk onderwijs. *Opleiding en Ontwikkeling*, 13(3), pp. 5-6.

Klarus, R., Tillema, H. & Veenstra,, J. (1999), Beoordelen met competentieprofielen of kwalificatiestructuren. *Opleiden en Ontwikkeling*. 11, pp. 15-19.

Klarus, R., (1998), *Competenties erkennen; Een studie naar modellen en procedures voor leerwegonafhankelijke beoordeling van beroepscompetenties*. 's-Hertogenbosch, CINOP.

Koper, R, e.a., (1998), *Eindrapportage werkpakket 1.1.,* Heerlen, OUNL, OTEC.

Koper, R., (2000), *Van Verandering naar Vernieuwing: onderwijstechnologische grondslagen voor elektronische leeromgevingen*. Heerlen, OUNL, OTEC.

Linn, R.L., Baker, E.L. & Dunbar, S.B., (1991), Complex performance based assessment: Expectations and validation criteria. *Educational Researcher*, 20(8), pp. 15-21.

Manderveld, J., e.a., (1999), *Eindrapportage deelproject onderwijsaanpak. Didactische scenario's*. Heerlen, OUNL, OTEC.

McDowell, L. & Sambell, K., (1999), The Experience of Innovative Assessment: Student Perspectives. In : Brown, S. & Glasner A., *Assessment Matters in Higher Education*. Suffolk, St Edmindsbury Press, SRHE & OU, pp. 71-82.

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

Messick, S., Validity. In: Linn, R.L. (ed.), (1983), *Educational Measurement*. New York, (3rd ed.), pp. 13-104.

Meyer, R.E., (1992), *Problemsolving: Teaching and Assessment*.

Moerkerke, G., (1996), *Assessment for Flexible Learning*, Utrecht, Lemma.

Moerkerke, G., (1998), Toetsing van academische vaardigheden: een curiculum perspectief. *TVHO*, 16(3), pp. 178-193.

Moerkerke, G. & Dochy, F., (1995), Het toetsen van complexe vaardigheden. In : ten Dam, G.E., e.a., (eds), *Handboek onderwijskunde voor het hoger onderwijs*. Assen, van Gorcum, pp. 214-236.

Nedermeyer, J, & Pilot, A., (2000), *Beroepscompetenties en academische vorming in hoger onderwijs*. Groningen, Wolters Noordhoff, HOR.

Projectplan Elektronische Leeromgeving, Heerlen, OUNL, OTEC, 1998.

Schuwirth, L.W.T. & van der Vleuten, C.P.M., (1999), Toetsen van probleemoplossend vermogen : Computergestuurde Casusgerichte Toetsen. In : Heijnen, G. & Meeder, S. (eds), *Toetsen en ICT in het hoger onderwijs*. Utrecht, Surf, pp. 155-161.

Segers, S.M.R., (1998), Het toetsen van probleemoplossende vaardigheden. Ervaringen met de OverAll Toets. *Tijdschr. vr Hoger Onderwij*s. 16, pp. 155-177..

Segers, M, Dochy, F. & Dierick, S., (2000), Een onderwijsmodel … een ander toetsmodel ? In: Wold, A., *Leren in perspectief*. Boston, Dordrecht, London, Kluwer, pp. 1-12.

Sluijsmans, D. & Dochy, F., (1998), Alternatieve toetsmethoden in studentgericht onderwijs. *Tijdschrift voor Hoger Onderwijs*, 1998, 16 (4), pp. 298-314.

Sluijsmans, D., Dochy F. & Moerkerke G., (1998), *The use of self-, peer- and co-assessment in higher education*. Heerlen, OUNL, OTEC.

Sluijsmans, D., Martens, R. & Verheijen, H., (2000), Peer-Assessment en onderwijsontwerp. *O.I.,* 2(1), pp. 17-24.

Snippe, M.D., & Smit, G.N., (1997), De Assessment Centermethode als feedbackinstrument in het onderwijs. *Tijdschrift voor Hoger Onderwijs*, 15, pp. 339-364.

Stiggins, R.J., (1992), Het ontwerpen en ontwikkelen van performance-assessment-toetsen. In : Kessels, J.W.M. & Smit, C.A. (eds), *Opleiders in organisaties/Capita Selecta*, Afl. 10, Kluwer Bedrijfswetenschappen.

Tillema, H.H., (1996), *Development centers: ontwikkelen van competenties in organisaties*.

Tillema, H.H., Kessels, J.W.M., & Meijers, F., (2000), Competencies as building blocks for integrating assessment with instruction in vocational education: a case from The Netherlands. *Assessment & Evaluation in Higher Education*, 25*,* pp. 265-278.

Voeten, M.J.M., (2000), Evalueren van Leervorderingen. In: *Onderwijskundig Lexicon*. Ed. III, Alphen. aan de Rhijn, Samson, pp. 15-40.

van der Vleuten, C.P.M., (1998), Assessment in Problem-based Learning. In: van Merrienboer, J. & Moerkerke, G., (eds), *Instructional Design for Problem-based Learning*. Maastricht, Proceedings of the Third Workschop of the EARLI SIG Instructional Design, pp. 303-311.

van der Vleuten, C.P.M., & Driessen, E.W., 2000), *Toetsen in probleemgestuurd onderwijs*. Groningen, Wolters Noordhoff, HOR.

van der Vleuten, C.P.M., Verwijnen, G.M. & Wijnen, W.H.F.W., (1996), Fifteen years of experience with progress testing in a problem-based learning curriculum. *Medical Teacher*, 18(2), pp. 103-110.

van der Vleuten, C.P.M., e.a., 1999), De rol van ICT in studietoetsen: een verkenning. In : Heijnen, G. & Meeder, S. (eds), *Toetsen en ICT in het hoger onderwijs*. Utrecht, Surf, pp. 7-15

van Merriënboer, J.J.G., 1999), *Cognition and multimedia Design*. Heerlen, OUNL, OTEC, Inaugurale rede.

van Tarwijk, J. & Wubbels, Th., (2000), Evalueren van leervorderingen met portfolio's. In : *Onderwijskundig Lexicon*. Ed. III, A. aan de Rhijn, Samson, pp. 41-58

van Tarwijk,J.W.F., Pilot, A. & Wubbels, T., (1999), Naar een digitale portfolio, In: Heijnen, G. & Meeder, S. (eds), *Toetsen en ICT in het hoger onderwijs*. Utrecht, Surf, pp. 136-146.

Vos, H.J. & Knuver, J.W.M., (2000), Standaarden in onderwijsevaluatie. In : *Onderwijskundig Lexicon*. Ed. III, A. aan de Rhijn, Samson, pp. 59-76.

Wolf, K. & Dietz, M., (1998), Teaching portfolios: purposes and possibilities. *Teacher Education Quarterly*, 25, pp. 9-22.

Wolters, L., (2000), Toetsen van academische vaardigheden. In: Nedermeyer, J, & Pilot, A., *Beroepscompetenties en academische vorming in hoger onderwijs*, Groningen, Wolters Noordhoff, HOR.