Wat is er mis met beoordelen?

H. Tillema

Samenvatting

"Waarop letten assessoren als ze beoordelen?"; "Wat is de kwaliteit van nieuw in te zetten beoordelingsinstrumenten?"; "Hoe kunnen we (in de opleiding dan wel in het beroep) blijvend werken aan kwaliteitsverbetering in het beoordelen van docenten?" Deze en andere vragen klinken door in de vier bijdragen van dit themanummer. Centraal staat daarin de vraag naar de validiteit rond docentenbeoordelingen. In deze discussiebijdrage wil ik laten zien dat er nogal wat kwesties spelen rond beoordelen van docenten. Om dit te kunnen doen neem ik een ruimer perspectief om duidelijk te kunnen maken welke bijdrage elk van de artikelen in dit themanummer in het bijzonder levert aan het opbouwen en verbeteren van de kwaliteit in beoordelingspraktijken van (aanstaande) docenten. "Er is niets mis met beoordelen zolang we weten waarover we praten" zou daarbij het motto kunnen zijn.

1 Beoordelen gaat ergens over

Niet alleen in de wereld van leraren(opleidingen), maar ook in andere beroepsvelden is men geruime tijd bezig met de constructie, de invoering en de beproeving van nieuwe vormen van beoordelen (Grootendorst & Tillema, 2002) en het vinden van andere methoden van beoordelen rond competent handelen van professionals (Snoek, 2002). Deze ontwikkeling, die een aantal jaren terug met enthousiasme is ingezet (Tillema, 2003), heeft niet alleen nieuwe vragen over kwaliteit en deugdelijkheid rond het beoordelingsproces opgeroepen (Onderwijsraad, 2004), maar heeft ook enkele lastige dilemma's opgeleverd over wat men rekent tot competent handelen (Ben Peretz, 2001; Gilroy, Edwards, & Hartley, 2002). Dit is bijvoorbeeld te zien in het dilemma van de lerarenopleider die, enerzijds als begeleider en anderzijds als beoordelaar, gevangen is tussen formatief en summatief beoordelen, tussen certificeren en ontwikkelen (Heilbronn, 2003). Vanuit een *accountability*-perspectief (Cochran-Smith, 2001) is de beoordelaar gehouden om extern te verantwoorden wat de kwaliteit is van functioneren, vastgelegd in redelijk uniforme standaarden (SBL, 2004). Vanuit een ontwikkelingsperspectief dient beoordelen juist primair gericht te zijn op het ondersteunen en stimuleren van professionele groei (in het beroep dan wel de beroepsvoorbereiding). De vraag bij het beoordelen van (aanstaande) docenten is dan ook hoe men twee heren kan dienen.

Ik noem dit dilemma om te laten zien dat vraagstukken die verband houden met de kwaliteit van beoordelen niet alleen instrumenteel-technisch van aard zijn of enkel betrekking hebben op procedures, maar plaatsvinden in een context. Een eerste indruk die kan ontstaan bij lezing van de bijdragen is dat zij ver af lijken te staan van de discussie die nu plaatsvindt in en rond beoordelen van leraren (in opleiding) (Cochran Smith & Zeichner, 2005). In dit verband is het illustratief te letten op wat Marieke Dresen, lerarenopleider van Fontys Hogescholen, in een e-mail naar mij opmerkte (18 mei 2006). Haar verslag biedt een realistische context van problemen die spelen in het beoordelen van docenten (in het citaat, toegespitst op het portfolio; cursivering, auteur).

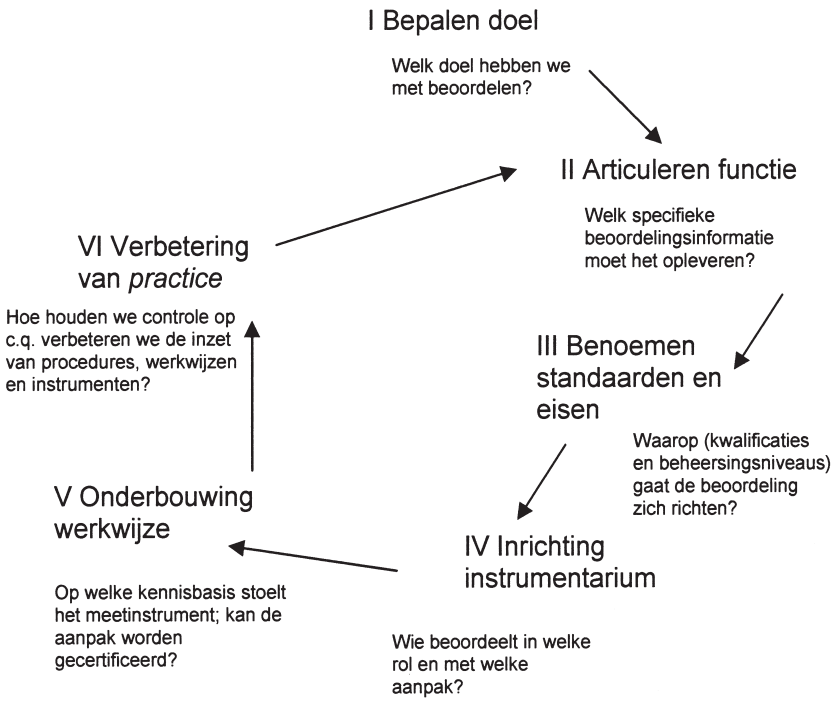
> Als ik kijk naar mijn eigen functioneren gedurende de laatste jaren en mezelf dan beperk tot het beoordelen van portfolio's dan heb ik, naast fantastische ervaringen, toch verschillende problemen ondervonden. Ik zal in eerste instantie spreken over mijn functioneren in de deeltijd waar ik gedurende zes jaar tegelijkertijd de rol van begeleider en beoordelaar bij studenten had.
> Allereerst kwam voor mij het probleem van *het soort portfolio*. Gedurende de jaren negentig had men het over ontwikkelingsgericht portfolio, reflectieport-

folio, beoordelingsportfolio, startportfolio, waarbij het onderscheid niet altijd even duidelijk was en waarbij aangemerkt dient te worden dat de verschillende soorten portfolio's ook niet los gezien kunnen worden. Toch werd ik geacht als hogeschooldocente er 'iets' over te zeggen. Afhankelijk van het soort portfolio trof ik ook andere producten aan. Om een voorbeeld te geven: bij ontwikkelingsgerichte portfolio's waren er meer reflecties, persoonlijke verslagen, zelfevaluaties etc. Dus verschillende portfolio's *voor verschillende doeleinden* met verschillende producten waarop ik als hogeschooldocente adequaat diende te reageren. Er ontstond ook nog een discussie omtrent de *mate van voorstructurering*, m.a.w. hoeveel ruimte krijgt de student? Was er sprake van een open portfolio of een gesloten portfolio? Lagen er gerichte beoordelingscriteria of wat globalere? Tot overmaat van ramp kwam toen de vraag: in *welke rol* moest er gereageerd worden op het portfolio? Als begeleider, als assessor of als mentor?

Het was een periode van onzekerheid maar tegelijkertijd ook een periode van zinvolle dialogen met studenten. Samen met hen werden de *vereiste bekwaamheden* besproken, zodat er zicht was op wat er van studenten gevraagd werd. Samen met hen keken we naar *geschikte bewijzen* om aan te tonen dat men aan die bekwaamheid voldeed. Er lag *een criterialijst* waarmee de studenten zelf konden bepalen of het portfolio voldeed aan de eisen. Deze lijst werkte met een score van 1 tot en met 10. De dialogen met studenten waren voor mij zeer verhelderend. Ik kreeg daardoor inzicht in de onduidelijkheden, hun werkwijze en noem maar op. Studenten gaven ook aan dat de gesprekken hen meer inzicht hadden gegeven in wat er geëist werd maar tegelijkertijd ook inzicht in hun eigen functioneren, bijvoorbeeld hun reflectievermogen of hun planningsvaardigheden.

Toen kwam het moment waarop ik de portfolio's moest beoordelen. De criterialijst werd eerst niet alleen door de student zelf, maar ook door één (of meerdere) medestudent(en) *gescoord* en daarna werd het portfolio bij mij ingeleverd. Ik moet toegeven dat ik het moeilijk vond om objectief te zijn, want ik wist wat de student allemaal had gedaan en waarom. De oplossing meende ik toen gevonden te hebben in een tweede beoordelaar. Onafhankelijk van elkaar bekeken we het portfolio en scoorden de criterialijst. We verschilden wel eens maar kwamen altijd tot *overeenstemming,* waarbij gezegd moet worden dat de tweede beoordelaar de student niet kende. Een andere oplossing meende ik te vinden in het laten zien van good practice, dus portfolio's die een goede beoordeling hadden gekregen. Zodoende hoopte ik de studenten een *richting te kunnen geven*. Na de beoordeling van het portfolio was er altijd een gesprek, mede om te kunnen vaststellen dat die student ook echt de maker is van het portfolio. Gedurende het gesprek is het enkele malen voorgekomen dat de student in kwestie zelf vroeg om zijn portfolio aan te vullen, omdat hij (of zij) het zelf wel erg mager of niet voldoende vond. Aangezien ik vond dat de beoordeling ook een leermoment voor de student zou moeten zijn, werkte ik met een korte rapportage waarin de gescoorde criterialijst zat met daarnaast een verslag van onze gezamenlijke bevindingen met sterke punten en ontwikkelpunten.

Het verslag van Marieke Dresen mag verduidelijken dat in het beoordelen keuzen moeten worden gemaakt, die om een positiebepaling vragen over wat men met beoordelen wil. Met name in de empirische bijdragen van dit themanummer is het gemis voelbaar van een duidelijke positiebepaling rond de vraag waartoe beoordelen dient. Vragen rond validiteit van beoordelingsinstrumenten houden immers direct verband met doelrealisering in relatie tot ingezette middelen (Moss, 2005; Preedy, Glatter, & Wise, 2002). Anders gesteld, kwaliteit in beoordelen van docenten heeft betrekking op vraagstukken rond geschiktheid van een beoordelingssystematiek voor specifieke doeleinden. Eerder heb ik (Tillema, 2003) gepleit voor een integrale benadering van docentenbeoordeling, bijvoor-



Figuur 1.

Stappen in een auditproces naar de deugdelijkheid van beoordelingspraktijken.

beeld door het (laten) uitvoeren van audits naar de deugdelijkheid van beoordelingspraktijken. Een weergave van dit totale proces van keuzen rond kwaliteit van beoordelen is te vinden in Figuur 1.

De aspecten die Mieke Dresen noemt (in cursivering) laten zich goed plaatsen binnen het schema. In een dergelijk auditproces, zoals weergeven in Figuur 1, gaat het om het bepalen van de deugdelijkheid van een proces door (beoordelings)doelen in relatie te bezien tot de (assessment)middelen die worden ingezet. Op deze manier beschouwd zijn de bijdragen aan dit themanummer, elk op een eigen wijze, doende de vraag naar deugdelijkheid van beoordelen te onderzoeken en oplossingen aan te dragen voor onderkende problemen in het beoordelen van docenten. Het stappenschema van het auditproces in Figuur 1 kan daarom houvast bieden bij het positioneren van de studies. Verder kan het hopelijk de discussie rond de vele aspecten van kwaliteit die spelen in beoordelen meer toespitsen en enigszins concreter maken.

2 Beoordelen – een kwestie van articulering

Plaatsing van de artikelen in het schema van Figuur 1 laat zien dat de drie empirische studies met name handelen over de instrumentering en onderbouwing van beoordelingspraktijken (stappen IV en V in de audit), terwijl de notitie van Roelofs vooral betrekking heeft op het articuleren van functie en verbetering van beoordelingspraktijken (stappen II en VI) (zie Tabel 1).

Ook is met behulp van Figuur 1 duidelijk dat de empirische bijdragen niet zozeer ingaan op de inhoudelijke kant van beoordelen, namelijk stap III (bepaling van standaarden en eisen), noch in het bijzonder een praktische bijdrage (willen) leveren aan de uitvoering of hantering van beoordelingspraktijken (stap VI). Het artikel van Roelofs wil, door aandacht te vragen voor doel en functie (stappen I en II), juist een kader bieden voor instrumentering en werkwijzen. Is er wat mis met deze ongelijke aandacht in de artikelen

Tabel 1

Bijdragen in dit themanummer rond beoordelen opgevat als audit

Onderdeel van het auditproces dat expliciet aan de orde is	In de bijdrage van
IV inrichting instrumentarium V onderbouwing werkwijze	M. van der Schaaf, K. Stokking en N. Verloop
IV inrichting instrumentarium V onderbouwing werkwijze	M. Bakker, P. Sanders, D. Beijaard, E. Roelofs, D. Tigelaar en N. Verloop
IV inrichting instrumentarium V onderbouwing werkwijze	M. Nijveldt, M. Brekelmans, D. Beijaard, Th. Wubbels en N. Verloop
II Articuleren functie III Standaarden VI verbetering practice	E. Roelofs

voor vraagstukken van kwaliteit in beoordelen? In het navolgende wil ik bezien hoe de artikelen, vanuit een integrale benadering van kwaliteit, elk hun bijdrage leveren aan de deugdelijkheid van docentenbeoordeling.

2.1 Doel en functie van beoordelen – stappen I en II

De vraag die bij docentenbeoordeling voorop moet staan, is die naar de vaststelling van wat men precies met beoordelen wil, de *shared vision* rond beoordeling (Preedy et al., 2002; Wilson & Youngs, 2005); een vraag die weinig of niet pregnant wordt gesteld in de empirische bijdragen van dit themanummer. Het gaat hierbij om profielbepalende keuzen ten aanzien van:

- Wat we willen met beoordelen, dus waarvoor het dient, en
- Welke focus en welk accent we daarbij aanbrengen (ontwikkelingsgericht of geschiktheidbeoordeling).

Zowel in de bijdrage van Bakker en collega's als die van Van der Schaaf e.a. en Nijveldt e.a. is impliciet gehouden waarvoor en met welk doel portfolio's dan wel video(dossiers) worden beoordeeld. Het maakt echter een wezenlijk verschil of men validering van instrumenten onderneemt voor een formatieve dan wel een summatieve functie (Tillema & Smith, 2007). Zo valt bijvoorbeeld op dat de beoordeelde docenten niet zelf zijn betrokken in het beoordelingsproces van hun porfolio dan wel hun eigen videodossier. Het zijn externe beoordelaars (andere docenten, en leerlingen) die beoordelen in de onderzoeken van de genoemde drie bijdragen. De vraag is waarom de beoordelingen van deze beoordelaars meer legitiem zouden zijn dan die van de docenten zelf. Het vermoeden rijst dat de onderzoekers enkel een summatieve rol voor hun instrumenten in docentenbeoordeling zien weggelegd. Toch zijn de consequenties van de keuze voor een of ander doel van beoordelen verregaand. Ze werken door in de volgende stappen van het beoordelingsproces: standaardsetting; instrumentering (met vaste richtlijnen en bewijsverzameling tegenover persoonlijke, situatieve inrichting van afnames); scoring (open versus gesloten) en validering (met consequenties voor *ability testing* tegenover *monitoring progress in development*; Moss, 2005). Daarom kan validering van instrumenten, zoals in de empirische bijdragen aan de orde is, niet los worden gezien van het doel van beoordelen. De bijdrage van Roelofs maakt dit duidelijk, wanneer hij pleit voor een samenhangende beoordeling van kritische beroepstaken. Toch kan ook hier de vraag gesteld worden in wiens dienst de beoordeling staat. Zo valt bijvoorbeeld moeilijk in te zien hoe assessoren middels 'hun' interpretatieve benadering recht kunnen doen aan niveaus van competent handelen die een docent wil nastreven om zo 'oorzaken van adequaat of minder adequaat handelen' te beoordelen.

Wel beschouwd heeft geen van de bijdragen een expliciet standpunt ingenomen ten gunste van een op leren en ontwikkeling gerichte manier van beoordelen, dienend ter verbetering van docenthandelen in het beroep. Er is genoeg reden om een dergelijke positie sterk te benadrukken in het beroep en de opleiding van leraren (Stiggins, 2002). In de *state of the art*-studie van de AERA (in het

volumineuze werk: “Studying teacher education” door Cochran Smith & Zeichner, 2005), in de *position paper* van de EARLI-assessment groep (Birenbaum, Breuer, Cascallar, Dochy, Dori, Ridgway, Wiesemes, & Nickmans, 2006), in het OECD-rapport “Improving learning” (2005) en in het werk van de Assessment Reform Group (2006) zien we dat er nadrukkelijk aandacht wordt gevraagd voor beoordelen (assessment) als *tool for learning*. Hier wordt nadrukkelijk onderkend hoe beoordelen kan bijdragen aan (beroeps)vorming. Op een internationale conferentie over assessment in Portland, Oregon, September 2005, is nagegaan hoe beoordelen daadwerkelijk kan bijdragen tot het creëren van betekenisvolle leerervaringen van docenten in hun beroepscontext. Linda Allal (University of Geneva), Janet Looney (OECD, Paris), Kari Smith (Universiteit Bergen), Harm Tillema (Universiteit Leiden), en Joke Voogt (Universiteit Twente) hebben deze positiebepaling in een brochure over *powerful assessment* samengevat (zie Tabel 2).

Deze uitgangspunten (zie: *www.assessment-reform-group.org.uk*.) kunnen bijdragen aan een expliciete positionering van beoordelen, dat wil zeggen aan de bepaling van standaarden, het onderbouwen van functie en doel van procedures en een gerichte benutting van beoordelingsuitkomsten; iets waar de gerapporteerde studies in dit themanummer mijns inziens (te) weinig aan refereren. Het zou in elk geval de (*consequential*) validiteit van onderzochte instrumenten ten goede komen, wanneer deze kunnen beantwoorden aan bovengenoemde criteria; in plaats van, zoals nu, de instrumenten doel- of perspectief neutraal op te vatten.

2.2 Onderbouwen van instrumenten en werkwijzen – stap IV

Beoordelen vindt niet plaats in een vacuüm, maar brengt verantwoordelijkheid met zich mee, omdat het kan leiden tot ver reikende beslissingen over (studie)loopbaan dan wel professionaliteit van (aanstaande) leraren. Veel te weinig nog wordt onderkend dat er legitimering nodig is van gehanteerde beoordelingspraktijken en dientengevolge van een weloverwogen keuze voor bepaalde beoordelingsinstrumenten (Delandshere & Petroski, 1998; Van Minden, 2002; Task Force on Assessment Centres, 2000;). Een primaire zorg is dan ook: wat is eigenlijk de kennisbasis waarop het beoordelen stoelt?

In een audit zou dit een aantal concrete vragen betreffen, zoals:

- Oordelen assessoren betrouwbaar en vanuit en gemeenschappelijk referentiekader?
- Zijn de procedures uitgeschreven en herhaalbaar voor anderen?
- Zijn er gegevens verzameld over het gebruik van de assessment instrumenten?
- Wordt bijgehouden hoe assessoren tot een afgewogen advies komen?
- Wat is de aard van de feedback die kandidaten ontvangen?

Dergelijke valideringsvragen zijn met name aan de orde in de artikelen van Nijveldt en collega’s, Bakker e.a. en Van der Schaaf e.a. Hun bijdragen bevatten waardevolle uitkomsten over kwaliteit (mogelijkheden en beperkingen) van instrumenten en de inzet ervan in beoordelen van docenten. De studies geven een redelijk positief beeld over de deugdelijkheid van onderzochte beoordelingsinstrumenten.

Bakker en collega’s. hebben beoordelaars van videodossiers een beoordelingskader aangeboden om het interpreteren en het beoordelen volgens persoonlijke constructen en criteria zoveel mogelijk uit te sluiten. Zij concluderen dat beoordelaars op een acceptabel niveau competenties van docenten kunnen beoordelen, hoewel sommige beoordelaars extreme scores toekennen. Het ontwikkelde beoordelingskader (met competentieniveaus, scoringsvoorschriften, training) heeft daarbij over het algemeen een positieve invloed gehad.

Van der Schaaf en collega’s zijn nagegaan wat de relatie is tussen opvattingen van docenten over hun competenties, zoals neergelegd in hun portfolio, en hun gedrag afgezet tegen dat gedrag, zoals beoordeeld door leerlingen en externe beoordelaars. Hoewel zij tot de conclusie komen dat er geen duidelijke relatie is tussen de opvattingen van de docenten en hun gedrag, blijken de beoordelingen door leerlingen van het gedrag van hun docent goed te sporen met die van de externe beoordelaars.

Nijveldt en collega's zijn nagegaan hoe gezamenlijke beoordeling van assessoren rond eenzelfde docent/competentie al dan niet leidt tot een coherente beoordeling. Uit hun onderzoek blijkt dat het moeilijk is voor beoordelaars om in de vaststelling van de eindbeoordeling elkaar kritisch aan te vullen dan wel uit te dagen, zodat de onderzoekers

Tabel 2

Een op ontwikkeling en leren gerichte oriëntatie op beoordelen

Gelet op het doel van beoordelen		Gelet op de manier van beoordelen	
Assessment can make a powerful contribution to effective teaching and successful student learning. It should:		Student teachers, teachers and teacher educators (pre-service, in-service, professional development activities for teachers and teacher educators), need to be given ample opportunity, time and guidance in order to:	
1	focus on how students learn	1	master the principles of assessment and learn how to integrate assessment in the learning/teaching process
2	be included in the planning of teaching and learning	2	experience the power of assessment as learners in teacher education programs
3	be an integral part of classroom practice	3	implement various forms of assessment and develop personal conceptions of assessment practices that can empower learning
4	be aimed at significant educational goals and objectives	4	undertake diversified activities that allow critical reflection about assessment (analysis of examples of students' work, self-observation through video-recording, reciprocal observation with peers and colleagues)
5	recognize the full range of all students' achievements	5	discuss beliefs and attempts to implement assessment, with the support of experienced and knowledgeable mentors
6	promote shared understanding of learning goals and assessment criteria	6	carry out informed, evidence-based professional decisions about assessment having lasting positive effects on student learning and achievements
7	be constructive and sensitive to students' feelings	7	foster the common goal of improving learning for each and every learner across the different levels of the educational system (preschool, primary, secondary, vocational)
8	take into account student diversity	8	explain the power of assessment to stakeholders in the educational system
9	foster student motivation for learning		
10	encourage positive relations and productive exchanges among students and between teachers and students		
11	entail forms of regulation (appropriate feedback, differentiated guidance, adaptive instructional activities) that promote student learning		
12	encourage active student involvement in assessment (self-assessment, peer and teacher and student co-assessment)		
13	be based on varied sources of information (written/oral, individual/group, quantitative/qualitative)		
14	situate student learning outcomes with respect to educational objectives		
15	be communicated in a transparent and coherent way to concerned parties (students, other teachers, students' parents)		
16	enhance students' capacity to undertake both independent learning and shared learning in group settings		

moeten concluderen dat afwegingen in het beoordelingsproces (inbrengen van tegenbewijs en/of alternatieve interpretaties) maar moeilijk te realiseren zijn in de praktijk.

Zonder twijfel hebben deze studies elk belangwekkende resultaten opgeleverd, die bijdragen aan een valide kennisbasis en verantwoorde inzet van instrumenten. De resultaten stemmen echter ook tot bezorgdheid bij nadere beschouwing.

In het onderzoek van Bakker e.a. is geprobeerd om het interpreteren en het beoordelen volgens persoonlijke constructen en criteria zoveel mogelijk uit te sluiten door standaardiseren en aanbieden van meerdere assessmenttaken. Ook zijn aan de praktijk aangepaste beoordelingscriteria aangeboden. Toch blijken persoonlijke voorkeuren – collega-docenten als beoordelaar geven extremere beoordelingen – de overhand te hebben, ondanks de voorstructurering door de onderzoekers. Bovendien blijkt er tussen beoordelaars over het algemeen een laag niveau van overeenstemming te zijn in hun eindbeoordeling.

In het onderzoek van Van der Schaaf e.a. blijkt de inhoud van docentportfolio's (hun neergelegde doelen) niet adequaat door externe beoordelaars gebruikt te worden. Integendeel, zij maken gebruik van eigen schemata om de opvattingen te beoordelen die docenten in hun portfolio neerleggen. Dit terwijl het portfolio instrument juist bedoeld is om recht te doen aan het perspectief van de portfolio-ontwikkelaar.

In het onderzoek van Nijveldt e.a. komt naar voren dat, ondanks een gemeenschappelijk beoordelingskader, beoordelaars niet komen tot een gezamenlijk gedragen eindevaluatie. Kennelijk durven ze niet kritisch te staan ten opzichte van het eigen oordeel en regresseert sociale druk de beoordeling naar een 'gemiddelde'.

Conclusie van de drie onderzoeken lijkt te moeten zijn dat er iets mis is met beoordeling door beoordelaars. Deze conclusie staat niet op zich; elders in onderzoek naar beoordeling van kwaliteit van instructie (Clausen, 2007) zien we eenzelfde resultaat: verschillende beoordelaars beoordelen verschillend over eenzelfde proces (ondanks pogingen tot standaardisatie). Kennelijk heeft elke assessor eigen perspectieven, observaties en criteria om eenzelfde handeling of gebeurtenis te waarderen (Kane, 2006). Dit gegeven kan evenwel ook positief worden bezien, zoals bijvoorbeeld gebeurt in *multi-rater*-feedback beoordelingen (360 graden feedback; Jellema, 2003; Lievens, 1998). Dus niet overeenstemming, maar inclusie van diversiteit in de *overall assessment rating* zou dan een optie kunnen zijn.

2.3 Standaarden en criteria in beoordeling – stap III

De aansluiting bij het auditproces mag ook duidelijk maken dat bij ontbreken van standaarden en criteria een overeenstemming of afstemming in beoordeling zal ontbreken. Gelet op wat men met beoordelen wil (stappen I en II) is het daaropvolgend nodig standaarden aan te geven, dus de eisen te benoemen die men hanteert in het uitvoeren van een beoordelingstraject. De centrale vraag is hierbij: Wat is een acceptabele toetssteen in de beoordeling? Het gaat in deze stap om de bepaling van:

- het niveau en de graad van detaillering waarmee men wil beoordelen;
- de standaarden die zijn aanlegd in het beoordelingstraject (en van welk niveau), bijvoorbeeld rond assessoren training, afnameprocedures, terugrapportage en advies, en
- de eisen die zijn gesteld aan de afname en instrumentering.

De drie empirische studies hebben, bij ontstentenis van een in de beoordelingspraktijk gegronde standaard, zelf een voor het onderzoek geschikte criteriumbepaling moeten ondernemen. Het heeft geleid tot een, door onderzoekers bepaalde, lijst van indicatoren waaraan beoordelaars zich al of niet committeerden. Interessant is dan ook te zien hoe de bijdrage van Roelofs, die immers een procesmodel aanbiedt, kan voorzien in de bepaling van competent handelen vanuit een gemeenschappelijk referentiekader. Kern van zijn aanpak is het beoordelen van prestaties (de 'gevolgen van handelen') bij complete taken uit een taakdomein op een aantal kwaliteitscriteria en via een interpretatieve redenering afleiden wat verantwoorde handelwijzen en afwegingen van docenten zijn. Kwaliteitscriteria dienen de wenselijke gevolgen van

handelen op een taakdomein te benoemen, hetgeen door waarderend rapporteren wordt teruggekoppeld. Met andere woorden: de basis van docentenbeoordeling is een 'protocolisering' van het handelen. Criteria voor 'correct' handelen zijn dan te ontlenen aan een (de?) professionele kennisbasis van het onderwijzen. Klemmende vraag is natuurlijk of er reeds een 'samenhangend procesmodel van competent handelen' is dat kan beschrijven:

- welke processen van denken en handelen een (aankomend) professional met succes uitvoert;
- hoe competenties worden ingezet op taakniveau;
- welke taakuitvoering in het beroep aan de orde is, en
- welke vormen van bewijsvoering moet worden verzameld om competente taakuitvoering aan te tonen.

Vooral het ontbreken van een bestaand, door onderzoek geleid, 'protocol' van competent handelen in diverse, complexe beroepssituaties maakt de agenda van Roelofs' voorstellen onzeker, zo het ooit zal komen tot een dergelijke professionele kennisbasis van onderwijzen (Gilroy, et al., 2002). Men neme ter illustratie de middelste kolom uit Tabel 1 van zijn bijdrage om al snel te ontdekken dat juist de andere bijdragen uit dit themanummer materiaal aandragen die de moeilijkheid en meer nog de onwaarschijnlijkheid van specificatie hebben aangetoond. Dit wil niet zeggen dat het benoemen van standaarden of het stellen van criteria ondoenlijk is. Terecht wijst Roelofs in zijn bijdrage op de noodzaak tot specificatie van competent handelen en met recht wijst hij op het gevaar van losstaande 'checklijstjes' van competenties (vgl. ook Tillema, 2004). Punt is echter: wie is bevoegd en bekwaam om een integraal beeld van competent handelen te formuleren en wat is de positie van onderwijskundig onderzoek daarbij (Cohran-Smith & Zeichner, 2006; Gilroy, et al., 2002).

3 Met perspectief beoordelen

Een vraag die niet onbesproken kan blijven is of beoordelen van docenten en dat van aanstaande docenten een en hetzelfde is. Eigenlijk valt dit onderscheid, dat dwars door dit themanummer loopt, grotendeels samen met het dilemma van de lerarenopleider/assessor dat aan het begin van mijn bijdrage is genoemd, namelijk die tussen ontwikkelen en certificeren. In het geval van beoordelen vanwege een *accountability*perspectief is de vraag aan de orde welke docent geschikt is voor (stadia in) het beroep, waarbij rekenschap valt te geven van de verworven kwalificaties op enig moment. Bij beoordelen vanuit een ontwikkelingsperspectief gaat het er juist om geschikte feedbackinformatie te verzamelen die (bij)sturend kan werken voor te nemen vervolgstappen in leren en groei tijdens het leraarschap.

Het onderscheid is van groot belang voor de inrichting van beoordelingstrajecten (Zeichner & Wray, 2002), omdat men in beide gevallen toetst met eigen, voor dat doel geschikte, instrumenten (Dottin, 2001; Smith & Tillema, 2003). Vermenging van beoordelingsdoelen of althans onduidelijkheid daarover bij degene die beoordeeld wordt, dan wel degene die een selectiebeslissing moet nemen, kan ernstige gevolgen hebben – niet alleen in termen van motivatie tot leren (ontvangen van al dan niet adequate feedback), maar ook in termen van de kwaliteit van geleverde prestaties (voldoen aan standaarden) (Heilbronn, 2003; Zuzowsky & Libman, 2002). Formatief en summatief beoordelen zijn twee trajecten met elk hun eigen eisen aan beoordelingsinstrumenten; de bijdragen in dit themanummer adstrueren dit nogmaals. Want, formatief beoordelen vraagt om een integratie van opleiden en beoordelen, summatief beoordelen niet. Formatief beoordelen richt zich op feedback en ondersteuning na assessment; summatief beoordelen niet. En instrumenten die geschikt zijn voor summatief beoordelen, zijn dat niet zondermeer voor formatief terugkoppelen (bijv. 360 graden feedback, Jellema, 2004; of de gekozen portfoliovariant, Tillema & Smith, 2007). Kwaliteitscriteria voor beide manieren van beoordelen zijn dan ook verschillend (Darling Hammond & Bransford, 2005). Ingeval van certificerend beoordelen is dit taak en doel georiënteerd, consistent met tevoren opgestelde criteria en afgebakend op specifieke

kwalificaties. Formatief beoordelen kent andere functies: identificatie van sterkten en zwakten, gevolgd door positieve steun en constructieve feedback; ondersteuning ten aanzien van verder leren, binnen een klimaat van vertrouwen; periodiek; en continue gepositioneerd in een leerperiode, waarbij een gemeenschappelijk intentie aanwezig is bij beoordelaar en beoordeelde (Tillema, 2004).

4 Hoe dan verder

Beoordelen is een cruciale HRM(Human Resource Management)-functie in het beroep van docenten. Het speelt een rol bij de intake (werving, selectie, plaatsing), de beloning (motivatie, inzet, positie), functioneren (niveau differentiatie, loopbaanontwikkeling) en ontwikkeling (bevorderen van kennisontwikkeling en docentcompetentie). Deugdelijkheid van beoordelen (opgevat als een totale kwaliteitsmanagement-TQM-vraag, Lawler, 2001), bij voorkeur bepaald aan de hand van een audit-procedure, is daarom van groot belang. De studies in dit themanummer hebben daartoe op een aantal manieren bijgedragen en bevatten waardevolle inzichten, zoals:

- Beschikbare assessment*instrumenten* (middelen) kunnen door onderzoek worden aangevuld met een breder scala van meer specifieke en (HRM-)doelgerichte aanpakken, zoals het videodossier en het portfolio. Onderzoek kan verder aanreiken wat de precieze (doel)mogelijkheden en (gebruiks)beperkingen zijn van die instrumenten. Zo blijkt de beoordeling van het videodossier en het portfolio, ondanks problemen in afstemming tussen beoordelaars, nadrukkelijker het gedrag van docenten onder de aandacht te kunnen brengen.
- De studies maken duidelijk dat de rol van *assessoren* van wezenlijk belang is, ondanks het gebrek aan overeenstemming tussen beoordelaars. Want, zo hebben de onderzoeken van Nijveldt e.a. en Bakker e.a. laten zien, er zijn procedures te ontwikkelen die helpen de betrouwbaarheid te verhogen, zoals training, standaardisatie van taken of via de samenstelling van het assessorenteam.
- Het gemis aan, in het beroep, erkende *beoordelingscriteria* en -standaarden laat zich in de bijdragen pijnlijk voelen, hetgeen wijst op het belang van een voortgezet debat over een praktijkrelevant beoordelingskader. De bijdrage van Roelofs schetst een aanzet en biedt een perspectief, maar zal een praktijklegitimering nodig hebben wil het als beoordelingskader erkenning krijgen. De bijdrage van Bakker en collega's heeft laten zien welke validiteitsvragen bij een dergelijk debat aan de orde moeten komen.
- De *conditie*s waaronder beoordeling van docenten plaats vindt, is in de onderzoeken (noodgedwongen soms) artificieel (dat wil zeggen, ten behoeve van het onderzoek geconstrueerd). Het laat daarmee zien dat de precieze inrichting van assessmentpraktijken de nodige aandacht verdient. De vraag is dan onder welke richtlijnen, afnameprocedures, afstemmings- en scoringsregels een beoordeling tot stand komt. Deze discussiebijdrage heeft, meer nog, willen bepleiten dat de inbedding van dergelijke specificaties in wat met beoordeling wordt nagestreefd onder de aandacht moet staan, dat wil zeggen afgestemd dient te zijn op de functie van beoordelen.

De toegevoegde waarde van assessment, zo heeft dit themanummer duidelijk gemaakt, is dat er niets mis is met beoordelen zolang deugdelijkheid in het beoordelingsproces in relatie staat tot het beoordelingsdoel dat men voor ogen heeft.

Literatuur

Assessment Reform Group. (2006). T*he role of teachers in the assessment of learning*. Geraadpleegd op 20 april 2007, op: *www.assessment-reform-group.org*.

Ben-Peretz, M. (2001). The impossible role of teacher educators in a changing world. *Teacher Education, 52*(1), 48 - 56.

Birenbaum, M, Breuer, K., Cascallar, E., Dochy, F., Dori, Y., Ridgway, J., Wiesemes, R., & Nickmans, G. (2006). A learning integrated assessment system. *Educational Research Review, 1*, 61 - 65.

Clausen, M. (2007, augustus. *Instructional Quality, integrating diverging measurement of classroom environments*. Paper ge[resenteerd op de tweejaarlijke bijeenkomst van de EARLI, Budapest, Hongarije.

Cochran-Smith, M. (2001). The outcomes question in teacher education. *Teaching and Teacher Education, 17*, 527 - 546.

Cochran-Smith, M., & Zeichner, K. (2005) (Eds). *Studying Teacher education, report of the AERA panel on research and teacher education.* Washington: Mahwah.

Darling Hammond, L., & Bransford, J. (2005). *Preparing teachers for a changing world, what teachers should learn and be able to do*. San Francisco :Jossey Bass.

Delandshere, G., & Petrosky, A. (1998). Assessment of complex performances: Limitations of key measurements assumptions. *Educational Researcher*, *27*(2), 14 - 24.

Dottin, E. (2001). *The development of a conceptual framework*. AACTE: University press of America

Gilroy, P., Edwards, A., & Hartley, D. (2002). *Rethinking Teacher Education, Collaborative responses to uncertainty*. London: Routledge Falmer.

Grotendorst, A., & Tillema, H. (Red.) (2002). *Passen en meten; naar deugdelijke assessments in onderwijs en organisaties.* HRD thema papers nr 4. Alphen aan den Rijn, Nederland: Kluwer.

Heilbronn, R. (2003, augustus). *Standards are not enough*. Paper gepresenteerd op de tweejaarlijke bijeenkomst van de EARLI, Padua, Italië.

Jellema, F.(2003). *Measuring training effects: The potential of 360-degree feedback*. Dissertatie. Universiteit Twente, Enschede, Nederland.

Kane, M. T. (2006). Validation. In R.L. Brennen (Ed.), *Education Measurement* (4th ed.). Westport, CT: Praeger Publishers.

Lawler, E. (2001). *Organizing high performance, employee involvement and TQM*. San Francisco: Jossey Bass.

Lievens, F. (1998). Factors which improve the construct validity of assessment centers. *International Journal of Selection and Assessment*. *6*, 141 - 152.

Minden, J. van. (2002). *Alles over psychologische tests*. Rotterdam, Nederland: Business Contact.

Moss, P. A. (2005). Understanding the other/ understanding ourselves: Toward a constructive dialogue about principles in educational research. *Educational Theory, 55*, 263 - 283.

OECD (2005). *Formative assessment: Improving learning in secondary classrooms*. Geraadpleegd op 14 maart 2006, op Internet site *www.oecdbookshop.org*.

Onderwijsraad. (2004). Examinering in het hoger onderwijs, transparantie en kwaliteitsgarantie -advies. Den Haag, Nederland: Onderwijsraad.

Preedy, M, Glatter, R., & Wise, M (2002). *Strategic leadership and educational improvement*. New York: Sage.

SBL (Stichting Beroepskwaliteit leraren en ander onderwijspersoneel). (2004). Bekwaamheidseisen leraren. Geraadpleegd op 11 mei 2006, op: *http://ww.learanweb.nl/bijlagen/inleiding 20mei.doc*.

Smith, K., & Tillema, H. (2003). Clarifying different types of portfolio use, *Assessment & Evaluation in Higher Education, 26*, 625 - 648.

Snoek, M. (2000). Aardverschuivingen in de lerarenopleiding. *VELON Tijdschrift voor lerarenopleiders, 21*(3), 5 - 17.

Stiggins, R. J. (2002). Assessment srisis: The absence of assessment for learning. *Phi Delta Kappan, 83*, 758 - 765.

Task force on Assessment Centers. (2000). Guidelines and ethical considerations for assessment center operations. *Public Personnel Management*, *29*, 315 - 331.

Tillema, H. H. (2003). Auditing assessment practices; establishing quality criteria in the appraisal of competencies in organisations. *International Journal of Human Resource Development and Managemen*t, *3*, 359 - 369.

Tillema. H. H. (2004). Gericht werken met competenties in de opleiding. *VELON Tijdschrift voor lerarenopleiders, 25*(2), 28 - 35.

Tillema, H. H., & Smith, K. *(2007).* Portfolio assessment, in search of criteria. *Teaching and Teacher Education, 23, 442 - 456.*

Wilson, S., & Youngs, P. (2005). Research on accountability processes in teacher education. In M. Cochran-Smith & K. Zeichner (Eds.) *Studying teacher education, report of the AERA panel on research and teacher education* (pp. 591 - 645). Washington: Mahwah.

Zeichner, K. & Wray, S. (2001). The teaching portfolio in US teacher education programs: What

we know and what we need to know. *Teaching and Teacher Education*, 17, 613 - 621

Zuzowsky, R., & Libman, Z. (2002, augustus). *Standards of teaching performance and teacher tests; where do they lead us.* Paper gepresenteerd op de jaarlijkse bijeenkomst van de ATEE, Warschau, Polen.

Manuscript aanvaard: 19 mei 2008

Auteur

Harm Tillema is als Universitair hoofddocent werkzaam bij de Universiteit Leiden op het terrein van Opleiding en Ontwikkeling. Zijn specialisme ligt op het gebied van de competentie-ontwikkeling en leren van professionals.

Correspondentieadres: Harm Tillema, Pedagogische Wetenschappen, Faculteit Sociale Wetenschappen, Universiteit Leiden, Wassenaarseweg 53, 2333 AK Leiden, e-mail: *tillema@fsw.leidenuniv.nl*.

Abstract

What is wrong with assessment?

This discussion paper critically examines the contributions made to the thematic issue from the perspective of quality assurance in assessment. In order to appraise what recent research has to offer in judgmental practices, the discussion is focused on the integration of both content (competencies to be appraised) and method (ways of appraising) to conclude that clarity of procedures as well as specification of criteria and conditions would enhance warranty of assessment methods. The contributions of this paper, thus interpreted, provide challenging findings.