Validiteit in paarsgewijze beoordelingen van docentcompetenties[1]

M. Nijveldt, M. Brekelmans, D. Beijaard, Th. Wubbels en N. Verloop

Samenvatting

Omdat beoordelingen van docentcompetenties doorgaans gebaseerd worden op kwalitatief, niet-gestandaardiseerd materiaal uit verschillende bronnen en contexten, hangt de validiteit van die beoordelingen voornamelijk af van de beoordelingsprocessen van beoordelaars. Verschillende auteurs hebben gesuggereerd dat de kwaliteit van beoordelingsprocessen kan worden versterkt door samenwerking tussen beoordelaars, maar tot nu toe is weinig empirisch onderzoek beschikbaar over de aard van gezamenlijke beoordelingsprocessen en de wijze waarop samenwerking de validiteit van de beoordeling kan bevorderen. In deze studie beoordeelden 24 beoordelaars paarsgewijs dezelfde docent-in-opleiding. De aard van hun beoordelingsprocessen werd gekarakteriseerd aan de hand van de door hen ondernomen communicatieve activiteiten. Vier typen gezamenlijke beoordelingsprocessen konden op deze manier worden onderscheiden en voor elk type konden specifieke sterke punten en valkuilen worden vastgesteld. De resultaten hebben gevolgen voor de waarborging van validiteit in competentiebeoordelingen en de training van beoordelaars.

1 Inleiding

Als gevolg van de toegenomen aandacht voor de kwaliteit en professionaliteit van docenten wordt beoordeling van docentcompetenties steeds gebruikelijker. Voor vele vormen en niveaus van onderwijs zijn inmiddels competenties gespecificeerd die docenten nodig hebben om goed te kunnen functioneren in deze specifieke onderwijscontexten. Zowel het formuleren van competenties als het beoordelen ervan kunnen bijdragen aan de professionalisering van docenten: ze bevorderen discussie over de essentie van goed onderwijzen evenals reflectie van docenten op het eigen functioneren (Darling-Hammond & Snyder, 2000; Delandshere & Arens, 2003; Dwyer & Stufflebeam, 1996). Om de kwaliteit van competentiebeoordelingen te waarborgen en optimale leereffecten te realiseren, moeten beoordelingsprocedures recht doen aan het feit dat het geven van onderwijs complex, contextspecifiek en persoonsgebonden is (Cochran-Smith, 2003; Darling-Hammond & Snyder, 2000; Dwyer, 1995; Hager, Gonczi, & Athanasou, 1994; Uhlenbeck, Verloop, & Beijaard, 2002). Onderwijs geven is *complex,* omdat er in een klas veel tegelijkertijd gebeurt en docenten voortdurend afwegingen maken, bijvoorbeeld tussen aandacht geven aan een individuele leerling of instructie geven aan de klas als geheel. Onderwijs geven is *contextspecifiek,* omdat docenten hun lesgeven moeten afstemmen op de specifieke behoeften en kenmerken van hun leerlingen, de aard van de lesstof en de specifieke leerdoelen. Dit vraagt om een breed repertoire aan strategieën voor onderwijzen en om het vermogen om in te schatten welke strategie of doceerstijl het best past bij een gegeven situatie. De keuzes die docenten maken zijn *persoonsgebonden*, dat wil zeggen afhankelijk van hun persoonlijke stijl, hun persoonlijke theorieën over onderwijzen en leren, hun persoonlijke doelen en persoonlijke interpretaties van specifieke lessituaties.

Op basis van deze kenmerken van onderwijzen is een aantal aanbevelingen gedaan voor valide beoordeling van docenten (Darling-Hammond & Snyder, 2000; Dwyer, 1998; Uhlenbeck et al., 2002). Beoordelingsprocedures moeten bijvoorbeeld zo veel mogelijk authentiek zijn: de beoordeling moet plaatsvinden binnen de context van de eigen onderwijssituaties van docenten. Daarnaast is het van belang dat een beoordeling niet alleen het gedrag van docenten omvat, maar ook de onderliggende cognities. Docenten kunnen bijvoorbeeld gevraagd worden om aan te geven wat zij proberen te bereiken in een specifieke onderwijssituatie, op welke manier en hoe de speci-

fieke context hun aanpak heeft beïnvloed. Ten slotte moeten beoordelingen worden gebaseerd op meerdere databronnen, zoals lesopnames, lesplannen, lesmateriaal en reflecties van de docent op zijn of haar eigen functioneren. Geen enkele bron kan op zichzelf een volledig beeld geven van docentcompetentie.

In dergelijke vormen van competentiebeoordeling – gebaseerd op rijk, kwalitatief materiaal uit verschillende bronnen en verzameld binnen verschillende onderwijscontexten – spelen beoordelaars een cruciale rol. Hoewel een geëxpliciteerde beoordelingsprocedure en geëxpliciteerde beoordelingscriteria belangrijke condities zijn voor het realiseren van een valide oordeel, hangt de kwaliteit van de beoordeling als geheel voor een groot deel af van het vermogen van de beoordelaars om het beschikbare materiaal op een accurate wijze te *interpreteren.* Bij het interpreteren van het materiaal is het ten eerste van belang dat de beoordelaars de specifieke onderwijscontext in overweging nemen (Delandshere & Petrosky, 1998), aangezien deze context bepalend is voor de keuzes die een docent maakt. Bovendien is het van belang dat het materiaal op een holistische manier geïnterpreteerd wordt. Delandshere en Petrosky (1994, p.13) verwoorden dit als volgt: "pieces of performance can only be analyzed, interpreted and evaluated in the context of the whole performance, because their significance is determined by that context." Hoewel in de beoordeling het functioneren van docenten in zijn totaliteit beschouwd moet worden, is het echter van belang dat expliciet aandacht wordt besteed aan de bijdrage van specifieke aspecten van het functioneren aan dit totaal.

Bij beoordeling gebaseerd op meerdere databronnen kunnen de volgende essentiële beoordelingstaken of -processen worden onderscheiden: a) aanwijzen van 'bewijs van competentie' uit de afzonderlijke bronnen en b) combineren van dit bewijs tot een totaaloordeel over de kandidaat. In de literatuur worden doorgaans de volgende bedreigingen van de validiteit van deze beoordelingsprocessen onderscheiden. Ten *eerste* kunnen beoordelaars bewijsmateriaal in overweging nemen dat niet relevant is voor de te beoordelen competentie, en op deze manier 'construct-irrelevante variantie' introduceren, of simpelweg niet alle beoordelingscriteria en/of al het relevante bewijsmateriaal in overweging nemen, wat resulteert in 'construct onderrepresentatie' (Heller, Sheingold, & Myford, 1998; Messick, 1989, 1996). Een duidelijk beoordelingskader en een uitgebreid trainingsprogramma kunnen deze bedreigingen nooit volledig wegnemen. Een *tweede* belangrijke bedreiging komt voort uit het gegeven dat beoordelaars al heel snel een voorlopig oordeel hebben van een kandidaat en ongewild de neiging hebben om vooral te zoeken naar bewijs dat dit voorlopige oordeel bevestigt (Moss, Schutz, & Collins, 1998; Schutz & Moss, 2002). Verschillende auteurs hebben gesuggereerd dat de bovengenoemde bedreigingen van validiteit zouden kunnen worden verkleind en de kwaliteit van het beoordelingsproces kan worden vergroot door samenwerking in duo's of grotere teams (bijv. Johnston, 2004; Moss et al., 1998; Tigelaar, Dolmans, Wolfhagen, & Van der Vleuten, 2005). Samenwerking wordt in dit geval gezien als dialoog c.q. discussie tussen beoordelaars nadat zij individueel tot een oordeel zijn gekomen over een kandidaat.

In theorie worden gezamenlijke beoordelingsprocessen als veelbelovend gezien en ook in de praktijk komt gezamenlijke beoordeling regelmatig voor – bijvoorbeeld in de lerarenopleiding, waar vaak meerdere opleiders betrokken zijn bij beoordelingsbeslissingen. Tot nu toe heeft slechts een aantal studies zich expliciet gericht op de aard of kwaliteit van gezamenlijke beoordelingsprocessen en de manier waarop dialoog c.q. discussie tussen beoordelaars de validiteit van het beoordelingsproces zou kunnen vergroten. In dit artikel beschrijven we de gezamenlijke beoordelingsprocessen van beoordelaars die paarsgewijs een docent-in-opleiding (dio) beoordeelden op basis van een specifieke beoordelingsprocedure. De aard en kwaliteit van de beoordelingsprocessen worden beschreven op basis van de communicatieve activiteiten die de beoordelaarsparen ondernemen.

2. Gezamenlijke beoordelingsprocessen

Om meer zicht te krijgen op de aard van essentiële beoordelingsprocessen wordt in deze

studie geput uit literatuur vanuit een hermeneutische, interpretatieve benadering van beoordeling (Delandshere & Arens, 2003; Delandshere & Petrosky, 1998; Johnston, 2004; Moss, 1994; Tigelaar et al., 2005). Binnen deze benadering wordt erkend dat beoordeling afhankelijk is van subjectieve, menselijke interpretatie. Kenmerkend voor een interpretatief beoordelingsproces is daarom het voortdurend testen, uitdagen en herzien van interpretaties, net zo lang tot al het beschikbare bewijsmateriaal in overweging genomen is (Moss et al., 1998). Dit wordt gezien als een belangrijk middel om tot een accurate, valide interpretatie te komen. Voorlopige interpretaties zouden niet alleen moeten worden uitgedaagd op basis van tegenvoorbeelden uit het materiaal, maar ook door alternatieve interpretaties of perspectieven van een andere beoordelaar (Guba & Lincoln, 1989; Johnston, 2004; Moss et al., 1998; Tigelaar et al., 2005). Het is hierbij van belang dat het interpretatieproces zorgvuldig wordt gedocumenteerd en dat beoordelaars aangeven op basis van welke specifieke gegevens zij tot hun conclusies zijn gekomen (vgl. Kane, 1992).

Op basis van empirische studies vanuit een interpretatieve benadering van beoordeling gaan we hieronder in op de aard van de twee eerder onderscheiden essentiële beoordelingsprocessen: a) aanwijzen van bewijs van competentie uit de afzonderlijke bronnen en b) combineren van dit bewijs tot een totaaloordeel. We besteden hierbij specifiek aandacht aan de moeilijkheden die deze processen omvatten.

Het *aanwijzen van bewijs van competentie* omvat het toepassen van een beoordelingskader op het concrete, contextgebonden materiaal van de kandidaat. Hierbij kunnen twee belangrijke moeilijkheden worden onderscheiden. Ten eerste is het onvermijdelijk dat beoordelaars persoonlijke opvattingen hebben over de essentie van de te beoordelen competentie. Deze opvattingen kunnen in tegenspraak zijn met het beoordelingskader, en daarmee een bedreiging vormen voor de validiteit van het oordeel, maar kunnen ook gebaseerd zijn op praktijkkennis die de kwaliteit van de beoordeling kan vergroten (Moss et al., 1998). Ten tweede kunnen beoordelaars bij het benoemen van bewijsmateriaal de neiging hebben om zich te beperken tot het herkennen van oppervlakkige kenmerken van competentie. Zij verwijzen dan bijvoorbeeld voornamelijk naar abstracte sleutelbegrippen uit het beoordelingskader en niet naar de kenmerkende bijzonderheden van de specifieke kandidaat (Delandshere & Arens, 2003; Delandshere & Petrosky, 1998; Moss et al., 1998). De kwaliteit van het aanwijzen van bewijsmateriaal staat hier ter discussie.

Het *combineren van bewijs tot een totaaloordeel* behelst de volgende problemen voor beoordelaars. Studies van Moss en collega's (1998) en Schutz en Moss (2004) – uitgevoerd in de context van beoordeling van docentportfolio's – illustreerden dat beoordelaars de neiging hebben om hun eindoordeel te baseren op een *selectie* van het beschikbare bewijsmateriaal, in plaats van een allesomvattende afweging te maken. Daarnaast illustreerden deze studies dat beoordelaars specifieke onderdelen van het portfolio interpreteren op basis van een meer algemene totaalindruk. Met andere woorden, beoordelaars zoeken bewust of onbewust naar patronen in de portfoliodata. Op het moment dat zo'n patroon zich gevormd heeft, wordt nieuwe informatie geïnterpreteerd in termen van dit patroon: zelfs als de nieuwe informatie in strijd is met het aanvankelijke patroon is de kans groot dat deze – tegenstrijdige – informatie geïnterpreteerd wordt als bevestiging van dit patroon. Deze neiging om bevestiging te zoeken voor je aanvankelijke indruk kan gezien worden als een vorm van *bias* die nadelig kan zijn voor de validiteit van de beoordeling.

Op basis van deze bevindingen introduceerden Moss e.a. de volgende principes om sturing te geven aan het beoordelingsproces. Het *eerste* principe houdt in dat beoordelaars coherentie zouden moeten zoeken tussen het beschikbare bewijsmateriaal en expliciet na zouden moeten gaan of al het relevante bewijsmateriaal in overweging genomen is. Een *tweede*, verwant principe houdt in dat beoordelaars aangemoedigd zouden moeten worden om expliciet te zoeken naar bewijs dat de zich ontwikkelende indruk zou kunnen ontkrachten, en om expliciet na te gaan of alternatieve interpretaties van het beschikbare be-

wijs mogelijk zijn. Deze principes worden gezien als kernprincipes voor een valide beoordeling. Samenwerking ofwel discussie tussen beoordelaars, nadat zij een voorlopige individuele indruk hebben gevormd, zou beoordelaars bij uitstek kunnen stimuleren om te voldoen aan de genoemde principes. De validiteit van het beoordelingsproces kan met name worden gewaarborgd wanneer beoordelaars elkaar met betrekking tot bewijsmateriaal en argumentatie zowel aanvullen als actief uitdagen: "the validity of the conclusion is warranted, in part, in the consensus among [assessors] who are empowered to challenge one another's developing interpretations in light of the cases at hand" (Moss et al., p. 142). Hoewel in dit artikel de *validiteit* van beoordelingen centraal staat, moet opgemerkt worden dat bovenstaande principes tevens gerelateerd kunnen worden aan de *betrouwbaarheid* van de beoordeling. Met Mabry (1999) en Uhlenbeck e.a. (2002) zijn we van mening dat het onderscheid tussen validiteit en betrouwbaarheid in competentiebeoordelingen minder scherp is dan in traditionele vormen van toetsing. Het criterium van betrouwbaarheid wordt gezien de complexiteit van het beoordelingsproces door steeds meer auteurs uitgebreid van overeenstemming over het eindoordeel, c.q. het niveau van de kandidaat, tot overeenstemming over de argumentatie die ten grondslag ligt aan dit oordeel. We veronderstellen dat wanneer voldaan wordt aan de voornoemde principes van valide beoordeling, de kwaliteit – en eenduidigheid – van de argumentatie toeneemt en daarmee de betrouwbaarheid van het oordeel.

In het onderhavige onderzoek worden de gezamenlijke beoordelingsprocessen van beoordelaars gekarakteriseerd aan de hand van de specifieke 'communicatieve' activiteiten die beoordelaarsparen ondernemen. In literatuur over gezamenlijke kennisconstructie worden communicatieve activiteiten doorgaans geanalyseerd op basis van het type bijdrage aan een discussie en het type reactie hierop. Vier typen communicatieve activiteiten kunnen op deze manier worden onderscheiden: a) inbreng leveren, b) accepteren van inbreng, c) bediscussiëren van inbreng en d) negeren van inbreng (vgl. Barron, 2003). Vertaald naar de context van gezamenlijke beoordelingsprocessen kan 'inbreng' bijvoorbeeld zijn: inbrengen van bewijsmateriaal voor competentie, inbrengen van een interpretatie van dit bewijsmateriaal of toetsing van de kandidaat tegen een specifieke norm. Accepteren van een inbreng omvat instemming met het gepresenteerde bewijsmateriaal, het inbrengen van aanvullend bewijsmateriaal, of instemming met aanvullend bewijsmateriaal. Bediscussiëren van een inbreng omvat bijvoorbeeld het ter discussie stellen of afwijzen van een interpretatie, het presenteren van tegenbewijs, of het ter discussie stellen van het belang of gewicht van ingebracht bewijsmateriaal. In deze studie beperken we de categorie negeren van inbreng tot negeren van een *discussiebijdrage*, zoals niet in discussie gaan over ingebracht tegenbewijs.

De onderzoeksvraag die in deze studie centraal staat is: Wat is de aard en kwaliteit van beoordelingsprocessen van beoordelaars die paarsgewijs een docent beoordelen? En meer specifiek:

a) Wat zijn de communicatieve activiteiten die beoordelaars uitvoeren bij de gezamenlijke beoordeling van docentcompetenties?
b) In welke mate representeren deze activiteiten de twee beoordelingsprincipes 'coherentie zoeken tussen het beschikbare bewijsmateriaal en expliciet nagaan of al het relevante bewijsmateriaal in overweging genomen is' en 'expliciet zoeken naar tegenbewijs en /of alternatieve interpretaties'?
c) Welke verschillen tussen de afzonderlijke beoordelingsparen zijn in dit verband te onderscheiden?

3. Methode

3.1 Beoordelingsprocedure

In deze studie analyseerden we de beoordelingsprocessen van beoordelaars die de interpersoonlijke competentie van een docent-in-opleiding (dio) beoordeelden aan de hand van een hiervoor ontwikkelde beoordelingsprocedure. Er is gekozen voor beoordeling van interpersoonlijke competentie – het kun-

nen creëren van een goede werksfeer en daarmee een goede relatie met leerlingen – omdat dit voor veel dio's een belangrijk aandachtspunt is.

De ontwikkelde beoordelingsprocedure is gebaseerd op drie bronnen: 1) een door de dio geselecteerde video-opname van een les, 2) de Vragenlijst Interpersoonlijk Leraarsgedrag (VIL) die een valide en betrouwbaar beeld geeft van de percepties van leerlingen van de interpersoonlijke relatie met hun docent (Wubbels, Brekelmans, Den Brok, & Van Tartwijk, 2006), en 3) een zelfevaluatie van de dio waarin hij of zij de eigen interpersoonlijke competentie analyseert met zo veel mogelijk verwijzing naar de lesopname en de VIL-resultaten.

Om de beoordelaars te ondersteunen bij het aanwijzen van bewijsmateriaal en het combineren van bewijsmateriaal tot een totaaloordeel, werd een beoordelingskader aangeboden waarin zes essentiële aspecten van interpersoonlijke competentie worden onderscheiden: 1) sturing geven/structuur bieden aan leerlingen, 2) normen en regels stellen, 3) corrigeren van ongewenst gedrag, 4) aandacht geven en belonen, 5) ruimte en verantwoordelijkheid geven aan leerlingen, en 6) reflecteren op eigen interpersoonlijk functioneren. Voor elk aspect werd ter illustratie een aantal indicatoren en contra-indicatoren van competentie uitgewerkt. Dit beoordelingskader werd gebaseerd op literatuur over interpersoonlijk leraarsgedrag, en meer specifiek op het Model voor Interpersoonlijk Leraarsgedrag (Wubbels et al., 2006). Voor meer details over het beoordelingskader – en het ontwerp en de evaluatie van de beoordelingsprocedure – wordt verwezen naar Nijveldt, Beijaard, Brekelmans, Verloop, & Wubbels (2005).

Om beoordelaars te ondersteunen bij het systematisch noteren van bewijsmateriaal voor interpersoonlijke competentie werd voor elk instrument een formulier ontworpen dat kolommen bevatte voor het noteren van bewijs bij elk van de zes aspecten. Deze formulieren bevatten bovendien voor elk aspect een aantal vragen om de analyse van de beoordelaars aan te sturen. Voor het aspect sturing geven / structuur bieden waren dit bijvoorbeeld de volgende vragen: Hoe biedt de docent structuur aan de leerlingen? Hoe handelt de docent tijdens plenaire momenten? De vragen voor het aspect reflecteren op eigen interpersoonlijk functioneren waren: Kan de docent zijn eigen sterke en zwakke kanten benoemen? Weet de docent een goede analyse te maken van zijn eigen interpersoonlijk gedag? Ziet de docent in op welke punten hij zichzelf verder zou kunnen ontwikkelen? Heeft de docent adequate ideeën over de manier waarop hij deze verdere ontwikkeling zou kunnen realiseren?

Een typisch kenmerk van de ontwikkelde beoordelingsprocedure, ten slotte, is discussie met een andere beoordelaar nadat beoordelaars een individueel oordeel hebben geformuleerd. Volgens de procedure noteren beoordelaars eerst individueel bewijsmateriaal uit elk van de drie databronnen. Dan integreren zij dit bewijs tot een individueel oordeel in de vorm van een 'evaluatieve samenvatting' waarin zorgvuldig uiteengezet wordt op basis van welk bewijsmateriaal zij tot welk oordeel gekomen zijn (vgl. Delandshere & Petrosky, 1994, 1998; Moss et al., 1998). Ten slotte gaan beoordelaars in gesprek met een andere beoordelaar en formuleren zij een gezamenlijke, uiteindelijke evaluatieve samenvatting. Gedurende dit proces worden beoordelaars aangemoedigd om niet alleen ondersteunend bewijsmateriaal te noemen, maar ook tegenstrijdig bewijs en tegenvoorbeelden, alternatieve interpretaties te bediscussiëren en interpretaties te herzien tot al het relevante bewijsmateriaal in overweging genomen is (vgl. Moss et al., 1998). Opgemerkt moet worden dat de beoordelaars geen vaste richtlijnen kregen aangeboden voor de wijze waarop de drie databronnen in de evaluatieve samenvatting gecombineerd moesten worden; de complexiteit van de data maakt een strikt schema ongepast (Wolf, 1995). Het idee achter de evaluatieve samenvatting is dat de beoordelaars zelf beschrijven hoe zij een coherent beeld gevormd hebben uit de afzonderlijke gegevens. Dit is kenmerkend voor een interpretatieve benadering van beoordeling. Ook omvatte het beoordelingskader geen expliciete beoordelingscriteria of standaarden voor de zes aspecten van interpersoonlijke competentie. De beoordelaars werd gevraagd om de criteria die zij hanteerden te

expliciteren in de evaluatieve samenvatting. Wanneer meer ervaring is opgedaan met de gehanteerde beoordelingsprocedure kunnen de criteria voor bijvoorbeeld de niveaus onvoldoende, voldoende en excellent worden uitgewerkt, en geïllustreerd op basis van authentieke cases, ofwel *benchmarks* (vgl. Mabry, 1999).

3.2 Beoordelaars en training

In totaal werden vierentwintig personen geselecteerd om als beoordelaar deel te nemen aan de studie. Onder hen waren lerarenopleiders (N = 4) en zogenoemde begeleiders-op-school van dio's (N = 20). Allen hadden ervaring met het evalueren van dio's en het geven van feedback. Vijf van hen hadden daarnaast ervaring als beoordelaar van competenties van zij-instromers. Ter voorbereiding op de studie volgden alle deelnemers een beoordelaarstraining van twee avonden. De eerste avond werd de ontwikkelde procedure voor het beoordelen van interpersoonlijke competentie geïntroduceerd en werd het beoordelingskader uitvoerig bediscussieerd. De tweede avond beoordeelden de beoordelaars een dio op basis van door een dio ter beschikking gesteld materiaal. Volgens de procedure vormden zij eerst individueel een indruk; daarna ging iedere beoordelaar in discussie met een andere beoordelaar op basis waarvan zij een gezamenlijke indruk formuleerden. Nadat alle beoordelaars deze taak hadden uitgevoerd werden de ervaringen uitgewisseld in een plenaire discussie, met specifieke aandacht voor het aangewezen bewijsmateriaal, de interpretatie van dit bewijs, de combinatie van dit bewijs tot een totaalindruk en de ervaren moeilijkheden. In de training werden de beoordelaars gewezen op de bovengenoemde principes van coherentie zoeken tussen bewijsmateriaal en nagaan of al het relevante bewijsmateriaal in overweging genomen is, en expliciet zoeken naar bewijsmateriaal dat de zich ontwikkelende indruk zou kunnen ontkrachten. In aanvulling hierop werden de beoordelaars gewezen op het belang van nagaan of alle conclusies kunnen worden onderbouwd met duidelijke argumentatie, gebaseerd op de oorspronkelijke data. De beoordelaars kregen geen specifieke richtlijnen aangeboden voor de manier waarop zij deze principes in de praktijk moesten brengen. Aangezien de literatuur hiertoe nog weinig praktische aanknopingspunten biedt, waren we voornamelijk geïnteresseerd in de strategieën die beoordelaars zelf gebruiken.

3.3 Dataverzameling en -analyse

Na afronding van de beoordelaarstraining kregen de vierentwintig beoordelaars de opdracht om paarsgewijs een nieuwe case te beoordelen, gebaseerd op materiaal dat ter beschikking was gesteld door een dio. Alle twaalf beoordelaarsparen beoordeelden dezelfde dio. Volgens de procedure tekenden de beoordelaars eerst individueel bewijsmateriaal aan bij de drie informatiebronnen. Vervolgens formuleerden zij een individuele evaluatieve samenvatting. Op basis van deze individuele voorbereiding werden zij gevraagd om in tweetallen een definitieve, gezamenlijke evaluatieve samenvatting op te stellen. Van dit gezamenlijk proces werden audio-opnamen gemaakt.

De gesprekken tussen beoordelaars werden volledig getranscribeerd en geanalyseerd volgens de volgende vier stappen. De eerste stap bestond uit het samenstellen van een categorieënsysteem voor de communicatieve activiteiten die werden ondernomen door de beoordelaarsparen. Op basis van een voorlopige lijst van categorieën, gedestilleerd uit de hierboven aangehaalde literatuur, werden door de eerste auteur vijf gesprekken gecodeerd die een duidelijke variatie lieten zien in de aard en de frequentie van communicatieve activiteiten. Gedurende dit proces werden bestaande categorieën verfijnd en nieuwe toegevoegd. Het aanvankelijke, op theorie gebaseerde categorieënsysteem werd dus verfijnd op basis van de beschikbare data.

In een tweede stap werden twee van de vijf gecodeerde gesprekken gecodeerd door een onafhankelijke onderzoeksassistent. De verschillen tussen de toegekende codes werden bediscussieerd tot overeenstemming werd bereikt over de toe te kennen code. Op basis van deze discussies werden de definities van enkele codeercategorieën aangescherpt en werden voorbeelden toegevoegd om de codes te illustreren. Opgemerkt moet worden dat de fragmenten werden afgebakend door de eerste auteur. Een nieuw frag-

ment begon met een nieuwe communicatieve activiteit. In een zeer beperkt aantal gevallen werd deze sectionering op basis van discussie met de onderzoeksassistent aangepast.

In de derde stap werden de overige tien transcripten gecodeerd door zowel de eerste auteur als de onafhankelijke onderzoeksassistent, met behulp van Atlas.ti, een softwareprogramma voor kwalitatieve data-analyse. Ook in deze stap werden de fragmenten afgebakend door de eerste auteur. Aan 90% van de fragmenten werd dezelfde code toegekend. De codering van de fragmenten die niet overeenstemde (10%) betrof alle (sub)categorieën. De betreffende coderingen werden bediscussieerd tot overeenstemming werd bereikt.

In de vierde stap vergeleken we de aard van de beoordelingsprocessen van de twaalf beoordelaarsparen. We hebben ervoor gekozen om de beoordelingsprocessen te vergelijken op basis van de frequenties van de verschillende communicatieve activiteiten die de beoordelaarsparen ondernamen. We bekeken eerst hoe vaak de verschillende activiteiten voorkwamen in de afzonderlijke paren. De paren werden vervolgens gecategoriseerd door te kijken hoe vaak activiteiten voorkwamen die het meest kenmerkend waren voor de twee onderscheiden beoordelingsprincipes. Op basis van de frequenties van deze kenmerkende activiteiten werden de afzonderlijke paren onderverdeeld in verschillende typen gezamenlijke beoordelingsprocessen. Voor elk type proces gingen we vervolgens na welke communicatieve activiteiten binnen dit type frequent of juist minder frequent voorkwamen in vergelijking met de totale groep. Wanneer de gemiddelde frequentie van een communicatieve activiteit voor een bepaald type meer dan 0,8 standaarddeviatie hoger of lager lag dan het gemiddelde van de totale groep (vgl. Cohen, 1988), werd dit als kenmerkend beschouwd voor dit type samenwerking. In aanvulling hierop werd de aard van de verschillende typen beoordelingsprocessen meer kwalitatief en holistisch beschreven, op basis waarvan specifieke sterke en zwakke punten met betrekking tot de validiteit van de beoordeling werden geïdentificeerd.

4 Resultaten

4.1 Overzicht van communicatieve activiteiten

De eerste twee stappen van de hierboven beschreven analyse resulteerden in het categorieënsysteem, zoals gepresenteerd in Tabel 1. Dit categorieënsysteem omvat de communicatieve activiteiten die de verschillende beoordelaarsparen lieten zien tijdens het beoordelen van docentcompetenties. De tabel laat zien dat de communicatieve activiteiten zijn gegroepeerd in vier algemene categorieën: A) inbreng leveren (16 categorieën), B) accepteren van inbreng (3 categorieën), C) bediscussiëren van inbreng (3 categorieën, 8 subcategorieën) and D) negeren van (confronterende) inbreng. Voor elke categorie of communicatieve activiteit worden achtereenvolgens de totaalfrequentie, de gemiddelde frequentie per paar, het bereik en de standaarddeviatie weergegeven.

Tabel 1 laat zien dat het beoordelingsprincipe *zoeken van coherentie tussen de verschillende databronnen en nagaan of al het relevante bewijs in overweging is genomen* in de gesprekken tussen beoordelaars vooral tot uitdrukking komt in de activiteit *aandragen van aanvullend bewijsmateriaal* (B2). Het beoordelingsprincipe *uitdagen van de zich ontwikkelende totaalindruk door actief op zoek te gaan naar tegenbewijs of alternatieve interpretaties* komt het duidelijkst tot uitdrukking in de activiteit *uitdagen van een interpretatie door aandragen van confronterend bewijsmateriaal of een alternatieve interpretatie* (C3). Deze twee communicatieve activiteiten kwamen relatief vaak voor – respectievelijk 124 en 72 maal – terwijl de totaalfrequentie van een aantal andere activiteiten die samenhangen met de onderscheiden beoordelingsprincipes relatief weinig voorkomen.

In de categorie inbreng leveren kwamen de volgende activiteiten bijvoorbeeld weinig voor: constateren van incoherentie tussen de beschikbare databronnen (A7), constateren van een meningsverschil tussen de beoordelaars (A9), expliciet nagaan of al het relevante bewijs in overweging genomen is (A11) en ter discussie stellen van de volledigheid van het bewijsmateriaal (A16). In de categorie

Tabel 1

Totaalfrequenties, gemiddelden, bereik en standaarddeviaties van communicatieve activiteiten van beoordelaars

Activiteiten / codes	Totaal	Gem.	Bereik	*SD*
A. Inbreng leveren				
1. Aandragen van bewijsmateriaal	309	25,8	13-51	10,5
2. Aandragen van een interpretatie van bewijs	133	11,1	3-19	5,1
3. Inbrengen van een (eind)oordeel	143	11,9	8-18	3,1
4. Beoordelen van de kandidaat tegen een norm	40	3,3	0-10	3,7
5. Inbrengen van een oordeel op een tien-puntsschaal	48	4,0	2-6	1,5
6. Constateren van coherentie tussen de beschikbare databronnen	21	1,8	2-4	1,5
7. Constateren van incoherentie tussen de beschikbare databronnen	3	0,3	0-2	0,6
8. Constateren van overeenstemming tussen de beoordelaars	23	1,7	0-6	1,9
9. Constateren van een meningsverschil tussen de beoordelaars	1	0,0	0-1	0,3
10. Expliciteren van een definitie van interpersoonlijke competentie	7	0,6	0-3	1,0
11. Expliciet nagaan of al het relevante bewijs in overweging genomen is	8	0,8	0-3	1,0
12. Inbrengen van een voorstel voor de te volgen aanpak / procedure	32	2,6	0-6	2,3
13. Vragen om onderbouwing of verheldering	20	1,7	0-5	1,9
14. Herformuleren van aanvankelijke inbreng (bewijs, interpretatie, oordeel)	7	0,6	0-1	0,5
15. Vragen om aanvulling of confrontatie	34	2,8	0-6	1,9
16. Ter discussie stellen van de volledigheid van het bewijsmateriaal	4	0,3	0-4	1,1
B. Accepteren van inbreng				
1. Bevestiging van inbreng met argumentatie /toelichting	78	6,5	0-13	4,9
2. Aandragen van aanvullend bewijsmateriaal (AANVULLEN)	124	10,3	0-16	5,3
3. Bevestiging van inbreng zonder argumentatie / toelichting	198	16,5	8-29	5,6
C. Bediscussiëren van inbreng				
1. Afwijzen van inbreng (bewijs, interpretatie, oordeel)				
a. Afwijzen van inbreng op basis van construct-irrelevantie	7	0,6	0-2	0,8
b. Afwijzen van inbreng op basis van de kwaliteit van het bewijsmateriaal	3	0,3	0-1	0,5
c. Afwijzen van inbreng op basis van de norm	4	0,3	0-2	0,7
d. Afwijzen van inbreng zonder toelichting	16	1,3	0-4	1,5
2. Bediscussiëren van het gewicht van inbreng				
a. Op basis van de kwaliteit van het bewijsmateriaal	38	3,2	0-11	2,8
b. Op basis van de norm	31	2,5	0-7	2,4
c. Op basis van de specifieke omstandigheden / context	25	2,1	0-8	2,3
d. Zonder toelichting	33	2,8	0-9	2,9
3. Uitdagen van een interpretatie door aandragen van confronterend bewijsmateriaal of een alternatieve interpretatie (CONFRONTEREN)	72	6,0	0-16	5,0
D. Negeren van confronterende inbreng	10	0,8	0-3	0,9

bediscussiëren van inbreng, kwamen afwijzingen op basis van constructirrelevantie (C1a), de kwaliteit van het bewijs (C1b) en de norm (C1c) weinig voor. Hoewel beoordelaars frequent het gewicht van inbreng bediscussiëren op basis van kwaliteit (C2a), standaard (C2b) en/of de specifieke omstandigheden of context (C2c), leidden zulke discussies zelden tot expliciete afwijzing van de betreffende inbreng.

4.2 Vier typen gezamenlijke beoordelingsprocessen

In deze paragraaf worden de gesprekken van de twaalf beoordelaarsparen gekarakteriseerd aan de hand van het voorkomen van de activiteiten aandragen van aanvullend bewijsmateriaal (*aanvullen*) en uitdagen van een interpretatie door aandragen van confronterend bewijsmateriaal of een alternatieve interpretatie (*confronteren*), waarin de twee beoordelingsprincipes het meest prominent tot uitdrukking komen. De gesprekken van de beoordelaars verschilden aanmerkelijk met betrekking tot zowel het aanvullen als het confronteren. Vier typen gezamenlijke beoordelingsprocessen konden worden onderscheiden: een gezamenlijk proces waarin beoordelaars elkaar zowel aanvullen als confronteren (Type I), voornamelijk confronteren (Type II), voornamelijk aanvullen (Type III), of aanvullen noch confronteren (Type IV). Hieronder beschrijven we welke communicatieve activiteiten bovengemiddeld of ondergemiddeld voorkwamen binnen de vier typen beoordelingsprocessen.

Type I: Aanvullen en confronteren

De gezamenlijke beoordelingsprocessen van drie van de twaalf beoordelaarsparen konden worden gekarakteriseerd als Type I. De activiteiten aanvullen en confronteren kwamen vaker dan gemiddeld voor in de gesprekken van de paren A, B en C. Daarnaast kwam een aanzienlijk aantal andere communicatieve activiteiten vaker voor dan gemiddeld (zie Tabel 2). Geen van de onderscheiden activiteiten kwam binnen deze duo's minder vaak voor dan gemiddeld. Verder kan worden opgemerkt dat de paren A, B en C verantwoordelijk zijn voor alle drie de afwijzingen van inbreng op basis van de kwaliteit van het bewijsmateriaal die voorkomen in de gehele groep (Tabel 1), en voor vijf van de zeven af-

Tabel 2

Communicatieve activiteiten die boven- en ondergemiddeld voorkomen in Type I beoordelingsprocessen (aanvullen en confronteren)

Activiteiten		Gemiddelde	
		Type I	Totaal
Activiteiten die **bovengemiddeld** voorkomen:			
Inbreng	Beoordelen van de kandidaat tegen een norm	8,7	3,3
	Expliciteren van een definitie van interpersoonlijke competentie	1,7	0,7
	Inbrengen van een voorstel voor de te volgen aanpak / procedure	5,0	2,6
	Vragen om onderbouwing of verheldering	3,0	1,7
	Ter discussie stellen van de volledigheid van het bewijsmateriaal	1,3	0,3
Acceptatie	Aandragen van aanvullend bewijsmateriaal	13,0	10,3
Discussie	Afwijzen van inbreng op basis van construct-irrelevantie	1,7	0,6
	Afwijzen van inbreng op basis van de kwaliteit van het bewijsmateriaal	1,0	0,3
	Uitdagen van een interpretatie door aandragen van confronterend bewijsmateriaal of een alternatieve interpretatie	10,7	6,0
Activiteiten die **ondergemiddeld** voorkomen:			
Geen			

wijzingen gebaseerd op constructirrelevantie. Zulke afwijzingen reflecteren duidelijk het beoordelingsprincipe van het confronteren van de zich ontwikkelende totaalindruk.

De frequentie van communicatieve activiteiten die *discussie* omvatten was voor Type I-paren groter dan gemiddeld. Nadere bestudering van de discussies van de drie Type I-paren laat het volgende zien. In het gesprek van paar A konden vier interacties worden onderscheiden die gebaseerd waren op discussie of, met andere woorden, op C1, C2 en/of C3 activiteiten. Drie van deze vier discussies resulteerden in een expliciete conclusie. In het gesprek van paar B waren twee van de drie interacties waarin discussie voorkwam, gebaseerd op slechts één discussiebijdrage. Deze discussiebijdrage werd in beide gevallen genegeerd door beide beoordelaars. De derde discussie resulteerde niet in een expliciete conclusie, maar had als resultaat dat het bewijs dat geleid had tot een bepaald oordeel werd gespecificeerd. Een fragment van deze discussie is opgenomen in Tekstbox 1. In het gesprek van paar C, ten slotte, resulteerden drie van de vier discussies in een expliciete conclusie en was de vierde discussie gebaseerd op een enkele discussiebijdrage die werd genegeerd door beide beoordelaars. De analyse van de beoordelingsprocessen gekarakteriseerd als Type I wees uit dat de argumentatie van Type I-beoordelaarsparen relatief uitgebreid en transparant is voor alle onderscheiden aspecten van interpersoonlijke competentie. In de argumentatie die ten grondslag ligt aan het eindoordeel krijgen ook tegenbewijs en tegenargumenten een plek.

Type II: Confronteren

Slechts één van de twaalf beoordelaarsparen, paar D, liet een gezamenlijk beoordelingsproces zien dat voornamelijk gebaseerd was op confrontatie. De communicatieve activiteiten die binnen dit paar vaker of minder vaak voorkwamen dan gemiddeld zijn opgenomen in Tabel 3.

In het gesprek van paar D konden drie duidelijke discussies worden onderscheiden. Eén van deze discussies was zeer uitgebreid en omvatte vrijwel het gehele gesprek. Deze discussie resulteerde in een uitgebreide conclusie waarin de beoordelaars ingingen op zowel de interpretaties waarover ze duidelijk overeenstemming hadden bereikt als de interpretaties waarover ze geen overeenstemming

Tekstbox 1

Voorbeeld van een Type I gezamenlijk beoordelingsproces: Aanvullen en confronteren

B1	Deze docent moet volgens mij vooral werken aan haar onzekerheid. Want als zij wat zekerder is, dan treedt zij op heel veel terreinen volgens mij krachtiger op. (A2)
B2	Vond jij haar zo'n onzekere indruk maken? Ik vond dat zelf niet echt namelijk. (C1d)
B1	Nou, in die zin onzeker dat ze zo onzichtbaar was in de les. (A1)
B2	Ja, dat ben ik wel met je eens (B3) Maar ik vind dat ze niet heel onzeker was. Ze heeft het er wel over [zelf-evaluatie] en je ziet het ook in de VIL [zelfbeeld docent], maar ik heb zelf het idee dat dat bij haar een beetje een pose is. Dat je niet wilt dat anderen negatiever over je zijn dan jij zelf. Begrijp je? (C3)
B1	Dat is een vorm van onzekerheid (A10)
B2	Ja, maar je ziet dat niet in de les. het is meer iets dat zich in haar hoofd afspeelt. Want ze heeft best wel overwicht. Bij het begin van de les dacht ik even dat wordt helemaal niks: ze komt binnen en doet eerst een hele tijd helemaal niets. Dat vond ik merkwaardig. Maar op een gegeven moment gaat ze beginnen met praten en ze doet er geen enkele moeite voor en die leerlingen luisteren wel naar haar. En ze weet ze ook, soms met een heel klein tikje, uitermate goed onder controle te houden. Ze hoeft maar even gebaartjes te maken, of ze zegt even "dames", en dan onmiddellijk ophouden met praten. Of wil je ook even meedoen, die dame met dat haar, draait zich onmiddellijk om. En dat verandert pas op het moment dat die uitleg te lang gaat duren. Dan draaien de zaken om en dan weet ze volgens mij niet meer wat ze moet doen. Dus dat begin heeft ze heel goed onder controle, maar op een gegeven moment verliest ze die controle. Volgens mij zit dat in haar correcties. Ze is heel goed in het geven van lage intensiteit correcties, hè, maar als dat niet meer werkt, dan weet ze het niet meer. Dan gaat het dus fout, en dan gaat ze zichzelf overschreeuwen. (C3)
B1	Nou, ik denk dat ik daar mijn meningen over onzeker zijn op heb gebaseerd: het niet ingrijpen daar waar het nodig is op een gegeven moment. Ze geeft wat dat betreft nog te weinig richtlijnen bij de verschillende lesonderdelen, en ze komt dus nauwelijks de klas in, staat achter die bank hè? Ik vond haar af en toe echt onzichtbaar, dus daar heb ik die onzekerheid op gebaseerd. (B2)

Tabel 3

Communicatieve activiteiten die boven- en ondergemiddeld voorkomen in Type II beoordelingsprocessen (confronteren)

Activiteiten		Gemiddelde	
		Type II	Totaal
Activiteiten die **bovengemiddeld** voorkomen:			
Inbreng	Expliciteren van een definitie van interpersoonlijke competentie	2,0	0,7
Discussie	Bediscussiëren van gewicht op basis van de kwaliteit van het bewijsmateriaal	5,0	2,5
	Uitdagen van een interpretatie door aandragen van confronterend bewijsmateriaal of een alternatieve interpretatie	15,0	6,0
Activiteiten die **ondergemiddeld** voorkomen:			
Acceptatie	Bevestiging van inbreng met argumentatie /toelichting	2,0	6,5
	Aandragen van aanvullend bewijsmateriaal	3,0	10,3

konden bereiken en/of waarvoor ze naar hun mening onvoldoende data ter beschikking hadden. Een fragment van deze discussie is opgenomen in Tekstbox 2. De analyse van het beoordelingsproces van paar D laat zien dat hun oordelen met betrekking tot *die* aspecten van interpersoonlijke competentie waarover discussie ontstond zorgvuldig zijn onderbouwd met bewijsmateriaal. Zowel bevestigend bewijs als tegenbewijs krijgen een plaats in de argumentatie, en constructrelevantie, de kwaliteit van het bewijsmateriaal en het relatieve gewicht van het bewijsmateriaal worden bediscussieerd. De overige aspecten, dus die aspecten waarover op het eerste gezicht geen meningsverschillen leken te bestaan, kregen echter minder aandacht in het gesprek. De argumentatie met betrekking tot deze overige aspecten blijft vrij abstract, met vrijwel geen verwijzing naar relevant bewijsmateriaal.

Tekstbox 2

Voorbeeld van een Type II gezamenlijk beoordelingsproces: Confronteren

B1	De VIL… Ik vind het leerlingbeeld eigenlijk niet overeenstemmen met wat ik zie in de les. In haar zelfevaluatie geeft de docent aan dat ze nog wat onzeker is, en in de les zag ik dit terug op alle mogelijke manieren. Maar het lijkt erop dat de leerlingen dit simpelweg niet in de gaten hebben. (A7)
B2	*Ik* haal uit de zelfevaluatie dat de docent haar onzekerheid onder controle heeft nu. Ze moet nog werken aan leiding geven, maar ze is zich daarvan bewust, en kan daar dus aan gaan werken. (C3) Maar jij zegt dat je daarvan niet overtuigd bent.... (A13)
B1	Inderdaad. Ik betwijfel of ze hieraan kan werken: ten eerste wordt op allerlei plaatsen in de zelfevaluatie onzekerheid zichtbaar en ten tweede zoekt ze vaak een excuus om ergens niet aan hoeven te werken [geeft enkele concrete voorbeelden uit de zelfevaluatie]. (C3)
B2	Nou ja, ik heb wel de vraag geformuleerd of ze uit sociaal wenselijkheid zegt dat ze bepaalde dingen wil veranderen, of dat ze deze ook *echt* wil veranderen." (C2)
B1	…op een bepaald moment geeft ze aan dat ze tijdens plenaire momenten de aandacht van de *hele* klas vast wil houden door niet constant persoonlijke aandacht te geven aan individuele leerlingen. Dus hier presenteert ze een duidelijk plan. Maar vervolgens zegt ze "dat is moeilijk en eigenlijk vind ik het helemaal niet prettig om het anders te doen". Dus ik heb erachter gezet: "dus het gaat mislukken!" (C3)
B2	Ja, oké… Maar ik zie dit niet als hopeloos, ze is nog in opleiding, ze heeft nog tijd om dit te ontwikkelen. (C1c)
B1	[leest een andere passage uit de zelfevaluatie als voorbeeld van excuses zoeken] (C3)
B2	Ik ben het met je eens dat dat niet erg sterk is (B3), maar ik vind het ook wel weer logisch dat je een zelfevaluatie opdracht gebruikt om dat soort onzekerheden te benoemen. Ik denk dat ze op het gebied van structuur bieden aan leerlingen op de hoogte is van haar sterke en zwakke punten en ik vind dat ze ook inzicht heeft in de manier waarop ze zichzelf kan ontwikkelen. Ik denk dus dat ze een onwijze verbetering kan maken. (C3)

Tabel 4

Communicatieve activiteiten die boven- en ondergemiddeld voorkomen in Type III beoordelingsprocessen (aanvullen)

Activiteiten		Gemiddelde	
		Type III	Totaal
Activiteiten die **bovengemiddeld** voorkomen:			
Inbreng	Expliciet nagaan of al het relevante bewijs in overweging genomen is	1,4	0,8
Acceptatie	Aandragen van aanvullend bewijsmateriaal	13,8	10,3
Activiteiten die **ondergemiddeld** voorkomen:			
Discussie	Uitdagen van een interpretatie door aandragen van confronterend bewijsmateriaal of een alternatieve interpretatie	3,2	6,0

Type III: Aanvullen

De beoordelingsprocessen van vijf van de twaalf beoordelaarsparen, paar E tot en met I, bleken voornamelijk gebaseerd te zijn op aanvullen, ofwel aandragen van aanvullend bewijs voor een specifieke interpretatie of specifiek oordeel. Tabel 4 laat zien dat *expliciet nagaan of al het relevante bewijs in overweging genomen is* een andere activiteit is die in de gesprekken van deze duo's vaker voorkwam dan gemiddeld. De vijf Type III paren zijn verantwoordelijk voor zeven van de totaal acht keer dat deze activiteit voorkomt in de data (zie Tabel 1). Nagaan of al het relevante bewijs in overweging genomen is in het totaaloordeel behelst een van de kernprincipes van valide beoordeling.

De beoordelingsprocessen van Type III paren hadden bovendien gemeenschappelijk dat inbreng minder vaak dan gemiddeld werd bediscussieerd. Nadere inspectie van de voorkomende discussies laat zien dat deze in het algemeen bestaan uit slechts één discussieactiviteit (in tegenstelling tot een reeks, zoals te zien bij Type I- en II-discussies). Het onderwerp van discussie betreft in de meeste gevallen slechts een klein detail. Vier van de tien discussies die werden geteld binnen de gesprekken van deze vijf paren resulteerden in een expliciete overeenstemming, twee bleven onbeslist en vier werden genegeerd. De analyse van de beoordelingsprocessen gekarakteriseerd als Type III laat zien dat de gesprekken van deze paren een erg concreet niveau hebben – het noemen van concrete observaties staat centraal. Een kenmerkend fragment voor een Type III interactie is opgenomen in Tekstbox 3.

Tekstbox 3

Voorbeeld van een Type III gezamenlijk beoordelingsproces: Aanvullen

B1	Ik vind haar analyse van haar eigen gedrag en de klassensituatie niet erg sterk. Ze maakt haar analyses nooit echt *af*. (A1)
B2	Nee, ik vind haar analyse niet zo overtuigend. Ze spreekt zichzelf ook constant tegen. (B2)
B1	Kunnen we daar een voorbeeld van geven? Ehm, (…) denk je dat regel 83 een voorbeeld is? (B2)
B2	Ehm, even kijken (…). Ik heb regel 85 tot 87 als voorbeeld gegeven. O ja, oké, ja, dat is hetzelfde voorbeeld als dat van jou. Ehm, een ander voorbeeld is te vinden rond regel 125, waar de docent zegt dat ze het goed vind dat leerlingen overleggen tijdens haar uitleg, maar dat ze wel wil dat ze opletten. (B2)
B1	Wat is dat toch hè, dat dio's zo graag willen toestaan dat leerlingen discussiëren tijdens de uitleg?
B2	Discussiëren en opletten is in mijn ogen nogal tegenstrijdig.
B1	Ja, precies! Ze lijkt echt te denken dat leerlingen *de lesstof* bespreken…
B2	Dus ons punt is dat de docent haar analyses niet afmaakt, en dat zij zichzelf regelmatig tegenspreekt. Ook is ze niet in staat om concrete plannen voor verbetering te formuleren (A3)
B1	Nee, ze formuleert in het geheel geen plannen. En dat sluit dus naadloos aan op het feit dat ze nog niet weet wat ze wil verbeteren. Ik bedoel, als je analyses niet adequaat zijn…(A6)

Tabel 5

Communicatieve activiteiten die boven- en ondergemiddeld voorkomen in Type IV beoordelingsprocessen (aanvullen noch confronteren)

Activiteiten		Gemiddelde	
		Type IV	Totaal
Activiteiten die **bovengemiddeld** voorkomen:			
Inbreng	Constateren van overeenstemming tussen de beoordelaars	3,3	1,5
Activiteiten die **ondergemiddeld** voorkomen:			
Inbreng	Inbrengen van een voorstel voor de te volgen aanpak / procedure	0,3	2,6
Acceptatie	Aandragen van aanvullend bewijsmateriaal	4,3	10,3
Discussie	Uitdagen van een interpretatie door aandragen van confronterend bewijsmateriaal of een alternatieve interpretatie	1,7	6,0

Type IV: Aanvullen noch confronteren

De gesprekken van drie van de twaalf beoordelaarsparen werden gekarakteriseerd door aanvullen noch door confrontatie. Een activiteit die binnen de gesprekken van deze duo's vaker voorkwam dan gemiddeld is *constateren van overeenstemming tussen de beoordelaars* (zie Tabel 5). De bevinding dat deze beoordelaars veelal in een vroeg stadium overeenstemming constateerden verklaart wellicht dat confrontatie binnen deze duo's ver onder het gemiddelde lag.

Net als in de gesprekken van Type III-paren, kwam discussie in de gesprekken van Type IV-paren lager dan gemiddeld voor. Ook was de *aard* van de discussies van Type IV-paren vergelijkbaar met die van Type III-paren: discussies omvatten doorgaans niet meer dan één discussie-activiteit. In totaal vijf van de elf discussie interacties die voorkwamen in de gesprekken van de Type IV-paren, resulteerden in overeenstemming, drie bleven onbeslist en drie werden genegeerd. Een fragment van een typische Type IV-interactie is opgenomen in Tekstbox 4. De analyse van de beoordelingsprocessen gekarakteriseerd als Type IV wees uit dat gesprekken van deze paren ofwel erg abstract bleven, met relatief weinig verwijzing naar concreet bewijsmateriaal, of juist erg concreet – met het karakter van opnoemen van concrete observaties, zonder interpretatie hiervan.

Tekstbox 4

Voorbeeld van een Type IV gezamenlijk beoordelingsproces: Aanvullen noch confronteren

B1	Maar wat ook heel wezenlijk is, uit de VIL-resultaten blijkt een behoorlijke discrepantie tussen wat de leerlingen vinden en wat zij vindt. Ze heeft eigenlijk een heel matig zelfbeeld. (A2)
B2	Een matig zelfbeeld? Nee, dat ben ik toch niet met je eens. (C1d)
B1	Zij vindt zichzelf minder leidend dan de leerlingen haar ervaren, zij vindt zichzelf niet streng genoeg in dit geval hè? Dus dit is een typisch geval van iemand die begeleiding behoeft op dit gebied. (A1, A2)
B2	Ja, en op een aantal gebieden andere gebieden ook.(B3)
B1	Oké, maar nu even wat haar zelfbeeld betreft.
B2	Ik vind ook dat ze begeleiding moet krijgen op het gebruik van haar stem om maar eens wat te noemen. Ik weet niet of dat allemaal aan die video heeft gelegen, en op het gebruik van hoe zullen we het noemen? Lichaamstaal? Mimiek? Ehm, verteltechniek? Want zij heeft een uitermate boeiend onderwerp voor die leerlingen, aids in Afrika, eh, demonstratie van een condoom, spannender kun je het eigenlijk niet maken, en toch lukt het haar niet om de leerlingen geïnteresseerd te houden. Voor jou en mij zou dat een eitje zijn, met dit onderwerp. Ik zie dus een paar punten waar ze heel nadrukkelijk op gecoached moet worden. (A2, D)
B2	Ja, ja. Oké.(B3)

5 Conclusie en discussie

In dit artikel presenteerden we een overzicht van de communicatieve activiteiten die beoordelaars ondernemen wanneer zij gezamenlijk een docent-in-opleiding beoordelen. Uit de resultaten van de studie blijkt dat beoordelaars duidelijk activiteiten uitvoeren conform de genoemde beoordelingsprincipes: a) zoeken naar coherentie tussen de verschillende databronnen en nagaan of al het relevante bewijs in overweging is genomen en b) uitdagen van de zich ontwikkelende totaalindruk door actief op zoek te gaan naar tegenbewijs of alternatieve interpretaties. De resultaten laten zien dat deze twee beoordelingsprincipes het best worden gerepresenteerd door de activiteiten aanvullend bewijsmateriaal aandragen (*aanvullen*) en uitdagen van een interpretatie door het aandragen van tegenbewijs of het presenteren van een alternatieve interpretatie (*confronteren*). Het blijkt echter moeilijk voor beoordelaars om elkaar in de discussie zowel aan te vullen als uit te dagen: slechts drie van de twaalf paren lieten een gezamenlijk beoordelingsproces zien waarin zij elkaar zowel *aanvullen* als *confronteren* (Type I). De discussie van de overige paren was voornamelijk gericht op het elkaar *confronteren* (Type II), *aanvullen* (Type III) of *aanvullen* noch *confronteren* (Type IV). We veronderstelden dat een samenwerking die gericht is op zowel aanvullen als confronteren (Type I) leidt tot de meeste valide oordelen, dat wil zeggen oordelen gebaseerd op *al* het relevante bewijsmateriaal, zowel bewijs als tegenbewijs. Hieronder besteden we per type beoordelingsproces aandacht aan de specifieke sterke punten en valkuilen.

De kracht van een beoordelingsproces dat is gebaseerd op zowel aanvullen als confronteren *(Type I)* is dat dit in potentie leidt tot oordelen waarin al het relevante bewijs in overweging genomen is: zowel bewijs dat de aanvankelijke indruk ondersteunt als tegenbewijs. De resultaten laten zien dat de argumentatie van Type I-beoordelaarsparen relatief uitgebreid en transparant is voor alle essentiële aspecten van (interpersoonlijke) docentcompetentie. De gesprekken van Type I-beoordelingsparen omvatten een aantal echte discussies, waarvan het merendeel leidt tot een expliciete conclusie. Dit impliceert dat beoordelaars op basis van discussie met de medebeoordelaar hun aanvankelijke indruk bijstellen en/of aanscherpen, en dat geëxpliciteerd wordt hoe is omgegaan met tegenbewijs en alternatieve interpretaties.

Een duidelijke kracht van het gezamenlijke beoordelingsproces van paren die elkaar voornamelijk confronteren *(Type II)* is dat hun oordelen met betrekking tot *die* aspecten van interpersoonlijke competentie waarover discussie ontstaat uiterst zorgvuldig worden onderbouwd met bewijsmateriaal, waarbij zowel bevestigend bewijs als tegenbewijs een plaats krijgen in de argumentatie. Dit impliceert dat beoordelaars op basis van discussie hun argumentatie bijstellen en/of aanscherpen. Echter, beoordelaars lijken sterk gericht te zijn op het bereiken van overeenstemming over een selectie van aspecten waarover op het eerste gezicht meningsverschillen ontstaan. De overige aspecten krijgen minder aandacht in het gesprek en de argumentatie met betrekking tot deze aspecten blijft vrij abstract, met vrijwel geen verwijzing naar relevant bewijsmateriaal.

In een beoordelingsproces dat is gericht op het aanvullen van bewijsmateriaal *(Type III)* wordt expliciet welk specifiek bewijs geleid heeft tot een bepaalde interpretatie of een bepaald oordeel. Maar hoewel de activiteit *expliciet nagaan of al het relevante bewijsmateriaal in overweging genomen is* in Type III-gesprekken vaker voorkomt dan gemiddeld, zoeken beoordelaars voornamelijk naar bewijs dat de zich ontwikkelende indruk ondersteunt en komt vrijwel geen tegenbewijs ter sprake dat de zich ontwikkelende indruk zou kunnen uitdagen. Een ander gevaar in dit type beoordelingsproces kan zijn dat beoordelaars zich beperken tot het opnoemen van bewijsmateriaal, zonder werkelijke interpretatie hiervan of zonder coherentie te zoeken. Type III beoordelingsprocessen omvatten daarnaast relatief weinig discussie tussen beoordelaars. Discussie beperkt zich veelal tot geïsoleerde discussiebijdragen die betrekking hebben op details en deze discussies blijven relatief vaak onbeslist. Dit impliceert dat beoordelaars discussies niet gebruiken om hun argumentatie bij te stellen of aan te

scherpen en dat niet expliciet wordt hoe is omgegaan met tegenbewijs en alternatieve interpretaties.

In de samenwerking van paren die elkaar aanvullen noch uitdagen *(Type IV)* kunnen geen specifieke sterke punten worden onderscheiden. Hoewel een gebrekkige onderbouwing met bewijs ofwel een gebrekkige argumentatie voor bepaalde aspecten van het eindoordeel voorkomt bij alle beoordelaarsparen, is dit in het bijzonder voor Type IV-paren een valkuil. De gesprekken van deze paren blijven ofwel erg abstract, met relatief weinig verwijzing naar concreet bewijsmateriaal, of juist erg concreet – met het karakter van opnoemen van concrete observaties – zonder interpretatie hiervan. Net als Type III-beoordelingsprocessen omvatten Type IV-processen relatief weinig discussie tussen beoordelaars en leidt discussie zelden tot expliciete conclusies. Dit impliceert dat beoordelaars de discussie niet gebruiken om hun argumentatie bij te stellen of aan te scherpen.

Om discussie uit te kunnen lokken moet de discussie boven het niveau van het concrete bewijs uitstijgen (vgl. Delandshere & Petrosky, 1994; Mislevy, Steinberg, Breyer, Almond, & Johnson, 2002; Moss, et al, 1998). Onze data suggereren dat extreem holistische oordelen, met weinig verwijzing naar relevant bewijsmateriaal (gevaar in Type IV-processen) en oordelen die niet veel verder gaan dan een uitzetting van het gevonden bewijsmateriaal (gevaar in Type III- en Type IV-processen) weinig leiden tot discussie, het inbrengen van tegenbewijs en het inbrengen van alternatieve interpretaties. Binnen alle types kwam echter voor dat discussies onbeslist bleven of dat bijdragen die discussie omvatten zelfs volledig werden genegeerd. Zelfs de processen van de beoordelaars die erin slaagden om elkaar zowel aan te vullen als uit te dagen zouden in dit opzicht dus nog verder verbeterd kunnen worden. Discussies van deze beoordelaars leidden niet altijd tot duidelijke conclusies, wat zou kunnen betekenen dat het aangedragen tegenbewijs en alternatieve interpretaties niet werden meegenomen in het totaaloordeel. Dit betekent dat ook deze beoordelaars er niet geheel in zijn geslaagd om *alle* tegenbewijs en tegenargumenten volledig te verklaren en een expliciete plek te geven in de argumentatie die ten grondslag lag aan het eindoordeel.

Op basis van de resultaten van de studie kan worden geconcludeerd dat een beoordelingsproces waarin zowel a) expliciet coherentie wordt gezocht tussen het beschikbare bewijsmateriaal totdat al het relevante bewijsmateriaal in overweging genomen is als b) de zich ontwikkelende totaalindruk actief wordt uitgedaagd met tegenbewijs en/of alternatieve interpretaties van bewijs, moeilijk te realiseren is voor beoordelaars. Omdat een dergelijk proces al moeilijk te realiseren blijkt voor duo's, veronderstellen we dat dit nog moeilijker te realiseren zal zijn voor individuele beoordelaars. Duo's kunnen elkaar aanvullen in hun argumentatie en kunnen elkaar corrigeren voor bijvoorbeeld oneven-redige aandacht voor bepaalde aspecten van interpersoonlijke competentie en/of 'blinde vlekken'. Ook ligt het in de rede dat de kans dat mogelijk tegenbewijs wordt opgemerkt en alternatieve interpretaties van het bewijsmateriaal in overweging genomen worden groter is bij dialoog c.q. discussie tussen beoordelaars dan bij individuele oordeelsvorming.

De resultaten van deze studie hebben een aantal specifieke moeilijkheden c.q. valkuilen aan het licht gebracht, die naar onze mening zowel gelden bij paarsgewijze beoordeling als bij individuele beoordeling. Het bleek lastig voor beoordelaars om hun oordeel voor *alle* aspecten van interpersoonlijke competentie te onderbouwen op basis van zowel concreet bewijs als tegenbewijs (valkuil Type II-, III- en IV-processen). Eerder suggereerden we dat extreem holistische oordelen, met weinig verwijzing naar relevant bewijsmateriaal (gevaar in Type IV-processen) en oordelen die niet veel verder gaan dan een uitzetting van het gevonden bewijsmateriaal (gevaar in Type III- en Type IV-processen) weinig leiden tot discussie, het inbrengen en in overweging nemen van tegenbewijs, en het inbrengen en in overweging nemen van alternatieve interpretaties. Hier kan aan toegevoegd worden dat zowel in een erg holistische aanpak als in een aanpak waarin bewijsmateriaal wordt opgesomd zonder verdere interpretatie, geen sprake is

van coherentie zoeken in het specifieke bewijsmateriaal.

Een beperking van het onderzoek betreft het feit dat de vierentwintig beoordelaars slechts één dio beoordeelden. Om zoveel mogelijk variatie te genereren in de beoordelingsprocessen van beoordelaars kozen we voor een maximum aantal beoordelaars en dus niet voor een groter aantal te beoordelen dio's. De dio werd geselecteerd op basis van een 'gemiddeld profiel' op het gebied van interpersoonlijke competentie, met een aantal voor dio's kenmerkende moeilijkheden op dit gebied. Wanneer een groter aantal dio's beoordeeld zou worden, zou dit mogelijk een verfijnder inzicht kunnen geven in beoordelingsprocessen. Met name cases van dio's waarin duidelijke incongruentie bestaat tussen de verschillende databronnen, c.q. cases die veel onduidelijkheid oproepen, zouden interessante gegevens kunnen opleveren over de wijze waarop beoordelaars coherentie zoeken in het beschikbare bewijsmateriaal en de wijze waarop zij hun zich ontwikkelende indruk uitdagen met tegenbewijs. Om dieper inzicht te krijgen in de wijze waarop beoordelaars in de praktijk invulling kunnen geven aan deze beoordelingsprincipes zou in vervolgonderzoek tevens gebruik gemaakt kunnen worden van aanvullende manieren van dataverzameling. Gedacht kan worden aan hardop-denksessies en/of interviews met expert beoordelaars, oftewel beoordelaars die duidelijk in staat zijn om invulling te geven aan de beoordelingsprincipes.

Een tweede kanttekening bij het onderzoek betreft het gegeven dat het aantal paren in de verschillende types klein was: onder type II viel zelfs maar één beoordelingspaar. Dit heeft mogelijk gevolgen voor de kwaliteit van de omschrijving van de vier types en hun specifieke sterke punten en valkuilen. Het zou de moeite waard zijn om de karakterisering van de vier types te testen aan de hand van de analyse van de beoordelingsprocessen van een groter aantal beoordelaars of op basis van een groter aantal te beoordelen dio's.

Ten derde plaatsen we een kanttekening bij het gehanteerde categorieënsysteem. Dit systeem is deels ontleend aan de literatuur, deels aangevuld op basis van de data die in het onderzoek verzameld zijn. In vervolgonderzoek zou het systeem op grotere en meer gevarieerde schaal kunnen worden getest.

Ten slotte merken we op dat de beoordelingsparen in de onderhavige studie op basis van willekeur door beoordelaars zelf werden samengesteld. Hoewel beoordelaars met achtergronden in verschillende disciplines (alfa, bèta en gamma) en met verschillende expertise en karaktereigenschappen paarsgewijs samenwerkten, hebben we in de onderhavige studie niet onderzocht of de samenstelling gevolgen had voor de aard van het gezamenlijke beoordelingsproces. In vervolgonderzoek zou de vraag kunnen worden gesteld welke samenstelling van beoordelaars de meeste toegevoegde waarde heeft, oftewel de kans optimaliseert dat beoordelaars hun individuele oordelen verbeteren door elkaar aan te vullen en te confronteren.

6 Aanbevelingen

Om tot een valide beoordeling van docentcompetentie te komen achten we het niet alleen van belang dat beoordelaars weten aan welke kenmerken een valide *oordeel* zou moeten voldoen (bijvoorbeeld *geen* construct onderrepresentatie en/of constructirrelevante variantie), maar ook dat zij zicht hebben op de kenmerken van een valide beoordelings*proces*. In de onderhavige studie introduceerden we twee principes van een valide beoordelingsproces: a) zoeken naar coherentie tussen de verschillende databronnen en nagaan of al het relevante bewijs in overweging is genomen en b) uitdagen van de zich ontwikkelende totaalindruk door actief op zoek te gaan naar tegenbewijs of alternatieve interpretaties. Het overzicht van communicatieve activiteiten en de vier typen (gezamenlijke) beoordelingsprocessen, inclusief hun sterke en zwakke punten, lijken een bruikbare aanzet te bieden om zowel individuele als gezamenlijke beoordelingsprocessen te analyseren en na te gaan op welke wijze de kwaliteit van het beoordelingsproces verbeterd zou kunnen worden. Dit kan de discussie over een valide beoordeling en het voorkomen van *bias*, constructirrelevante variantie en constructonderrepresentatie concreter maken.

Hoewel de resultaten in principe gebruikt zouden kunnen worden om praktische richtlijnen te formuleren voor een valide beoordelingsproces, zijn we van mening dat het formuleren van al te specifieke richtlijnen moet worden voorkomen. De studie van Moss e.a. (1998) liet zien dat beoordelaars de neiging hebben om erg rigide om te gaan met dergelijke richtlijnen, wat ten koste kan gaan van diepgang in de interpretatie van het beschikbare bewijsmateriaal. Een aanbeveling is daarom om beoordelaars niet een al te strak keurslijf van richtlijnen aan te bieden, maar in plaats daarvan beoordelaars te stimuleren om kritisch naar hun eigen beoordelingsproces te kijken. Na een beoordeling of serie van beoordelingen zouden zij aangemoedigd kunnen worden om stil te staan bij vragen als: Hoe beoordelen zij de kwaliteit van hun eigen beoordelingsproces? Hebben zij al het relevante bewijsmateriaal terug laten komen in hun uiteindelijke beoordeling? Zijn zij actief op zoek gegaan naar tegenbewijs of alternatieve interpretaties? Zijn deze tegenvoorbeelden ook daadwerkelijk verklaard en geïntegreerd in het totaaloordeel? Wat zouden zij kunnen doen om de kwaliteit van hun proces en de transparantie van hun argumentatie te verbeteren?

Noot

1 Dit onderzoek werd gefinancierd door de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO-projectnummer 411-21-205).

Literatuur

Barron, B. (2003). When smart groups fail. *The Journal of the Learning Sciences*, *12*, 307 - 359.

Cochran-Smith, M. (2003). The unforgiving complexity of teaching: Avoiding simplicity in the age of accountability. *Journal of Teacher Education, 54*, 3 - 5.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Darling-Hammond, L., & Snyder, J. (2000). Authentic assessment of teaching in context. *Teaching and Teacher Education*, *16*, 523 - 545.

Delandshere, G., & Arens, S. A. (2003). Examining the quality of the evidence in preservice teacher portfolios. *Journal of Teacher Education*, *54*, 57 - 73.

Delandshere, G., & Petrosky, A. R. (1994). Capturing teachers' knowledge: Performance assessment a) and post-structuralist epistemology b) from a post-structuralist perspective c) and post-structuralism d) none of the above. *Educational Researcher*, *23*, 11 - 18.

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

Dwyer, C.A. (1995). Criteria for performance-based teacher assessments: Validity, standards and issues. In A.J. Shinkfield, & D. Stufflebeam (Eds.), *Teacher evaluation: Guide to effective practice* (pp. 62 - 80). Boston: Kluwer Academic Publishers.

Dwyer, C. A., & Stufflebeam, D. (1996). Teacher evaluation. In D. C. Berliner & R.C. Calfee (Eds.), *Handbook of Educational Psychology* (pp. 765 - 768). New York: Macmillan.

Evans, E.D., & Tribble, M. (1986). Perceived teaching problems, self-efficacy, and commitment of teaching among preservice teachers. *Journal of Educational Research, 80*, 81 - 85.

Evertson, C. M., & Weinstein, C.S. (2006). Classroom management as a field of inquiry. In C.M. Evertson & C.S. Weinstein (Eds.), *Handbook of classroom management: Research, practice, and contemporary issues* (pp. 3 - 16). Mahawn, NJ: Lawrence Erlbaum Associates.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. London: Sage.

Hager, P., Gonczi, A., & Athanasou, J. (1994). General issues about assessment of competence. *Assessment and Evaluation in Higher Education, 19*, 3 - 16.

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

Johnston, B. (2004). Summative assessment of portfolios: An examination of different approaches to agreement over outcomes. *Studies in Higher Education*, *29*, 395 - 412.

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

Mabry, L. (1999). *Portfolio's plus, a critical guide to alternative assessment.* Thousand Oaks, CA: Corwin Press.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13 - 103). New York: MacMillan.

Messick, S. (1996). Validity of performance assessments. In G.W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 1 - 18). Washington DC: Deparment of education, office of Educational research and Improvement.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher*, *23*, 5 - 12.

Moss, P. A., Schutz, A. M., & Collins, K. M. (1998). An integrative approach to portfolio evaluation for teacher licensure. *Journal of Personnel Evaluation in Education*, *12*, 139 - 161.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L.(2002). Making sense of data from complex assessments. *Applied Measurement in Education, 15*, 363 - 389.

Nijveldt, M., Beijaard, D., Brekelmans, M., Verloop, N., & Wubbels, Th. (2005). Assessing the interpersonal competence of beginning teachers: The quality of the judgement process. *International Journal of Educational Research*, *43*, 89 - 102.

Quinlan, K. M. (2002). Inside the peer review process: How academics review a colleague's teaching portfolio. *Teaching and Teacher Education*, *18*, 1035 - 1049.

Schutz, A., & Moss, P.A. (2004). Reasonable decisions in portfolio assessment: Evaluating complex evidence of teaching. *Education Policy Analysis Archives*, *12*, 33.

Tigelaar, D. E. H., Dolmans, D. H. J. M, Wolfhagen, I. H. A. P., & Vleuten, C. P. M. van der. (2005). Quality issues in judging portfolios: Implications for organizing teaching portfolio assessment procedures. *Studies in Higher Education, 30*, 595 - 610.

Uhlenbeck, A. M., Verloop, N., & Beijaard, D. (2002) Requirements for an assessment procedure for beginning teachers: Implications from recent theories on teaching and assessment. *Teachers College Record, 104*, 242 - 272.

Wolf, A. (1995). *Competence-based assessment.* Buckingham, UK: Open University Press.

Wubbels, Th., Brekelmans, M., Brok, P., den, & Tartwijk, J. van. (2006). An interpersonal perspective on classroom management in secondary classrooms in the Netherlands. In. C. Evertson & C. S. Weinstein (Eds.), *Handbook of classroom management: research, practice and contemporary issues* (pp. 1161 - 1191). New York: Lawrence Erlbaum Associates.

Manuscript aanvaard: 19 mei 2008

Auteurs

Mirjam Nijveldt is als onderzoeker werkzaam bij het Instituut voor Leraar en School (ILS), Radboud Universiteit Nijmegen.

Mieke Brekelmans is hoogleraar onderwijskunde, Faculteit Sociale Wetenschappen, Universiteit Utrecht.

Douwe Beijaard is als hoogleraar verbonden aan de Eindhoven School of Education (ESoE), een gemeenschappelijk instituut van Fontys Hogescholen en de Technische Universiteit Eindhoven.

Theo Wubbels is hoogleraar onderwijskunde en vice-decaan van de Faculteit Sociale Wetenschappen, Universiteit Utrecht.

Nico Verloop is als hoogleraar-directeur verbonden aan het Interfacultair Centrum voor Lerarenopleiding, Onderwijsontwikkeling en Nascholing (ICLON), Universiteit Leiden.

Correspondentieadres: M. Nijveldt, ILS, Radboud Universiteit Nijmegen, Postbus 9103, 6500 HD Nijmegen, e-mail: *m.nijveldt@ils.ru.nl*

Abstract

Validity in collaborative teacher assessment

Given that the assessment of student teachers is generally based on non-standardized, qualitative information derived from multiple sources, the validity of the assessment process largely depends on the judgement capacities of the assessors. Although it has recently been suggested that the quality of the assessment process can be improved via collaboration in pairs, there is, however, little empirical research on the nature of the collaborative assessment process or the ways in which such collaboration can enhance the validity of assessment. In the present study, twelve assessor pairs were asked to collaboratively judge the same student teacher. The assessment process was subsequently characterized in terms of the specific communicative activities engaged in. Four different types of collaborative assessment processes could be distinguished and, for each of these, a number of strengths and weaknesses could be identified. The implications of this information for the validity of collaborative assessment processes and the preparation of assessors are discussed.