Sabine Dierick
Filip Dochy
Gerard Van de Watering

Assessment in het hoger onderwijs

Over de implicaties van nieuwe toetsvormen voor de edumetrie

S. Dierick is als onderwijskundige/toetsdeskundige werkzaam bij de capaciteitsgroep Onderwijsinnovatie & IT van de Faculteit der Rechtsgeleerdheid, Universiteit Maastricht.
F. Dochy is hoogleraar onderwijskunde aan de Universiteit Maastricht en de KU Leuven.
G. Van de Watering is als onderwijskundige/toetsdeskundige werkzaam bij de capaciteitsgroep Onderwijsinnovatie & IT van de Faculteit der Rechtsgeleerdheid, Universiteit Maastricht.

Assessment is recentelijk een topic geworden in vele universiteiten en hogescholen die druk zijn met de innovatie van hun onderwijs. De kern daarbij is nieuwe vormen van leren: constructiegericht onderwijs, leren leren, probleemgestuurd leren, casusgeoriënteerd leren. Kortom alle mogelijke varianten van studentgericht onderwijs. Niet de traditionele tests of toetsen gericht op hoofdzakelijk reproductieve kennis sluiten aan bij dit soort onderwijs, wel de zogenaamde nieuwe toetsvormen of assessmentvormen. Deze verzamelnaam heeft betrekking op de toetsen die gericht zijn op het meten van kennis en vaardigheden in authentieke situaties en van competenties. In dit artikel wordt aandacht besteed aan de gevolgen van deze evoluties voor het screenen van de edumetrische kwaliteit van assessment. Er wordt geconcludeerd dat de traditionele testtheoretische betekenis van validiteit en betrouwbaarheid niet langer hanteerbaar is, maar minstens uitgebreid (c.q. vervangen) dient te worden. Bij assessment zijn kwaliteitsaspecten als transparantie van de toetsprocedure, eerlijkheid, cognitieve complexiteit, authenticiteit van taken, en de invloed van de assessment op het onderwijs van belang.

Inleiding

Recente veranderingen in het hoger onderwijs richten zich in grote mate op het realiseren van 'krachtige leeromgevingen' (De Corte, 1990). Met name die omgevingen waarin studenten coöperatief leren, uitgaande van concrete authentieke, uit de praktijk gegrepen problemen en gericht op het toepassen van kennis en vaardigheden. Deze veranderingen vloeien voort uit nieuwe inzichten in diverse stromingen in de onderwijskunde (Segers & Dochy, 1999). Daarnaast is er meer aandacht gekomen voor de aansluiting van het onderwijs bij de behoeften van de arbeidsmarkt (Boud, 1990). In de informatiemaatschappij van morgen moet de afgestudeerde naast een zekere basiskennis van de

wetenschap ook vaardigheden bezitten als probleemoplossen, analyseren, synthetiseren, coachen, leiding kunnen geven, presenteren, kritisch evalueren, enzovoort naast de wetenschapsspecifieke vaardigheden. Uit de organisatie van het onderwijs en het leren zal moeten blijken dat studenten niet alleen goede probleemoplossers worden die kunnen werken in multidisciplinaire teams, maar ook dat zij kunnen reflecteren op de kwaliteit van het eigen werk en op het werk van anderen. De combinatie van deze kennis en vaardigheden maakt dat de afgestudeerde als een 'reflective practitioner' aan de slag kan. Met name het basisgedachtegoed uit het constructivisme en de stroming van het coöperatief leren heeft de jongste 25 jaar geleid tot een brede implementatie van diverse nieuwe onderwijsvormen in een diversiteit van disciplines. Vormen van constructiegericht leren, leren leren, probleemgestuurd onderwijs, competentiegericht onderwijs, combinaties van werken en leren of coöperatief leren worden geïmplementeerd (Dochy e.a., 2000).
Bij een dergelijke verandering van onderwijskundige aanpak kan men de toetsing van de leerresultaten van de studenten niet langer uitsluitend baseren op kennisgerichte of reproductiegerichte toetsen. Erger nog, een overschakeling naar constructiegericht onderwijs (Dochy, 1999) met een status-quo bij de wijze van toetsing zal leiden tot een verdwijnen van de meerwaarde van de innovatie. Welke student spendeert immers tijd en moeite aan het analyseren van problemen wanneer enkel de reproductie van feitenkennis wordt getoetst? (Kessels, 2000a, b). Met andere woorden, ook voor de beoordeling zal een andere oriëntatie gezocht moeten worden. Veranderingen van onderwijskundige aanpak leiden tot de noodzaak aan innovatieve toetsvormen. Deze bewerkstelligen enerzijds dat de door de wetenschap en het werkveld gewenste kennis en vaardigheden beoordeeld zijn op voldoende niveau. Anderzijds sturen ze het leergedrag van de studenten in deze richting.

Kenmerken van assessment

Vanwege de algemene negatieve connotatie die wordt toegekend aan de begrippen 'examens' en 'testen', is de nieuwe stroming uitgegaan van de term 'assessment' (Birenbaum, 1996). Zeker ook omdat een aantal steeds meer gebruikte 'assessment'-vormen in het onderwijs afkomstig zijn uit de bedrijfswereld waar selectie en assessment al jaren worden gebruikt. In die context werd assessment eerst gebruikt als aanduiding van een selectie-instrumentarium, thans ook ruimer als een instrument voor potentieelmeting of als procedure voor het in kaart brengen van persoonlijke ontwikkelingsdoelen en -plannen (c.q. assessmentcenter; development center) (Dochy & De Rijke, 1995).
Met de term 'assessment' wordt niet alleen verwezen naar het 'meten' of 'in kaart brengen', er wordt evenzeer verwezen naar de basiskenmerken van assessment:

1 kennisconstructie is het uitgangspunt, niet kennisreproductie;
2 zowel basiskennis, het toepassen van kennis en vaardigheden zijn het doel van de toetsing (kennen en kunnen);
3 er worden authentieke of levensechte situaties gebruikt in bijvoorbeeld de vorm van casussen of problemen;
4 studenten worden actief betrokken bij het ontwerp en/of de uitvoering van toetsprocedures;
5 integratie van het toetsen in het leer- en instructieproces.

Daarnaast worden vaak nog andere kenmerken gegeven waaraan 'assessment' kan voldoen:

- Assessment is veel minder gestandaardiseerd dan testen.
- Bij assessment worden zowel het product als het proces dat geleid heeft tot het product opgenomen in het beoordelingsproces.
- Assessment wordt vooral gebruikt voor de beoordeling van complexere leerdoelen of leerprocessen. (procedurele kennis, methoden/vaardigheden en attitudes), dit wil zeggen dat assessment niet enkel uitkomsten moet weerspiegelen in een enge technische zin, maar in termen van basiskennis, verstaan, communicatieve en competentiedoelen die bereikt worden.
- Ook door de lerende zelf wordt beoordeeld. Dit is geen alleenrecht meer is van de onderwijsgevende. Hierdoor leren studenten niet alleen kritische zin en zelfreflectie, maar ook verkrijgen ze inzicht in de criteria.
- Assessment dient niet enkel om formele beslissingen te nemen, maar dient ook zoveel als mogelijk als input voor directe en gepaste feedback.
- Het assessmentproces moedigt selfassessment en reflectie aan.
- Assessment moet congruent zijn met het onderwijs.
- Assessment is zoveel mogelijk een continu gebeuren.
- De rol van de docent en de assessor/examinator worden zoveel mogelijk gescheiden.
- Assessment geeft aanleiding tot zelfwaakzaamheid, zelfwerkzaamheid en zelfsturing.

Studentgericht toetsen

Het ontwerp en de implementatie van studentgerichte krachtige leeromgevingen waarbinnen het leren leren en het verwerven van kennis en vaardigheden centraal staat, heeft consequenties voor wie beoordeelt. De veranderingen in de taken van de student en de taken van de onderwijsinstelling/docent geven aan dat de beoordeling niet meer alleen in de handen hoeft te liggen van de docent. Studenten worden opgevat als zelfstandige, autonome en initiatiefrijke individuen die in grote mate zelf hun leerproces sturen (Sluijsmans & Dochy, 1998). Kennis en competenties worden bij voorkeur niet overgedragen via reproductieve technieken, maar op actieve wijze door studenten verworven (geconstrueerd). Daarbij moet de student leren kritisch te kijken naar eigen handelen en kennis.

Meer dan kennis alleen toetsen

Curricula in het hoger onderwijs richten zich steeds meer op doelen zoals het opleiden van studenten die in staat zijn (beroepsrelevante) problemen op een adequate wijze op te lossen. In ons steeds wijzigend jargon heet dat tegenwoordig 'basiscompetenties'. Cognitief-psychologisch onderzoek toonde aan dat een succesvolle probleemoplosser over een goede kennis beschikt van de concepten en principes van de relevante disciplines. Men spreekt in dit geval van een goed gestructureerd kennisbestand: men begrijpt de concepten en principes, hun onderlinge relaties en de condities waarin ze bruikbaar zijn. Dit impliceert dat basiskennis een noodzakelijke voorwaarde is voor het kunnen gebruiken van deze kennis om problemen op te lossen. In dit verband is het toetsen van basiskennis relevant Het is echter niet voldoende. Als we studenten wensen op te leiden die deze kennis kunnen hanteren als instrument om op een professionele wijze problemen aan te pakken en die de basiscompetenties beheersen, dan dient dit ook getoetst te worden.

Integratie van leren en assessment

De beoordeling zal in een krachtige leeromgeving niet alleen gericht zijn op het 'afrekenen' van de student, maar zal ook een rol moeten hebben in het structureren en monitoren van het leerproces. De prioritaire aanpak bij assessment vraagt dan eerder om een analyse van de sterkten en zwakten van de kennis en vaardigheden, dan om toetsen die tot een simpele eendimensionele score leiden (Birenbaum & Dochy, 1996). Toetsen dienen niet alleen certificerend te zijn, maar ook ondersteunend voor het leerproces. Boud (1990) onderkent de kloof tussen datgene dat van de student wordt vereist in het hoger onderwijs en wat gebeurt in het professionele leven. Hij benadrukt de behoefte aan onderzoek naar assessment om te kijken of toetsen ook overeenkomen met de doelen van het hoger onderwijs. Dit betekent dat assessment moet leiden tot gegevens die mede een formatief karakter hebben en op deze wijze bijdragen tot de ontwikkeling van reflectie en de basiscompetenties.

Kennisconstructie

Krachtige leeromgevingen zijn onderwijssettings waarbij wordt uitgegaan van authentieke situaties, problemen, casussen, opdrachten, die tot doel hebben de studenten kennis en vaardigheden te leren toepassen in een proces van samenwerking met medestudenten. Er is niet noodzakelijk een objectieve waarheid met betrekking tot de leerstof, maar studenten worden geacht hun eigen perspectief te ontwikkelen en de vooropgestelde doelen te bereiken (Kessels, 2000a). Het gaat dus om 'knowledge building' in plaats van 'knowledge reproduction'. Bij de assessment zal dus de kennisconstructie worden genomen als uitgangspunt, niet de hoeveelheid die een student kan reproduceren. Het belangrijkste bij een innovatie in die richting is dat de assessment dan als bouwsteen fungeert. Anders gezegd: men moet ervan uitgaan dat examens het sterkst sturende aspect van de leeractiviteiten van studenten zijn. Eerder onderzoek heeft dit trouwens aangetoond (Birenbaum & Dochy, 1996).

Authentieke situaties en cases

In het huidige onderwijs zijn veel verschillende soorten toetsen te onderscheiden, veelal gericht op het toetsen of 'testen' van de reproductie van kennis. Voorbeelden zijn de toets met open vragen, het essayexamen, de voortgangstoets, de multiplechoicetest en combinaties hiervan.

Assessment is meer dan dat: studenten worden beoordeeld op basis van hun actieve prestatie om kennis te gebruiken op een creatieve manier en competentie te demonstreren; de uitgangssituaties zijn veelal reële problemen, authentieke representaties van problemen in de werkelijkheid; studenten krijgen een eigen verantwoordelijkheid in het toetsproces en zowel kennen als kunnen wordt in kaart gebracht. Nieuwe assessmentvormen wijken daarmee af van het traditionele 'testen'.

Hoe kunnen we nu de kennisbasis die studenten zelf hebben verworven toetsen? En hoe beoordelen we wetenschappelijke en algemene vaardigheden, zoals een balans kunnen interpreteren binnen een specifieke bedrijfscontext, een diagnose kunnen stellen op basis van tegenstrijdige informatie, probleemoplossen, informatie kunnen zoeken, discussiëren, vergaderingen leiden, notuleren, synthetiseren, analyseren, kritisch evalueren?

Nieuwe assessmentvormen zijn daartoe ontwikkeld en worden met succes toegepast. Voor een nadere beschrijving verwijzen we naar eerdere publicaties (Birenbaum, 1996; Sluijsmans & Dochy, 1998; Dochy e.a., 1999). Bij de implementatie van assessmentmethoden wordt nauwlettend gekeken naar de edumetrische kwaliteiten van de assessment. De vraag die we ons hier verder stellen is of we het kunnen laten bij de traditionele concepten als validiteit en betrouwbaarheid en of deze concepten niet uitgebreid dienen te worden in functie van de nieuwe assessmentcontext.

Op zoek naar nieuwe kwaliteitsindicatoren voor assessment

Om een oordeel te kunnen geven over de kwaliteit van een toets zal men deze evalueren op een aantal edumetrische kwaliteiten. Traditioneel werden de begrippen 'betrouwbaarheid' en 'validiteit' gebruikt als kwaliteitsindicator. De validiteitsvraag had betrekking op de mate waarin de toets meet wat men ermee wenst te meten. In hoeverre stemt de inhoud van de toets overeen met de doelstelling van het onderwijs. Betrouwbaarheid werd gedefinieerd als de mate waarin een toets consistent meet. Consistentie in toetsresultaten betekende objectiviteit in scoring: dezelfde resultaten worden verkregen als de toets beoordeeld wordt door een andere persoon of door dezelfde persoon op een ander moment.

De invulling van het begrip betrouwbaarheid werd bepaald door de toen heersende opvatting dat toetsen vooral een selecterende functie diende te vervullen. Eerlijkheid in beoordeling werd gelijk gesteld aan objectiviteit. Het streven naar objectiviteit in testen en het kunnen vergelijken van scores, resulteerde in het gebruik van gestandaardiseerde toetsvormen, zoals multiple-choicetesten. Dit soort testen, die in de praktijk vooral reproductiegerichte kennis meten, worden momenteel bekritiseerd voor hun negatieve invloed op instructie. Omdat aan deze toetsresultaten belangrijk gewicht werd toegekend op beleidsniveau, gingen docenten hun onderwijs afstemmen op de inhoud en het niveau van de kennis die bevraagd werd in deze test. Dit had tot gevolg dat voornamelijk lagere cognitieve kennisniveaus aan bod kwamen.

Als reactie op de negatieve effecten van multiple-choicetesten werden nieuwe toetsvormen ontwikkeld. Bij deze nieuwe toetsvormen worden studenten beoordeeld op basis van hun prestatie om kennis te gebruiken om op een creatieve manier problemen op te lossen (Dochy, 1999). Assessmenttaken zijn reële problemen of authentieke representaties van problemen in de werkelijkheid.
Gezien een belangrijke doelstelling van het huidige onderwijs is om studenten op te leiden die hun kennis kunnen gebruiken om reële problemen op te lossen, lijkt assessment (en de daarin horende nieuwe toetsvormen) meer valide te zijn dan gestandaardiseerde toetsen. Traditionele multiple-choicetesten vereisen immers steeds een interpretatie vanuit studentantwoorden naar competentie in een specifiek domein. Bij authentieke assessment is interpretatie van de antwoorden niet nodig, omdat assessment reeds een directe uiting is van competentie. Echter, precies omwille van hun authentiek, niet gestandaardiseerd karakter scoren deze toetsvormen ongunstig op een conventionele betrouwbaarheidsmeting. Het uitgangspunt is immers niet meer hetzelfde.

Door de unieke aard van de nieuwe assessmentvormen kunnen we de traditionele benadering voor het meten van betrouwbaarheid in vraag stellen. In de eerste plaats omdat bij assessmentvormen die, in tegenstelling tot traditionele testen niet gestandaardiseerd zijn, een conventionele betrouwbaarheidsmeting een vertekening van de resultaten geeft. Ten tweede omdat assessment onvermijdelijk inhoudt dat op verschillende momenten, door verschillende beoordelaars op verschillende manieren de diverse vaardigheden en kennis worden getoetst.

Het is evident dat een nieuwe assessmentcultuur niet zomaar kan geëvalueerd worden op basis van de bestaande criteria. Om recht te doen aan de eigenheid van assessmentmethoden dient er gezocht te worden naar een verruiming van de traditioneel gebruikte criteria of naar andere, meer gepaste kwaliteitscriteria.

Verschillende auteurs hebben voorstellen gedaan om de criteria, technieken en methoden die gebruikt worden in de traditionele psychometrie uit te breiden (Cronbach, 1989; Haertel, 1991; Kane,1992; Linn e.a., 1991; Messick, 1994). In de volgende paragrafen geven we een geïntegreerd overzicht van de diverse visies.

Vernieuwde edumetrische criteria voor het evalueren van assessment: een overzicht

Binnen de literatuur over kwaliteitscriteria voor het evalueren van assessment kan onderscheid gemaakt worden tussen auteurs die een verruimde visie op validiteit en betrouwbaarheid presenteren (Cronbach, 1989; Kane, 1992; Messick, 1994) en die specifieke criteria voorstellen om de unieke kenmerken van nieuwe toetsvormen beter tot hun recht te laten komen auteurs (Frederickson & Collins, 1989; Haertel, 1991; Linn e.a., 1991).

Constructvaliditeit als alomvattend criterium voor de kwaliteit van assessment
Binnen de klassieke traditie werd de validiteit van een toets nagegaan aan de hand van drie aspecten, een inhoudelijk -, een criterium -, en een constructvaliditeitsaspect.
Het inhoudelijk validiteitsaspect bestudeert in welke mate de reeks en het type taken die gebruikt worden bij toetsing een adequate reflectie zijn van het kennisdomein dat gemeten wordt.
De term construct is een meer abstract begrip dat binnen deze traditie refereert naar psychologische denkprocessen, die ten grondslag liggen aan een domeinkennis, zoals het denkproces, de redenering die gevolgd wordt om algebra op te lossen. Het meten van constructvaliditeit gebeurde vanuit een psychometrische benadering, waarbij het te meten construct binnen een conceptueel netwerk werd geplaatst en de verwachte relaties tussen metingen van gerelateerde constructen statistisch werden onderzocht. De mate waarin er een correlatie was tussen scores op testen die hetzelfde construct meten werd criteriumvaliditeit genoemd. Evidentie dat een beoordeling criteriumvalide is, werd dus als argument gebruikt voor de constructvaliditeit van een toets.

Omdat bij nieuwe assessmentvormen gewerkt wordt met complexe, vaak multidisciplinaire problemen is het niet evident om de validiteit van assessment psychometrisch te benaderen. Daarom wordt binnen de nieuwe toetscultuur gezocht naar meer realistische benaderingen voor het meten van constructvaliditeit. Indien afstand gedaan wordt van de psychometrische benadering, dient ook nagedacht te worden over een andere definiëring van het abstract begrip 'construct'. Beargumenteerd kan worden dat de term construct het best kan ingevuld worden door de term competentie. Immers, doelstellingen binnen het hoger onderwijs worden vandaag de dag steeds meer geformuleerd in termen van competenties die bereikt dienen te worden.

Daarnaast heeft onderzoek aangetoond dat het gebruik van een bepaalde toetsvorm niet alleen bepalend is voor het soort kennis en vaardigheden dat wordt gemeten, maar ook bredere effecten heeft op de aard en inhoud van het onderwijs (Birenbaum & Dochy, 1996). Illustratief is de negatieve invloed die gestandaardiseerde testen hebben gehad op het soort kennis dat bevraagd werd, op de wijze waarop onderwijs werd gegeven en bijgevolg ook op de wijze waarop studenten de leerstof benaderden. Nieuwe toetsvormen worden geïmplementeerd om deze negatieve effecten tegen te gaan. Beargumenteerd wordt, dat het toetsen van hogere orde-vaardigheden zal leiden tot het leren van gelijkaardige kennis en vaardigheden. Examens zijn immers het meest sturende aspect gebleken van de leeractiviteiten van studenten. Voor het evalueren van de geschiktheid van een toetsvorm is daarom niet alleen de vraag of de toets een goede meting is van de bedoelde kennis en vaardigheden belangrijk, maar ook de vraag of het gebruik van de toets de verwachte effecten heeft gerealiseerd (Messick, 1989).

Bovenstaande argumenten hebben aanleiding gegeven tot een kritische reflectie ten aanzien van de traditionele benadering voor het valideren van nieuwe toetsvormen. Dit heeft geresulteerd in het gebruik van de term constructvaliditeit als alomvattend criterium voor het beoordelen van de kwaliteit van assessment.

De auteurs van de 'Standards' (AERA, APA, NCME, 1985) definiëren constructvaliditeit als 'een geïntegreerd criterium waarbij de gepastheid, betekenisvolheid en bruikbaarheid van interpretaties die gemaakt worden op basis van toetsscores dienen gerechtvaardigd te worden' (AERA, APA, NCME, 1985, p. 9). In dit constructvaliditeitscriterium worden de drie traditionele aspecten voor het bestuderen van validiteit verenigd.
Voor het omschrijven van het constructvalideringsproces wordt onder meer aansluiting gezocht bij de interpretatieve onderzoekstraditie. Auteurs zoals Kane (1992) en Cronbach (1989) hanteren een argumentatieve benadering voor het valideren van assessment waarbij argumenten gezocht worden om de gegeven interpretatie te rechtvaardigen, en mogelijke alternatieve interpretaties te weerleggen.

Messick's opvatting (1994) biedt de meest verruimde visie over het begrip constructvaliditeit. Hij omschreef dit concept als een samenvatting, waarin evidentie voor de interpretatie van een assessment en alle andere mogelijke gevolgen van deze interpretatie en van het gebruik van een bepaalde assessmentvorm binnen een leeromgeving, wordt geëvalueerd. Volgens hem omvat dit concept zes onderscheiden aspecten, die gezamenlijk

fungeren als kwaliteitscriteria voor assessment. Bij het bestuderen van het inhoudelijk aspect wordt nagegaan of de reeks en het type taken die gebruikt worden bij assessment een adequate reflectie zijn van het constructdomein, zowel naar inhoud als naar cognitief niveau. Het bewaken van het inhoudelijk aspect bevordert het substantieel aspect: denkprocessen bij het oplossen van taken in een assessmentcontext zullen meer consistent zijn met die van experts, indien het inhoudelijk validiteitsaspect gewaarborgd is. Bij het intern structureel aspect wordt onderzocht of de gebruikte criteria, de relaties tussen deze criteria en het gewicht dat eraan wordt gegeven (reflecteert de interne structuur van assessment) consistent zijn met de structuur binnen het constructdomein. Het generaliserend aspect bestudeert of de score-interpretatie die gebaseerd is op één assessmenttaak ook kan veralgemeend worden naar andere domeinspecifieke taken. De mate waarin scores op assessment hoog correleren met domeinverwante en laag correleren met niet-verwante taken refereert naar het extern validiteitsaspect. Het consequentiële aspect evalueert zowel de gevolgen van de interpretatie over assessment, als de effecten van het gebruik van assessment binnen een leeromgeving (consequential validity).
De visie dat de drie traditionele validiteitsaspecten worden geïntegreerd binnen één constructvaliditeitscriterium voor het evalueren van assessment vinden we ook terug in Messick's opvatting. Het meest vernieuwende aspect van deze validiteitbenadering is echter de nadruk die gelegd wordt op het evalueren van de invloed van assessment op het onderwijs. Daarnaast wordt ook het begrip betrouwbaarheid een onderdeel van het constructvalideringsproces. Bij nieuwe toetsvormen gaat het er immers niet om de scores van studenten volgens een normaalverdeling te selecteren. De belangrijkste vraag luidt: in welke mate is de beslissing 'of iemand wel of niet competent is' betrouwbaar?
Het generaliserend validiteitsaspect bestudeert de mate waarin de beslissing of een student competent is, kan veralgemeend worden naar andere taken en dus betrouwbaar is. Het meten van betrouwbaarheid kan dan opgevat worden als het nagaan van de mate van nauwkeurigheid waarmee een test score kan gegeneraliseerd worden naar een breder competentiedomein.

Hierna zal verder beargumenteerd worden waarom niet de klassieke betrouwbaarheidstheorie, maar wel de generaliseerbaarheidstheorie kan gebruikt worden om de betrouwbaarheid van authentieke assessment uit te drukken.

Meten van betrouwbaarheid: de vraag naar de mate van nauwkeurigheid waarmee assessment kan gegeneraliseerd worden
Volgens de traditionele wijze van toetsen kan de meettechnische juistheid van een toets op twee manieren opgevat worden. Enerzijds wordt betrouwbaarheid opgevat als de mate waarin overeenstemming tussen beoordelaars wordt bereikt. In geval van performance assessment kan betrouwbaarheid worden vertaald in termen van overeenstemming tussen beoordelaars bij één bepaalde taak tijdens één bepaald moment. De betrouwbaarheid kan volgens deze definitie verhoogd worden door de beoordelingen te structureren door middel van gedetailleerde protocollen en antwoordmodellen gecombineerd met een extensieve training voor de beoordelaars.
Heller e.a. (1998) stellen echter dat metingen op het gebied van interbeoordelaarbetrouwbaarheid in authentieke assessment niet noodzakelijkerwijs aangeven of beoordelaars op de juiste wijze beoordelen en ook niet voorzien in een basis om de toetstechnische kwaliteit te

verbeteren. Immers, verschillen tussen beoordelaars in hun beoordelingen geven soms een nauwkeuriger en meer betekenisvolle meting weer dan absolute overeenkomsten dat doen. Om tot een betrouwbare beslissing te komen is het beter om naar meer bewijs te zoeken over de competentie van een student, dan te oordelen op basis van één test (Suen e.a., in press).

Anderzijds is betrouwbaarheid de mate waarin de scores van een toets bij een herhaalde meting en bij dezelfde beoordelaar consistent is. De betrouwbaarheid van een toets kan volgens deze opvatting verhoogd worden door de taken binnen de performance assessment meer met elkaar te laten overeenstemmen qua inhoud en formaat.
Volgens Bennet (1993) werkt deze definitie van betrouwbaarheid bij performance assessment ook niet om een andere reden: in de periode tussen de eerste taak en de tweede taak van de assessment leert een student gewoon door. Het is in de praktijk dan ook moeilijk in te zien hoe een beoordelaar kan verzekeren dat dezelfde aspecten van dezelfde of soortgelijke taak op verschillende momenten op dezelfde manier getoetst worden. Een beoordelaar zal eerder naar de vorderingen van een student zoeken dan rekening houden met de consistentie van de scores.

Gezien de bovenstaande bevindingen lijken betrouwbaarheid en de bedoeling van assessment tegenover elkaar te staan. Te veel concessies doen voor het bereiken van een hoge betrouwbaarheid op het gebied van wat studenten moeten kennen en kunnen zal dan ook tot vermindering van de inhoudsvaliditeit leiden.
Binnen de nieuwe assessmentcultuur heerst de visie dat het wegen van informatie over een student vanuit verschillende invalshoeken tot een meer adequate beslissing leidt dan informatie uit één invalshoek. Het is van belang studenten een eerlijke kans te geven om te tonen wat ze echt kennen en kunnen. Dit veronderstelt dat alle vaardigheden en kennis die relevant geacht wordt, getoetst wordt. Daarbij kan gebruikgemaakt worden van verschillende vormen van toetsing waarbij je de verschillende informatie gebruikt om beslissingen te nemen (en niet eerst reduceert tot één dimensie en dan een beslissing neemt).
Bij deze opvatting over assessment kun je niet zondermeer de klassieke betrouwbaarheidstheorie gebruiken om de betrouwbaarheid van een toets uit te drukken. Het gaat er immers om dat je met behulp van een reeks toetsen of opdrachten de daadwerkelijk competentie van studenten vaststelt en dat je niet probeert zoals gezegd om de scores van de studenten volgens een normaalverdeling te selecteren. Daarnaast kun je met behulp van de klassieke betrouwbaarheidstheorie alleen nagaan of een toets op een bepaald toetsmoment consistent meet of dat er overeenstemming is tussen de verschillende beoordelaars.
Assessment betekent op verschillende momenten en op verschillende wijze, eventueel door verschillende beoordelaars, de mogelijkheden van de studenten meten. Assessment dient dan ook als een geheel beschouwd te worden en niet als een reeks afzonderlijke onderdelen. Analoog hieraan is de betrouwbaarheid van het geheel interessant en niet meer de betrouwbaarheid van de afzonderlijke delen. Het is zinvoller om te onderzoeken in hoeverre het vertoonde gedrag te generaliseren (transfereren) is naar bijvoorbeeld de beroepswerkelijkheid en hoeveel taken, toetsmomenten en beoordelaars daarvoor nodig zijn.

De bredere kijk op betrouwbaarheid heeft tot het ontstaan van de generaliseerbaarheidstheorie geleid (Cronbach e.a., 1972). In plaats van af te vragen hoe nauwkeurig de geobserveerde scores een weerspiegeling zijn van de daarbijbehorende ware scores, vraagt de generaliseerbaarheidstheorie zich af hoe nauwkeurig de geobserveerde scores het toestaan het gedrag van een student te generaliseren naar een goed gedefinieerd universum.

Wat kan de generaliseerbaarheidstheorie wel wat de klassieke betrouwbaarheidstheorie niet kan?

De klassieke betrouwbaarheidstheorie probeert antwoord te geven op de vraag hoe nauwkeurig een geobserveerde score overeenkomt met de daarbij behorende ware score. Hoe meer de scores uit de toets overeenkomt met de hypothetische ware score hoe kleiner de fout (error), hoe hoger de betrouwbaarheid.
De geobserveerde score wordt opgesplitst in een gedeelte ware score en een gedeelte error. De error bevat hierbij alle mogelijke bronnen van variantie in de score (bijvoorbeeld de items, toetsmomenten en de beoordelaars). En omdat taken binnen assessment vaak complex zijn en een open eind hebben is de kans op een grote error aanwezig (Cronbach e.a., 1997).
De verschillende foutbronnen in de error kunnen in de klassieke theorie echter niet onderscheiden worden. De generaliseerbaarheidstheorie kan, mits goed gedefinieerd, wel onderscheid maken tussen de verschillende foutenbronnen. Met generaliseerbaarheidstheorie is het mogelijk om de grootte van de verschillende bronnen (en de interactie ertussen) van meetfouten simultaan te herkennen te schatten. Op deze manier krijgt de onderzoeker antwoord op de vraag in welke mate de geobserveerde data te generaliseren is naar een goed gedefinieerd ‘universum’. Of anders gezegd in welke mate de meting overeenkomst met de werkelijkheid en welke componenten of interacties tussen componenten de oorzaak zijn van onnauwkeurigheden.
Diverse onderzoekers proberen de werking en de voordelen van de generaliseerbaarheidstheorie aan te tonen. Shavelson e.a., (1989), Brennan en Johnson (1995) bijvoorbeeld geven aan welke componenten via de generaliseerbaarheidtheorie uit de geobserveerde scores te destilleren zijn en wat hoge of lage varianties van die componenten betekenen voor de meting (de zogenaamde G-studie). Met behulp van deze informatie kan de assessment geoptimaliseerd worden. Door te variëren met het aantal taken, het aantal beoordelaars of het aantal toetsmomenten kan men een schatting maken van de toetsbetrouwbaarheid bij verschillende meetprocedures en op deze manier een zo efficiënt mogelijke betrouwbare assessment construeren (de zogenaamde D-studie).
Een ander voorbeeld van de toepassing van de generaliseerbaarheidtheorie geven Fan en Chen (2000). Zij demonstreren dat interbeoordelaarsbetrouwbaarheid, gemeten met behulp van de klassieke betrouwbaarheidstheorie, vaak een overschatting is van de werkelijke betrouwbaarheid. De generaliseerbaarheidstheorie kan een veel nauwkeuriger waarde berekenen.

Specifieke criteria voor het evalueren van de kwaliteit van nieuwe toetsvormen
Naast een verruiming van de traditionele validiteits- en betrouwbaarheidscriteria kunnen er andere mogelijke criteria gesuggereerd worden voor het evalueren van de toetskwaliteit: de transparantie van de toetsprocedure, eerlijkheid, cognitieve complexiteit en authenticiteit van taken, en directheid van assessment (Haertel, 1991; Linn, e.a., 1991, Frederickson & Collins, 1989). Deze criteria werden ontwikkeld om de unieke kenmerken van nieuwe toetsvormen te belichten.

Een belangrijk kenmerk van nieuwe toetsvormen is het soort taken dat gebruikt wordt. Authentieke assessmenttaken bevragen hogere orde-vaardigheden, die een diepe leerbenadering bij studenten stimuleren. Een eerste criterium waardoor nieuwe toetsvormen zich onderscheiden van traditionele testen is dan ook de mate waarin de assessmenttaken het probleemoplossen, het kritisch denken en het redeneren meten. Dit criterium wordt cognitieve complexiteit genoemd (Linn, e.a., 1991). Of taken voldoen aan dit criterium kan bestudeerd worden door na te gaan of het oplossen van de taken dezelfde denkprocessen vereist als degene die experts gebruiken bij het oplossen van domeinspecifieke problemen. Daarnaast dient bewaakt te worden dat de geselecteerde taken 'nieuw' zijn voor studenten, in die zin dat studenten niet vertrouwdheid zijn met de betreffende problemen en de wijze waarop deze dienen opgelost te worden (Bateson, 1994).
Een ander criterium voor het evalueren van assessmenttaken is authenticiteit. Shepard (1991) omschrijft authentieke taken als 'de beste indicatoren voor het bereikt hebben van de leerdoelen'. Traditionele testen vereisen immers steeds een interpretatie vanuit studentantwoorden naar competentie in een specifiek domein. Bij authentieke assessment is interpretatie van de antwoorden niet nodig, omdat assessment reeds een directe uiting is van de competentie.
Het criterium 'authenticiteit van de taken' is nauw gerelateerd aan de 'directheid van assessment'. Powers e.a. (1994) beargumenteren dat de mate waarin docenten competentie onmiddellijk kunnen beoordelen relevante evidentie is voor de directheid van assessment. In hun onderzoek werd aan docenten gevraagd om een globaal oordeel te geven over de competentie 'algemene schrijfvaardigheid' van voorgelegde werken, zonder deze te scoren. Daarna werden deze gescoord volgens vastgelegde standaarden door getrainde beoordelaars. Resultaten tonen een duidelijke samenhang aan tussen de globale beoordeling van competentie door de docenten, en de toegewezen cijfers door de beoordelaars.
Bij het scoren van assessment speelt het criterium eerlijkheid een belangrijke rol. De vraag die hierbij centraal staat is of studenten een eerlijke kans hebben gehad om te tonen wat ze echt kennen en kunnen. 'Bias' kan enerzijds ontstaan doordat de taken niet aangepast zijn aan het niveau van het gegeven onderwijs of studenten niet vertrouwd zijn met de culturele inhoud die bevraagd wordt, anderzijds door vooroordelen in de beoordeling.
Een laatste criterium dat van belang is bij de evaluatie van assessment is de transparantie van de gebruikte beoordelingscriteria. Volgens Fredericksen en Collins (1989) vormt de mate waarin studenten zichzelf en medestudenten even betrouwbaar kunnen beoordelen als getrainde beoordelaars een goede indicatie voor het voldaan hebben aan dit criterium.

Hoewel deze criteria nieuwe kwaliteitsindicatoren lijken te zijn voor assessment, naast de reeds bestaande concepten validiteit en betrouwbaarheid, kunnen ze niet los gezien worden van de zes validiteitscriteria die door Messick reeds geformuleerd werden. Het verschil is, dat deze criteria een meer concrete invulling geven, die recht doet aan de unieke geaardheid van nieuwe toetsvormen en ook meer specifiek aangeeft hoe validiteit edumetrisch (in plaats van psychometrisch) kan nagegaan worden.

De criteria authenticiteit van taken en cognitieve complexiteit kunnen beschouwd worden als nadere aanvullingen bij Messick's inhoudelijk validiteitsaspect, dat de inhoud en het niveau van taken een adequate representatie dienen te zijn van reële problemen die zich voordoen binnen het kennisdomein dat gemeten wordt.
Voor het bestuderen van het criterium cognitieve complexiteit dient geanalyseerd te worden of het oplossen van assessmenttaken dezelfde denkprocessen vereist als degene die experts gebruiken bij het oplossen van domeinspecifieke problemen. Dit criterium komt overeen met wat Messick het 'substantieel validiteitsaspect' noemt.
De criteria directheid en transparantie zijn relevant om te bestuderen in het kader van de consequentiële validiteit van nieuwe toetsvormen. De wijze waarop competentie getoetst wordt, direct of via een interpretatie vanuit studentantwoorden, heeft immers een onmiddellijk effect op de aard en inhoud van het onderwijs en op het leerproces van studenten. Ook het al dan niet transparant zijn van de gebruikte beoordelingscriteria heeft een invloed op het leerproces van studenten. Immers, 'meeting criteria improves learning': indien studenten duidelijk weten welke criteria gebruikt worden bij het beoordelen van een opdracht zal dit hun prestatie verbeteren, omdat zij precies weten welke doelstellingen dienen bereikt te worden (Dochy, 1999).
Het criterium eerlijkheid maakt zowel onderdeel uit van het door Messick gedefinieerde 'inhoudelijk validiteitaspect', als van zijn 'intern structureel aspect'. Om studenten een eerlijke kans te geven om te tonen wat ze echt kennen en kunnen dient het aanbod van taken gevarieerd te zijn, zodat ze het hele spectrum van de competentie die gemeten wordt, bevatten. Daarbij is het belangrijk dat de criteria die gebruikt worden om een taak te beoordelen en het gewicht dat eraan wordt gegeven een juiste weerspiegeling zijn van de criteria die door experts worden gebruikt bij het beoordelen van competentie in een bepaald domein.

Op de vraag of er nu werkelijk een breuk is tussen de oude en nieuwe toetsingsstroom in de kwaliteitscriteria die gehanteerd worden voor een toets kan beargumenteerd worden dat er een duidelijk verschil in benadering en achtergrond is. Binnen de edumetrie krijgen de criteria validiteit en betrouwbaarheid een andere invulling.

Beoordeling van nieuwe toetsvormen volgens de 'nieuwe edumetrische benadering'

Indien we de belangrijkste veranderingen in het toetsingsveld ten aanzien van criteria voor het evalueren van assessment integreren, omvat het uitvoeren van een kwaliteitsanalyse de volgende stappen.

Evalueren van de
1 validiteit van de taken
2 validiteit van de beoordeling
3 generaliseerbaarheid van assessment
4 consequentiële validiteit van assessment.

Tijdens deze analyse zullen er argumenten gevonden worden die de constructvaliditeit van assessment ondersteunen of weerleggen.

Wat zijn de argumenten die de constructvaliditeit van nieuwe toetsvormen kunnen ondersteunen?

1 Validiteit van de gebruikte taken
Bij het bestuderen van de validiteit van de taken die gebruikt worden voor assessment zijn de volgende aspecten belangrijk. De taken die gebruikt worden dienen een adequate reflectie te zijn van het construct of, bij nieuwe toetsvormen, de competentie die gemeten wordt. Dit betekent naar inhoud toe dat de gebruikte problemen authentiek zijn, zodat ze een juiste weerspiegeling vormen van de wijze waarop kennis en vaardigheden door experts gebruikt worden. Hun cognitief niveau moet voldoende complex zijn, zodat ze dezelfde denkprocessen vereisen als degene die experts gebruiken bij het oplossen van problemen.
Nieuwe toetsvormen scoren beter op deze criteria dan gestandaardiseerde toetsen, precies omwille van hun authentiek en complex probleemkarakter.

2 Validiteit van de beoordeling
De volgende stap die dient onderzocht te worden is, of de beoordeling die gegeven wordt valide is. Het criterium eerlijkheid speelt hier een belangrijke rol. Dit veronderstelt enerzijds dat de criteria die gebruikt werden gepast zijn en juist gebruikt, dat wil zeggen een juiste reflectie zijn van de criteria die experts gebruiken en van het gewicht dat eraan wordt gegeven voor het beoordelen van competentie (intern structureel validiteitsaspect). Anderzijds betekent het dat studenten een eerlijke kans hebben om te tonen wat ze echt kennen en kunnen.

Mogelijke problemen die zich kunnen voordoen bij de beoordeling is enerzijds dat relevante beoordelingscriteria ontbreken, waardoor bepaalde competentieaspecten niet voldoende aandacht krijgen. Anderzijds kunnen er irrelevante, persoonlijke criteria worden toegepast bij het beoordelen. Doordat assessment de mogelijkheden van studenten op verschillende momenten en op verschillende wijze, door verschillende beoordelaars meet, is de kans dat deze problemen een rol spelen in de validiteit van de beoordeling klein. Mogelijke bias in de beoordeling wordt immers opgeheven. Hierdoor zal de totaliteit van de beoordeling bij assessment een meer nauwkeurig beeld geven van de werkelijke competentie van een student dan gestandaardiseerde toetsen, waar de beslissing of een student competent is wordt gereduceerd tot één beoordeling op één moment.

3 Generaliseerbaarheid van assessment

Deze stap in het validiteitsproces onderzoekt in welke mate assessment veralgemeend kan worden naar andere taken die hetzelfde construct meten. Dit is een indicatie van de mate van betrouwbaarheid van een score interpretatie en levert evidentie dat assessment ook werkelijk het bedoelde construct meet.

Problemen die zich kunnen voordoen zijn constructondervertegenwoordiging en constructirrelevante variantie. Constructondervertegenwoordiging betekent dat assessment te eng is, waardoor belangrijke constructdimensies niet gemeten worden. In het geval van constructirrelevante variantie is assessment te breed, en bevat hierdoor systematische variantie die irrelevant is voor het gemeten construct (Dochy & Moerkerke, 1997). In dit kader kan bediscussieerd hoe breed het construct of de bedoelde competentie dient gedefinieerd te worden, alvorens een interpretatie over compentie betrouwbaar en valide kan genoemd worden. Messick stelt dat de veralgemeenbaarheid van een interpretatie over de competentie van een persoon naar andere contexten waar de competentie gemeten wordt pas aantoont hoe stabiel en dus betrouwbaar deze interpretatie is. Frederick en Collins daarentegen doen afstand van het idee dat assessment pas betrouwbaar kan zijn, indien de interpretatie kan veralgemeend worden naar een breder domein. Zij hanteren een ander model waar de eerlijkheid (fairness) van de beoordeling cruciaal is voor betrouwbaarheid, maar generaliseerbaarheid niet.

In ieder geval kan beargumenteerd worden dat een assessment waarbij meerdere authentieke taken gebruikt worden die representatief zijn om een specifieke competentie te meten, minder gevoelig zijn voor constructondervertegenwoordiging of constructirrelevante variantie. Het bedoelde construct wordt immers direct gemeten. Authentiek betekent hier dan ‘realistisch’ zoals het werken met uitgebreid casusmateriaal in een Overall-toets, en niet bijvoorbeeld een beperkte casus waar alleen de relevante informatie wordt opgesomd.

4 Consequenties van assessment

Bij de laatste stap dient de vraag gesteld te worden wat de gevolgen zijn van het gebruik van een bepaalde toetsvorm voor het onderwijs en het leerproces van studenten. Consequentiële validiteit stelt de vraag of de consequenties van de assessment die men vaststelt in het onderwijs ook de bedoelde gevolgen zijn. Dit is een erg belangrijke vraag aangezien elke vorm van toetsing het leerproces ook gaat sturen. Dit kan in kaart gebracht worden door het voorleggen van statements van verwachte (en niet verwachte) gevolgen van de assessment aan de studentenpopulatie of door het afnemen van semi-gestructureerde ‘key group’ interviews. Via deze laatste methode komen ook vaker de onbedoelde effecten aan het licht. Bij consequentiële validiteit komen aspecten aan bod als: wat studenten begrijpen als vereisten voor de assessment; hoe de studenten zich voorbereiden; hoe er geleerd wordt; of de assessment gerelateerd is aan de authentieke taken; of de assessment studenten aanspoort tot het toepassen van kennis in realistische contexten; of de assessment de ontwikkeling van diverse vaardigheden in de hand werkt; of er langetermijneffecten gepercipieerd worden; of er daadwerkelijk inspanning wordt beloond, eerder dan geluk; of breedte en diepte in het leren wordt

beloond; of onafhankelijkheid wordt gestimuleerd door het expliciet maken van verwachtingen en criteria; of er relevante feedback wordt voorzien over de voortgang; of er eerder competenties worden gemeten, eerder dan memoriseren van feiten.

Conclusie

In dit artikel werd betoogd dat studentgericht onderwijs een nieuwe vorm van toetsing impliceert. 'Assessment' is dan ook niet meer het traditioneel toetsen van studenten. Nieuwe assessmentvormen wijken veelal op meerdere punten van deze kenmerken af van het traditionele 'testen'. Studenten worden beoordeeld op basis van hun actieve prestatie om kennis te gebruiken op een creatieve manier en om competentie te demonstreren. De uitgangssituaties zijn daarbij veelal reële problemen, authentieke representaties van problemen in de werkelijkheid. Studenten krijgen ook vaak zelf verantwoordelijkheid in het toetsproces en zowel kennis als vaardigheid wordt gemeten. Voorts zijn assessmentvormen, in tegenstelling tot traditionele testen, vaak niet gestandaardiseerd. Ten slotte houdt assessment veelal in dat op verschillende momenten, door verschillende beoordelaars diverse vaardigheden en kennis worden getoetst. Bij de nieuwe assessmentvormen luidt de belangrijkste vraag als het gaat om kwaliteit: in welke mate is de beslissing 'of iemand wel of niet competent is' terecht?

Door deze specifieke karakteristieken is het toepassen van traditionele testtheoretische kwaliteitscriteria niet voor de hand liggend. In dit artikel is daarom ook aandacht besteed aan de gevolgen van deze evoluties voor het screenen van de edumetrische kwaliteit van assessment. Er wordt geconcludeerd dat de traditionele testtheoretische betekenis van validiteit en betrouwbaarheid niet langer hanteerbaar zijn, maar minstens uitgebreid (c.q.vervangen) dienen te worden.
Bij assessment is het belangrijk volgende kwaliteitsaspecten in beeld te brengen: transparantie van de assessmentprocedure, eerlijkheid, cognitieve complexiteit, authenticiteit van taken, en de invloed van de assessment op het onderwijs.

Literatuur

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1985) *Standards for educational and psychological testing*. Washington, D.C.
Bateson, D. (1994) Psychometric and Philosophic Problems in 'Authentic' Assesment, Performance Tasks and Portfolios. *The Alberta Journal of Educational Research,* 11, (2), 233-245.
Bennet, Y. (1993) Validity and reliability of assessments and self-assessments of work-based learning assessment. *Assessment & Evaluation in Higher Education*, 18, (2), 83-94.

Birenbaum, M. (1996) Assessment 2000: Towards a pluralistic approach to assessment. In M. Birenbaum & F.J.R.C. Dochy (Eds.), *Alternatives in Assessment of Achievements, Learning Processes and Prior Knowledge.* (3-30) Boston: Kluwer Academic Publishers.

Birenbaum, M. & Dochy, F. (Eds.) (1996) *Alternatives in Assessment of Achievement, Learning Processes and Prior Knowledge*. Boston: Kluwer Academic.

Boud, D. (1990) Assessment and the promotion of academic values. *Studies in Higher Education*, 15, 101-111.

Brennan, R.L., & Johnson, E.G. (1995) Generalizability of performance assessments. *Educational Measurement: Issues and Practice,* 11, (4), 9-12.

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

Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. (1972) *The dependeability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Cronbach, L.J., Linn, R.L., Brennan, R.L., & Haertel, E.H. (1997) Generalizability analysis for performance assessments of students achievement or school effectiveness. *Educational and Psychological Measurement*, 57, (3), 373-399.

De Corte, E. (1990) *A State-of-the-art of research on learning and teaching*. Keynotlecture presented at the first European Conference on the First Year Experience in Higher Education, Aalborg University, Denmark, 23-25 April 1990.

Dochy, F. (1999) *Instructietechnologie en innovatie van probleemoplossen: over constructiegericht academisch onderwijs*. Utrecht: Lemma.

Dochy, F., Heylen, L., & Van de Mosselaer, H. (2000) *Cooperatief leren in een krachtige leeromgeving: Handboek probleemgestuurd leren in de praktijk.* Leuven/Leusden: Acco.

Dochy, F.J.R.C., & de Rijke, T.R. (Red.) (1995) *Assessment centers: Nieuwe toepassingen in opleiding, onderwijs en HRM.* Utrecht: Lemma.

Dochy, F., & Moerkerke, G. (1997) The present, the past and the future of achievement testing and performance assessment. *International Journal of Educational Research*, 27, (5), 415-432.

Dochy, F., Segers, M., & Sluijsmans, D. (1999) The use of self-, peer and co-assessment in higher education: a review. *Studies in Higher Education,* 24, (3), 331-350.

Fan, X., Chen, M. (2000) Published studies of interrater reliability often overestimate reliability: computing the correct coefficient. *Educational and Psychological Measurement,* 60, (4), 532-542.

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

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

Heller, J.I., Sheingold, K., & Myford, C.M. (1998) Reasoning about evidence in portfolios: cognitive foundations for valid and reliable assessment. *Educational Assessment, 5,* (1), 5-40.

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

Kessels, J.W.M. (2000a) De academie in bedrijf. De omstreden dualisering van het wetenschapelijk onderwijs. *Opleiding & Ontwikkeling*, 13, (3), 33-39.

Kessels, J.W.M. (2000b) Kennismanagement: een anachronisme. In: Halbertsma, E.H. (red.) (2000) *Dilemma's te lijf. Management van mensen en organisaties,* 53-58. Assen: Van Gorcum/De Baak.

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

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

Powers, D., Fowles, M., & Willard A. (1994) Direct assessment, direct validation? An example from the assessment of writing? *Educational assessment, 2*, (1), 89-100.

Segers, M., & Dochy, F. (1999) Een nieuw onderwijsmodel voor het hoger onderwijs in theorie en praktijk. In: M. Lacante, P. De Boeck, *Meer kansen creëren voor het Hoger Onderwijs. Handboek Leerlingen-begeleiding,* 153-180. Dordrecht: Kluwer.

Shavelson, R.J., Webb, N.M., & Rowley, G.L. (1989) Generalizability Theory. *American Psychologist,* 44, (6), 922-932.

Shepard, L. (1991) Interview on assessment issues with Lorie Shepard. *Educational Researcher*, 20, (2), 21-23.

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

Suen, H.K., Logan, C.R., Neisworth, J.T., & Bagnato, S. (in press) Parent-professional congruence. Is it necessary? *Journal of Early Intervention.*