De invloed van cognitieve representaties van beoordelaars op hun beoordeling van docentportfolio's

M. van der Schaaf, K. Stokking en N. Verloop[1]

Samenvatting

Tegenwoordig wordt bij het beoordelen van docenten vaak gebruikgemaakt van portfolio's. De betrouwbaarheid van portfoliobeoordelingen is een heikel punt. Zulke beoordelingen worden beïnvloed door cognitieve representaties van de beoordelaars. Inzicht daarin is cruciaal om de betrouwbaarheid van portfoliobeoordelingen te kunnen verbeteren. We onderzochten zowel de betrouwbaarheid van beoordelingen als de cognitieve representaties van beoordelaars zoals die tot uitdrukking komen in retrospectieve hardopdenkprotocollen en beoordelingsformulieren. Door beide aan elkaar te relateren konden we nagaan in hoeverre beoordelingen kunnen worden verklaard vanuit cognitieve representaties van beoordelaars. In het onderzoek beoordeelden zes beoordelaars paarsgewijs 18 portfolio's. Bij 12 van deze portfolio's was de interbeoordelaarsbetrouwbaarheid redelijk tot goed. Variantieanalyse wees nauwelijks op beoordelaarseffecten. We gebruikten de Associated Systems Theory (Carlston, 1992, 1994) en de Correspondent Inference Theory (Jones & Davis, 1965) voor analyse van de inhoud van de hardopdenkprotocollen en beoordelingsformulieren. De gegeven beoordelingen konden grotendeels worden verklaard door categorisering van de cognitieve representaties van de beoordelaars op twee dimensies: abstracte versus concrete opmerkingen en positieve versus negatieve evaluatie.

1 Inleiding

De toegenomen aandacht voor kwaliteitsbewaking en verantwoording in het onderwijs en voor docenten als beroepsgroep gaat internationaal gepaard met onderzoek naar de beoordeling van hun competenties. Daarbij wordt vaak gebruikgemaakt van portfolio's. Een portfolio bestaat uit een selectie van over een langere periode verzameld documentatiemateriaal dat een beeld geeft van de wijze waarop een docent onderwijstaken uitvoert. Afhankelijk van de inhoud en vormgeving van het portfolio en de inbedding ervan in de werkomgeving van de docent kan een portfolio recht doen aan de contextgebondenheid van het onderwijzen en aan het uitgangspunt dat docentgedrag onlosmakelijk is verbonden met docentcognities (Andrews & Barnes, 1990; Bird, 1990; Lyons, 1998).

Bij het beoordelen van portfolio's is sprake van interacties tussen het portfoliomateriaal, de condities waaronder het portfolio functioneert, de beoordelingscriteria, en kenmerken en interpretaties van beoordelaars. Onderzoeken hebben aangetoond dat in het algemeen de variantie in beoordelingen grotendeels is toe te schrijven aan cognitieve activiteiten van de beoordelaars tijdens het beoordelen (Landy & Farr, 1980; Feldman, 1981; DeNisi, Cafferty, & Meglino, 1984). Deze activiteiten kunnen worden omschreven vanuit de inhoud van de cognities (bijvoorbeeld gericht op de te beoordelen docent of op persoonlijke opvattingen van de beoordelaar), het type cognitie (bijvoorbeeld interpretatie of beoordeling) en de aard van de cognities (meer of minder onbewust, gesitueerd en persoonlijk). In dit onderzoek richten we ons op de inhoud en op het type cognitieve activiteiten, in het bijzonder de cognitieve representaties bij de beoordelaars.

Er zijn diverse algemene modellen van cognitieve beoordelingsactiviteiten ontwikkeld (Gilbert, 1989; Jones & Davis, 1965). Cognitieve representaties bij het beoordelen van docentportfolio's zijn echter nog nauwelijks onderzocht. We achten inzicht in cognitieve representaties van beoordelaars noodzakelijk om portfoliobeoordelingen beter te kunnen begrijpen en de kwaliteit van de beoordelingen te kunnen verbeteren. Bij beoordelingen op basis van portfolio's is met name de betrouwbaarheid een knelpunt.

Ons onderzoek richtte zich op de volgende vragen:

1 Wat is de betrouwbaarheid van docent-portfoliobeoordelingen?
2 Welke cognitieve representaties gebruiken beoordelaars tijdens het beoordelen?
3 In hoeverre zijn de cognitieve representaties gerelateerd aan de beoordelingen en aan de betrouwbaarheid van de beoordelingen?

Het onderzoek maakt deel uit van een omvangrijker onderzoek naar het beoordelen van docentcompetenties van ervaren docenten bij het instrueren, begeleiden en beoordelen van onderzoeksvaardigheden van leerlingen in de bovenbouw havo-vwo in de gammavakken. Onderzoeksvaardigheden zijn een nieuw verplicht onderdeel in de examens in de Tweede Fase havo-vwo. Deze vernieuwing is representatief voor de verandering in veel landen naar meer constructivistische visies op leren en een grotere nadruk op de ontwikkeling van vaardigheden, zelfstandig leren en samenwerkend leren.

2 Beoordelingscriteria en portfolio's

Eerder hebben we onderzocht welke taken docenten zouden moeten vervullen bij het aanleren van onderzoeksvaardigheden, en over welke vermogens (kennis, vaardigheden en houdingen) ze daartoe zouden moeten beschikken. Op basis daarvan hebben we vervolgens met behulp van een delphi-panel met 21 'stakeholders' beoordelingscriteria ontwikkeld (Van der Schaaf, Stokking, & Verloop, ter publicatie aangeboden). Tevens hebben we een format voor portfolio's ontwikkeld (zie Methode). Deze criteria (in Tabel 1 afgezet tegen de onderdelen van het door ons gebruikte portfolio) zijn:

1 *Langetermijndoelen formuleren (DOEL).* Het hanteren van langetermijndoelen met betrekking tot het aanleren van onderzoeksvaardigheden door leerlingen.
2 *Geschikte onderzoeksopdracht kiezen (OPDR).* De opdracht is gericht op vakinhoudelijke en algemene vaardigheden, is authentiek, sluit aan bij de voorkennis van de leerlingen en biedt voldoende keuzemogelijkheden. Leerdoelen, inhoud en vorm sluiten bij elkaar aan.
3 *Het werken aan de opdracht door leerlingen voorbereiden en organiseren (ORGA).* Het gaat hier om het creëren van de faciliteiten (tijd, ruimtes, hulpmiddelen en bronnen) die leerlingen nodig hebben voor de opdracht en het informeren van de leerlingen daarover.
4 *Vooraf nadenken over instructie en begeleiding (DENK).* Kennen en kiezen van instructie- en begeleidingsvormen die aansluiten bij de voorkennis van de leerlingen en passen bij het doel van het leren doen van zelfstandig onderzoek, het gekozen onderzoeksonderwerp en de manier van beoordelen.
5 *Zelfstandig onderzoek instrueren en begeleiden (INSTR).* Gebruiken van instructie- en begeleidingsvormen die zelfstandig onderzoek in de gammavakken bevorderen.
6 *Een goed pedagogisch klimaat scheppen (KLIM).* Een veilige en stimulerende leeromgeving creëren.
7 *Op een adequate manier beoordelen (BEO).* Het vaststellen van doelen van de beoordeling, het gebruiken van heldere beoordelingscriteria en het aan de beoordeling verbinden van de juiste consequenties.
8 *Reflecteren op het onderwijsprogramma en het eigen handelen ten aanzien van zelfstandig onderzoek (REFL).* Aangeven van sterke en zwakke punten van het onderwijsprogramma en het eigen handelen, en suggesties doen voor verbetering.

Daarna (Van der Schaaf, Stokking, & Verloop, 2003) ontwikkelden we per criterium met behulp van 'policy capturing' een standaard, die aangeeft hoe goed docenten aan het criterium moeten voldoen om daarop een voldoende beoordeling te verkrijgen. We berekenden verder met multipele regressieanalyse de aan de criteria toegekende wegingsfactoren, waarbij bleek dat het panel aan de criteria INSTR en KLIM het meeste gewicht gaf.

3 Betrouwbaarheid van de beoordelingen

3.1 Psychometrische kwaliteitseisen

Portfoliobeoordeling wordt vaak gebruikt om

Tabel 1

Beoordelingscriteria en portfolio-onderdelen

	Criteria							
Onderdelen	DOEL	OPDR	ORGA	DENK	INSTR	KLIM	BEO	REFL
Zelfbeschrijving	X							X
Serie opdrachten	X							
Interviews			X	X	X	X		X
Video opnamen			X		X	X		
Onderzoeksopdracht	X	X	X		X		X	
Beoordelingen							X	
Reflectie								X
Leerlingevaluaties		X	X		X	X		

vast te stellen in welke mate docenten voldoen aan gewenste competenties (summatief) en om richtlijnen te formuleren voor hun verdere professionele ontwikkeling (formatief). Wij richten ons op beoordeling bij ervaren docenten met zowel summatieve als formatieve oogmerken.

De psychometrische of edumetrische kwaliteitseisen die aan zulke beoordelingen moeten worden gesteld, zijn nog niet uitgekristalliseerd. De klassieke criteria van betrouwbaarheid en validiteit zijn inmiddels nader gedifferentieerd, en aangevuld met criteria van aanvaardbaarheid en praktische bruikbaarheid (Messick, 1989; Stokking, Van der Schaaf, Jaspers, & Erkens, 2004). Aanvaardbaarheid betreft onder meer objectiviteit, inzichtelijkheid, gelijkwaardigheid en non-discriminatie. Praktische bruikbaarheid heeft onder meer betrekking op de functionaliteit, uitvoerbaarheid en doelmatigheid van de beoordeling.

Naar onze mening verdient bij zowel summatieve als formatieve beoordeling het criterium *validiteit* prioriteit (vgl. Linn, 1994; Linn, Baker, & Dunbar, 1991). De inzichtelijkheid, gelijkwaardigheid en doelmatigheid moeten ook altijd voldoende zijn. Bij formatieve beoordelingen hebben betrouwbaarheid, objectiviteit en gelijkwaardigheid minder prioriteit, omdat de beoordeling voornamelijk is gericht op mogelijkheden voor verbetering. De praktische bruikbaarheid is wel belangrijk om frequente en bruikbare feedback te kunnen geven. Bij summatieve beoordeling zijn ook betrouwbaarheid, objectiviteit en gelijkwaardigheid belangrijk (vgl. Stokking et al., 2004).

Het is momenteel onduidelijk in hoeverre summatieve portfoliobeoordelingen aan deze criteria kunnen voldoen. Met name de betrouwbaarheid is vaak teleurstellend (Burns, 1999; Johnson, McDaniel, & Willeke, 2000; Linn, 1994; Reckase, 1995; Shapley & Bush, 1999). Deze wordt veelal uitgedrukt in de interbeoordelaarsbetrouwbaarheid, waarbij wordt geschat in welke mate beoordelaars portfolio-inhouden op gelijke wijze classificeren. Een maximale interbeoordelaarsbetrouwbaarheid is theoretisch gezien alleen mogelijk als de beoordelaars de portfolio-inhouden en de beoordelingscriteria vrijwel exact hetzelfde interpreteren en vervolgens precies dezelfde scores toekennen bij de beoordeling. Aangezien dergelijke condities niet kunnen worden gegarandeerd (alleen al vanwege het feit dat interpretaties van mensen doorgaans uiteenlopen) (Huot, 1993; Kane, 1992; Van der Schaaf et al., 2003), gebruiken we de interbeoordelaarsbetrouwbaarheid slechts als indicatie van de betrouwbaarheid.

Daarnaast is variantieanalyse een geëigende methode om een inschatting te maken van de betrouwbaarheid van beoordelingen. Variantie in beoordelingen komt vaak voort uit 'bias' van individuele beoordelaars (bijvoorbeeld "halo-effecten") en uit systematische verschillen tussen beoordelaars. Deze laatste worden vaak in verband gebracht met persoonsgebonden factoren, zoals ervaringen en verwachtingen. In generaliseerbaarheidsstu-

dies wordt variantieanalyse gebruikt om te schatten in hoeverre diverse "foutenbronnen" bijdragen aan de variantie in beoordelingen. Hoewel sommige onderzoeken naar 'performance assessments' aantonen dat de variantie gerelateerd aan beoordelaars weinig bijdraagt aan de totale variantie (bijv. Shavelson, Baxter, & Gao, 1993), blijkt dit niet het geval bij complexere beoordelingstaken (Dunbar, Koretz, & Hoover, 1991).

3.2 Cognitieve representaties van beoordelaars en beoordelaartraining

De vooronderstelling dat cognitieve representaties van beoordelaars van invloed zijn op de beoordeling wordt gevoed door de sociaal cognitieve psychologie, waarin diverse modellen zijn ontwikkeld om beoordelingsactiviteiten te beschrijven (bijv. DeNisi et al., 1984; Feldman, 1981; Landy et al., 1980). Gemeenschappelijk aan deze modellen is dat beoordelaars op basis van hun schemata het gedrag van een persoon beoordelen, voorspellen en begrijpen. De schemata zijn vergelijkbaar met persoonlijke constructen (Kelly, 1955). Dat brengt met zich mee dat observaties, interpretaties en beoordelingen van anderen worden gefilterd door persoonlijke ervaringen en opvattingen van de beoordelaar.

Veel gebruikte modellen zijn gerelateerd aan de Correspondent Inference Theory van Jones en Davis (1965). Deze theorie veronderstelt dat het beeldvormingsproces van andermans gedrag kan worden uiteengelegd in verschillende activiteiten (Gilbert, 1989; Krull, 2001). Beoordelaars categoriseren gedrag (bijvoorbeeld: Irene spreekt haar leerlingen streng toe) en verbinden dit aan daarmee corresponderende kenmerken of eigenschappen (bijvoorbeeld: Irene is een strenge docente). Tegelijkertijd corrigeren beoordelaars hun typering door er situationele informatie bij te betrekken (bijvoorbeeld: de leerlingen van Irene zijn erg druk, Irene is in deze situatie terecht streng. Misschien is ze helemaal niet zo'n strenge docente). Tijdens de beoordeling kunnen beoordelaars positieve en negatieve opmerkingen maken (Huot, 1993). Gilbert (1989) suggereert dat beoordelingsactiviteiten zich meer of minder bewust en expliciet kunnen voltrekken. Met name algemene indrukken van anderen blijven vaak impliciet (Carlston, 1994). Verder blijkt het rekening houden met de situatie waarin gedrag plaatsvindt, niet vanzelfsprekend (wanneer beoordelaars onoplettend zijn, wordt daaraan vaak voorbij gegaan).

Het is belangrijk bij de ontwikkeling van beoordelaarstrainingen en beoordelingsformulieren aan te sluiten bij de activiteiten die beoordelaars uit zichzelf geneigd zijn uit te voeren. Uitkomsten van onderzoeken naar effecten van beoordelaarstrainingen suggereren dat beoordelingsactiviteiten zijn te beïnvloeden. Voorbeelden van veel gebruikte trainingsmethoden zijn: 'dimensional training', 'rater error training', 'behavioral observation training' en 'frame-of-reference' (FOR)-training. Reviewstudies tonen aan dat met name FOR-trainingen succesvol zijn. Dergelijke trainingen zijn gericht op het ontwikkelen bij de beoordelaars van een omvattende theorie over het beoordelingsproces. In FOR-trainingen worden de verschillende aspecten van een beoordelingsproces met elkaar gecombineerd (Lievens, 2001; Sulsky & Day, 1992; Woehr & Huffcutt, 1994). Het blijkt echter ook dat FOR-trainingen tot minder concrete gedragsgerichte opmerkingen leiden en tot minder interpretaties die de gegeven beoordeling ondersteunen dan observatietrainingen (Lievens, 2001). Dat is een nadeel, want met name bij formatieve beoordelingen is het kunnen geven van accurate feedback noodzakelijk.

Sceptici veronderstellen dat juist doordat cognitieve representaties van beoordelaars bij het beoordelen zo'n grote rol spelen, trainingen weinig effect zullen hebben. Immers, beoordelaars ondergaan in hun leven en loopbaan allerlei socialisatieprocessen. Van beoordelaarstrainingen kan niet worden verwacht dat daardoor cognitieve representaties, die zich gedurende een lange periode hebben ontwikkeld, gemakkelijk worden vervangen (Huot, 1993).

Toch vormen de cognities van beoordelaars een domein dat verdere exploratie verdient om beoordelingen te kunnen begrijpen en verbeteren (Day & Sulsky, 1995; Sulsky et al., 1992). We spitsten ons onderzoek toe op de aspecten waar beoordelaars op letten bij het vormen van indrukken van docenten. Om

deze te onderzoeken, gebruikten we de Associated Systems Theory (AST) (Carlston, 1992, 1994). De AST richt zich op diverse vormen van menselijke cognitieve representaties die tegelijkertijd optreden wanneer mensen zich een beeld vormen van anderen. De AST is reeds eerder voor het onderzoeken van beoordelingsactiviteiten van beoordelaars bruikbaar gebleken (bijv. Schleicher & Day, 1998).

3.3 Associated Systems Theory

Voortbordurend op eerder onderzoek in de sociale psychologie en neurologie (Fiske, 1992; Martindale, 1991) is de AST gebaseerd op de principes "doen leidt tot denken" en "denken leidt tot doen". Het eerste principe gaat ervan uit dat cognitieve representaties zich ontwikkelen door ervaringen van mensen die voortkomen uit hun (mentale en fysieke) activiteiten. Vertaald naar ons onderzoek worden cognitieve representaties bijvoorbeeld beïnvloed door onderwijservaring, beoordelaarstraining, en ervaring in het beoordelen van docentportfolio's. Ten tweede kunnen cognitieve representaties worden gezien als intermediair tussen de input van externe stimuli en de output van gedrag (in ons geval het geven van een oordeel) (Norman, 1985). Cognitieve representaties spelen dus een belangrijke rol bij het uitvoeren van taken, zoals het beoordelen van docentportfolio's in ons onderzoek.

De AST biedt een startpunt voor het classificeren van representaties van beoordelaars. Carlston (1992, 1994) modelleert de AST op twee dimensies (zie Tabel 2).

1 *Concreet versus abstract.* Concrete representaties (linkerkolom) zijn gebaseerd op tijd- en situatiespecifieke observaties, bijvoorbeeld het waarnemen van iemands fysieke verschijningsvorm of gelaatstrekken na een avond flink stappen. De vormen in de middelste kolom zijn abstracter, omdat ze voortkomen uit een cluster van observaties (bijvoorbeeld het aanduiden van iemands aantrekkelijkheid op basis van zijn of haar uiterlijke verschijningsvorm). Abstracte representaties zijn generieker en bevatten door de beoordelaar gepercipieerde algemene kenmerken van de beoordeelde, bijvoorbeeld het toekennen van eigenschappen (zoals lui of ijdel) op basis van iemands uiterlijk. Het beschrijven van personen in algemene kenmerken vergt over het algemeen meer cognitieve inspanning dan het concreet beschrijven van iemands uiterlijk. In die zin vertegenwoordigt deze dimensie ook een toename in cognitieve activiteit van de beoordelaars.
2 *Gericht op het doel versus gericht op zichzelf.* Hoewel cognitieve representaties altijd meer of minder subjectief en beoordelaarsgebonden zijn, kunnen beoordelingen variëren in hun mate van doelgerichtheid. Doelgerichte representaties zijn primair gericht op het doel of de persoon die wordt beoordeeld. Op zichzelf gerichte representaties betreffen persoonlijke reacties van de beoordelaar op de beoordeelde. Omdat persoonlijke reacties vaak zijn gebaseerd op relatief stabiele mentale structuren (vergelijk attitudes), zijn ze moeilijk veranderbaar. Bij beoordelen is doorgaans sprake van een mix van doelgerichte en zelfgerichte representaties (middelste rij). Beoordelaars interacteren immers altijd mentaal met de beoordelingssituatie en de beoordeelde, omdat hun persoonlijke schemata filteren wat ze waarnemen, en hoe ze wat ze waarnemen interpreteren en uiteindelijk beoordelen (Tulving, 1983). Dus ook al is er geen fysieke interactie tussen de beoordeelde en de beoordelaar en/of de beoordelingssituatie, van mentale interactie is wel degelijk sprake.

De specifieke cognitieve representaties in de cellen van de matrix worden door Carlston (1992, 1994) als volgt beschreven:

1a *Visuele manifestaties*: fysieke indrukken van anderen (verschijningsvormen of getoonde gedragingen). In ons onderzoek blijken deze bijvoorbeeld uit verwijzingen naar in de portfolio's opgenomen docentgedrag.
1b *Categorisaties (typeringen)*: dit betreft de labels die we toekennen aan de indrukken die we van anderen hebben. In ons onderzoek blijken deze bijvoorbeeld uit interpretaties of verklaringen van gedragingen van docenten, zoals afgeleid uit hun portfolio's.

Tabel 2

Structurele representatie van de AST-taxonomie

	Concreet		Abstract
Doel	(1a) Visuele manifestaties	(1b) Categorisaties	(1c) Toegekende persoonskenmerken
	(2a) Observaties	(2b) Mix	(2c) Evaluaties
Zelf	(3a) Gedragsmatige reacties	(3b) Oriëntaties	(3c) Affectieve reacties

1c *Toegekende persoonskenmerken*: beschrijvingen van anderen in termen van (gepercipieerde) persoonskenmerken of karaktereigenschappen. In ons onderzoek verwijst dit bijvoorbeeld naar het beschrijven van docentkenmerken bij het beargumenteren van een holistisch oordeel.

2a *Observaties*: dit betreft een combinatie van doelgerichte en zelfgerichte representaties bij het in kaart brengen van fysieke indrukken van anderen. Deze categorie veronderstelt dat een beoordelaar mentaal en/of fysiek interacteert met de beoordeelde en de beoordelingssituatie, wat bij beoordelen doorgaans het geval is (Conway, 1990; Tulving, 1972, 1983).

2c *Evaluaties*: ingenomen standpunten (meer of minder negatief) ten aanzien van anderen. Evaluaties blijken uit opmerkingen van beoordelaars als: “Ik denk dat ze betrokken is bij haar leerlingen, wat een goede zaak is”.

3a *Gedragsmatige reacties*: fysieke handelingen van de beoordelaar gericht op de beoordeelde persoon. Deze categorie is niet relevant in ons onderzoek.

3b *Oriëntaties*: neigingen of predisposities van beoordelaars om op een bepaalde manier op de beoordeelden te reageren. Een voorbeeld is vermijdingsgedrag van beoordelaars. Deze categorie wordt in ons onderzoek niet onderzocht.

3c *Affectieve reacties*: affecties verbonden aan fysiologische structuren van de beoordelaars, (mogelijk geuit via bijvoorbeeld huilen of lachen). Deze categorie is niet relevant in ons onderzoek.

Voor de validiteit van beoordelingen is het belangrijk dat beoordelaars bij het beoordelen concrete en abstracte representaties afwisselen. Het gebruik van concrete representaties, bijvoorbeeld de indrukken van de beoordelaars van de onderzoeksopdrachten en de video-opnamen in ons onderzoek, draagt bij aan de validiteit van de beoordelingen. Deze bevorderen namelijk de aansluiting tussen de beoordeling en de portfolio’s. Ook omwille van de aanvaardbaarheid van de beoordeling moeten beoordelaars duidelijk maken op welke concrete data in een portfolio ze hun beoordeling baseren. Verder zijn voor het geven van feedback concrete voorbeelden nodig die de (abstracte) beoordelingen illustreren. Anderzijds zijn voor accurate beoordelingen op de beoordelingscriteria abstracte representaties nodig. Abstracte representaties zijn ook nodig om voldoende nauwkeurig te kunnen voorspellen in welke andere situaties dan getoond in het portfolio, kan worden verwacht dat de docent op een bepaalde manier handelt. Verder moeten de beoordelingen het mogelijk maken meer en minder competente docenten van elkaar te onderscheiden (specificiteit) en metingen van eenzelfde criterium gebaseerd op verschillend portfoliomateriaal te vergelijken (convergentie).

Daarnaast is het aannemelijk dat doelgerichte representaties de betrouwbaarheid van beoordelingen verbeteren. Doelgerichte representaties vergroten immers de kans dat beoordelingen van docentgedrag worden gebaseerd op de competenties van die docenten, en minder op persoonsgebonden opvattingen van de beoordelaars.

4 Methode

4.1 Selectie van beoordelaars

Beoordelaars die zelf een onderwijsachtergrond hebben, hebben vaak minder moeite

met het beoordelen van docentcompetenties (Pula & Huot, 1993). Beoordelaars die de te beoordelen docenten reeds kennen, kunnen echter bevooroordeeld zijn. We kozen daarom voor externe beoordelaars met een onderwijsachtergrond. Geen van hen had ooit eerder docentportfolio's beoordeeld. De beoordelaars namen eerst deel aan de eerdere studies waarin beoordelingscriteria, standaarden en procedures zijn ontwikkeld (Van der Schaaf et al., 2003, 2004). In één daarvan beoordeelden de beoordelaars de geformuleerde criteria en standaarden als volledig en helder (gemiddelden boven 3.3. op een vierpuntsschaal) en als voldoende herkenbaar in de praktijk (gemiddelden van 3.0 tot 3.4), waarbij er ten aanzien van alle beoordelingscriteria sprake was van een hoge mate van consensus. De beoordelaars die deelnamen aan de hier gepresenteerde studie, waren een schoolleider (tevens ervaren docent); twee ervaren aardrijkskundeleraren (tevens nascholers); twee ervaren geschiedenisdocenten; een ervaren aardrijkskundedocent (tevens docent geschiedenis); een ervaren docent in de gammavakken (tevens lerarenopleider economie). Allen kregen een financiële vergoeding.

4.2 De samenstelling van de portfolio's

In totaal stelden 21 docenten (aardrijkskunde, economie, geschiedenis), werkzaam op evenzoveel scholen, vrijwillig een portfolio samen over hoe ze onderzoeksvaardigheden van leerlingen instrueren, begeleiden en beoordelen. De docenten verzamelden elk in een paar maanden tijd het volgende materiaal (zie ook Tabel 1):

1 een zelfbeschrijving van de ervaring van de docent en zijn of haar visie op het ontwikkelen van onderzoeksvaardigheden;
2 een serie onderzoeksopdrachten die de docent de leerlingen in opeenvolgende leerjaren in het Studiehuis geeft;
3 de resultaten van twee interviews over de praktijkkennis van de docent, en zijn of haar intenties bij het instrueren en coachen van onderzoeksvaardigheden van leerlingen;
4 twee video-opnames van lessen waarin de docent leerlingen instrueert en coacht bij het doen van onderzoek;
5 een onderzoeksopdracht die centraal staat in het portfolio, inclusief de leerdoelen van de opdracht en de motieven van de docent voor de inhoud en vorm van de opdracht;
6 beoordelingen van het werk van leerlingen (inclusief doelen, criteria en scoringsregels);
7 reflecties op eigen zwakke en sterke kanten, en hoe het onderwijs kan worden verbeterd;
8 beoordelingen van de docent door de leerlingen op een vragenlijst;
9 ter illustratie (niet ter beoordeling) voorbeelden van leerlingenwerk.

Het specificeren van de inhoud van het portfolio maakte het voor de docenten in ons onderzoek die geen ervaring hadden met het werken aan een portfolio, gemakkelijker om een portfolio samen te stellen, en ondersteunde ook de betrouwbaarheid van de beoordeling. Drie portfolio's werden beoordeeld in een voorstudie. Daarna werden de overige 18 portfolio's beoordeeld in de hoofdstudie. Alle docenten kregen een uitgebreid feedbackrapport (ongeveer 10 pagina's) met daarin alle beoordelingen, en feedback waarin de beoordelingen werden onderbouwd en suggesties voor verbetering werden gegeven.

4.3 Beoordelaarstraining en voorstudie

Om te bevorderen dat de beoordelaars accuraat zouden beoordelen, kregen ze een instructie en training. De beoordelaars bestudeerden eerst een handleiding met een overzicht van de doelen, planning en procedures van het onderzoek en een volledige omschrijving van de beoordelingscriteria, de performancestandaarden en bijbehorende ankerpunten, en het portfoliomateriaal. De training verliep vervolgens in een aantal stappen:

1 De beoordelaars bestudeerden eerst individueel een voorbeeldportfolio.
2 Vervolgens namen ze deel aan een trainingssessie (vier uur plenair). De beoordelaars werden geïnstrueerd in de te volgen beoordelingsprocedure en de bijbehorende beoordelingsformulieren. Elke beoordelaar oefende individueel met het uitvoeren van analytische en holisti-

sche beoordelingen. In de training lag het accent op het kunnen geven van argumenten voor een beoordeling.

3 Na de training beoordeelden de beoordelaars individueel drie portfolio's (de voorstudie). Daarna verbaliseerde elke beoordelaar in een retrospectieve hardopdenksessie met de onderzoeker (de eerste auteur van dit artikel) de gedachten die hij tijdens de beoordeling had.

4 De beoordelaars ontvingen feedback over hun beoordelingen, met onder meer informatie over de betrouwbaarheid van de beoordelingen en suggesties om hun beoordelingen te verbeteren (betreffende het accuraat gebruik van de criteria en de beoordelingsprocedure, de interpretatie van en het onderscheid tussen de criteria, de beargumentering van de gegeven beoordelingen, het relateren van de beoordeling aan concrete voorbeelden uit het portfolio, en het geven van verschillende typen feedback (positief, neutraal en negatief).

4.4 Instrumentatie en gegevensverzameling

Om zicht te krijgen op de cognitieve representaties van de beoordelaars gebruikten we drie bronnen van gegevens.

Beoordelingsformulieren. We ontwikkelden beoordelingsformulieren gebaseerd op het model van de Correspondent Inference Theory (Jones et al., 1965). Ten eerste illustreerden de beoordelaars elk beoordelingscriterium met data uit de portfolio's. Ten tweede beschreven ze hun interpretaties. Ten slotte gaven ze een score op elk van de criteria op een vijfpuntsschaal met ankerpunten. Per portfolio voegden de beoordelaars een 'overall' holistische beoordeling toe op een vijfpuntsschaal. Ten slotte beschreven de beoordelaars voor elk portfolio in hoeverre ze bij het beoordelen de voorgeschreven beoordelingsprocedure hadden gevolgd.

Hardopdenkprotocollen. De beoordelaars verwoordden in een hardopdenksessie (van gemiddeld twee uur) hun gedachten tijdens het beoordelen. Hardopdenksessies kunnen plaatsvinden tijdens de uitvoering van een taak, of daarna, retrospectief. De eerste manier is vooral gericht op cognitieve processen in het kortetermijngeheugen, terwijl bij retrospectief hardop denken een beroep wordt gedaan op het langetermijngeheugen (Ericsson & Simon, 1984). Wij gebruikten retrospectief hardop denken omdat het beoordelen van portfolio's op zichzelf al complex en tijdrovend was (gemiddeld vier uur per beoordeling). Het daarbij tegelijkertijd ook nog uitvoeren van hardop denken zou te belastend zijn voor de beoordelaars, en mogelijk afbreuk doen aan de kwaliteit van de beoordelingen. De hardopdenksessies vonden steeds plaats binnen twee weken na het beoordelen van de portfolio's. Daarbij gebruikten de beoordelaars hun eerder ingevulde beoordelingsformulieren als houvast.

De twee voornaamste vormen van protocolinvaliditeit zijn reactiviteit en onjuistheid bij het verslaan van gedachten (Russo, Johnson, & Stephens, 1989). In beide gevallen treedt verschil op tussen de achteraf weergegeven gedachten en de oorspronkelijke cognities. Bij reactiviteit wordt dat voornamelijk veroorzaakt door het uitvoeren van het hardop denken of door het tijdsverloop tussen de beoordeling en het hardop denken, waarin beoordelaars zich een andere voorstelling vormen van hun oorspronkelijke gedachten. Bij onjuiste weergave is er doorgaans sprake van enerzijds onvolledig rapporteren, anderzijds rapporteren van gedachten die zich niet voordeden. Met name het laatste moet zoveel mogelijk worden vermeden, omdat ten onrechte gerapporteerde gedachten niet meer als zodanig in het protocol kunnen worden onderkend.

Verschillende onderzoekers waarschuwen voor het gebruik van retrospectief hardop denken, Met name 'stimulus-cued' methoden, waarbij beoordelaars tijdens het hardop denken houvast hebben aan bijvoorbeeld eerder ingevulde beoordelingsformulieren, zouden leiden tot protocolinvaliditeit (Ericsson & Simon, 1980; Ericsson et al., 1984; Russo et al., 1989). Wij gaan er echter van uit dat beoordelaars tijdens het hardop denken nooit puur rapporteren over hun oorspronkelijke gedachten. In plaats daarvan construeren ze informatie waarin ze cognities over het eigen beoordelingsproces selecteren, interpreteren en communiceren (Long & Bourgh, 1996).

Samengevat denken we dat hardopdenkprotocollen bruikbare en unieke informatie

bevatten over cognitieve activiteiten van beoordelaars, maar dat ze de oorspronkelijke gedachten tijdens het beoordelen niet louter en volledig weerspiegelen. Om die reden achten we hardopdenkprotocollen vooral waardevol wanneer ze worden gebruikt in combinatie met andere bronnen (vgl. Long et al., 1996). In ons onderzoek zijn dat de ingevulde beoordelingsformulieren en de open interviews.

Open interviews. Direct na de hardopdenksessies namen we bij de beoordelaars open interviews af over de wijze waarop ze de taak hadden aangepakt. De interviews waren gericht op de gevolgde procedure en op de (cognitieve) activiteiten van de beoordelaars om tot een 'overall' beeld van een docent te komen (resulterend in een holistische beoordeling). Ten slotte zijn we nagegaan in hoeverre beoordelaars tijdens het beoordelen doelgericht respectievelijk zelfgericht waren.

4.5 De betrouwbaarheid van de beoordelingen

In de voorstudie beoordeelden de beoordelaars individueel portfolio's van docenten economie, geschiedenis en aardrijkskunde. In de beoordelingsformulieren gaven alle beoordelaars aan dat ze daarbij de voorgeschreven beoordelingsprocedure volgden. De jury-α's van de analytische beoordelingsscores waren 0.75, 0.44 en 0.63. In 50% van alle holistische beoordelingen stemden de beoordelaars volledig met elkaar overeen, bij 33% van de beoordelingen was er een verschil van een half punt, en in 17% was er een verschil van een punt of meer. Deze resultaten zijn vergelijkbaar met die van LeMahieu, Gitomer en Eresh (1995), die rapporteren over volledige overeenstemming tussen docentbeoordelingen van leerlingportfolio's in 46% tot 57% van de gevallen.

De acht criteria vormden een betrouwbare schaal (Cronbachs α = .79), wat een voorwaarde is voor het berekenen van ongewogen en gewogen gemiddelde analytische scores. Variantieanalyse wees niet op significante verschillen tussen de beoordelaars (zie Tabel 3).

4.6 Codering van hardopdenkprotocollen en beoordelingsformulieren

In de voorstudie verbaliseerden alle beoordelaars hun beoordelingen van het geschiedenisportfolio. De beoordelaars kregen vooraf diverse op Ericsson en Simon (1984) gebaseerde instructies om tijdens de sessie te blijven verwoorden wat ze eerder tijdens de beoordelingen dachten. De sessies vonden plaats bij de beoordelaar thuis of op zijn school, en werden opgenomen op band en volledig uitgetypt.

We gebruikten de hardopdenkprotocollen en de ingevulde beoordelingsformulieren van de voorstudie om een coderingsschema te ontwikkelen. Later in het onderzoek analyseerden we daarmee de cognitieve representaties van de beoordelaars.

Ondanks het gefaseerde ontwerp van de beoordelingsformulieren (conform de Correspondent Inference Theory), bleken de beoordelaars tijdens het hardop denken regelmatig van de hak op de tak te springen: ze gingen bijvoorbeeld vaak moeiteloos van het ene naar het andere onderwerp over, om vervolgens onaangekondigd weer terug te keren bij het eerste onderwerp. Daardoor was het niet mogelijk om de protocoldata in te delen naar betekenisvolle episodes. We hebben er daarom voor gekozen fragmenten te onderscheiden per portfolio-onderdeel per criterium (dus per cel in Tabel 1). De interbeoordelaarsovereenstemming (Cohens κ) van de

Tabel 3

Enkelvoudige variantieanalyse tussen docenten en tussen beoordelaars op holistische beoordelingen (h) en gewogen gemiddeld analytische beoordelingen (wa) in de voorstudie

Effecten		*df*	Sum of squares	Mean square	*F* (*p*)
Docenten	h	2	2.028	1.014	4.244 (.03)
	wa	2	4.819	2.409	19.146 (.00)
Beoordelaars	h	5	.944	.189	.486 (.78)
	wa	5	.432	.086	.165 (.97)

Tabel 4

Cohens κ's in de hardopdenkprotocollen en beoordelingsformulieren

Categorie	Hardopdenk-protocollen	Beoordelings-formulieren
Dimensie concreet-abstract (codering per regel)	.71	.65
Dimensie positief-negatief (codering per regel)	.66	.79
De situatie waarin docenten handelen	.75	1.00
Het eigen beoordelingsproces van beoordelaars	.69	.48
De beoordelingsprocedure	.65	.80

fragmenten tussen twee beoordelaars (de eerste auteur en een onafhankelijke onderzoeksassistent) van drie 'at random' gekozen protocollen was 0.89. Dit resultaat komt overeen met de range van de minimaal vereiste overeenstemming van 0.80 tot 0.90 zoals bediscussieerd in Ericsson e.a. (1984).

Het coderen van de fragmenten was primair gericht op de AST (Carlston, 1992, 1994). We beschreven de verbalisaties van de beoordelaars op de dimensie *concreet versus abstract*. Daarbij kregen concrete representaties (in ons onderzoek visuele manifestaties en observaties) code 1, meer abstracte representaties (categorisaties) code 2, en abstracte representaties (toegekende persoonskenmerken en evaluaties) code 3. Ook codeerden we de verbalisaties op de dimensie *positieve versus negatieve opmerkingen* (Huot, 1993): positief = code 1, neutraal = code 2, negatief = code 3. In navolging van Jones en Davis (1965) codeerden we ten slotte de verbalisaties over de situaties waarin de docenten handelden.

Als check op het aanvankelijke coderingsschema codeerden de twee eerder genoemde codeurs drie at random gekozen hardopdenkprotocollen. Discussies tussen de codeurs resulteerden in een verdere uitwerking van het coderingsschema. Deze bestond onder meer uit het toevoegen van de categorieën *het eigen beoordelingsproces van de beoordelaar*, en *de beoordelingsprocedure*. Om voldoende greep te krijgen op de dimensies *concreet-abstract* en *positief-negatief* besloten we de protocollen en beoordelingsformulieren op deze dimensies per uitgeschreven regel te coderen. We maakten beschrijvingen per categorie om een betrouwbare codering te verkrijgen. Samengevat codeerden we dus de beide beschreven dimensies per regel en de overige drie categorieën (*de situatie waarin docenten handelen*, *het eigen beoordelingsproces van de beoordelaar*, en *de beoordelingsprocedure*) per fragment.

Het uiteindelijke coderingsschema bevatte vijf hoofdcategorieën met elk diverse subcategorieën (zie Appendix). De hoofdcategorieën konden voldoende worden onderscheiden in zowel de hardopdenkprotocollen als de beoordelingsformulieren. Op één uitzondering na lagen alle Cohens κ's boven de 0.65 (zie Tabel 4), wat een voldoende indicatie is voor interbeoordelaarsovereenstemming (Popping, 1983).

5 Hoofdstudie

Na de voorstudie verkregen we 18 portfolio's (9 geschiedenis, 6 economie, 3 aardrijkskunde). Elk portfolio werd beoordeeld door twee beoordelaars, onafhankelijk van elkaar. De beoordelingsparen werden samengesteld op basis van hun vakexpertise. Aangezien de meeste beoordelaars expertise hadden in meer vakken, varieerde de samenstelling van de beoordelingsparen mede al naar gelang hun in de tijd wisselende beschikbaarheid. Gedurende een periode van 9 maanden beoordeelden de beoordelaars de 18 portfolio's in de volgorde waarin ze beschikbaar kwamen. Beoordelaar 1 beoordeelde zes portfolio's (voornamelijk economie), beoordelaar 2 vier (geschiedenis), beoordelaar 3 negen (voornamelijk geschiedenis), beoordelaar 4 vijf (aardrijkskunde en geschiedenis), beoordelaar 5 zeven (aardrijkskunde, geschiedenis en economie), en beoordelaar 6 vijf (voornamelijk economie) portfolio's.

Vervolgens verwoordden alle beoordelaars hardop denkend hun gedachten bij het beoordelen van twee at random gekozen portfolio's in een vak van hun eigen expertise. De protocollen werden volledig uitgetypt, resulterend in 7216 regels en 310 fragmenten. Zoals eerder aangegeven, zijn de fragmenten gesplitst per criterium per portfolio-onderdeel (de cellen in Tabel 1). De onafhankelijke onderzoeksassistent (degene die eerder interbeoordelaarsovereenstemming bereikte met de onderzoeker) codeerde de hardopdenkprotocollen van in totaal 12 portfolio's (twee portfolio's per beoordelaar) en de 36 beoordelingsformulieren van alle 18 portfolio's. Daarbij gebruikte ze het eerder ontwikkelde coderingsschema.

6 Data-analyse

6.1 Betrouwbaarheid van de beoordelingen

De beoordelingen leidden tot drie typen scores: een ongewogen gemiddelde analytische, een gewogen gemiddelde analytische, en een holistische score. De ongewogen gemiddelde analytische scores betroffen het gemiddelde op de acht beoordelingscriteria. De gewogen gemiddelde analytische scores waren gebaseerd op de bovengenoemde, in een eerdere studie verkregen gewichten per criterium. De holistische scores waren afzonderlijke beoordelingen per portfolio van de overall competentie van de docent met betrekking tot het instrueren, begeleiden en beoordelen van onderzoeksvaardigheden van leerlingen.

Ten eerste analyseerden we de schaalbaarheid van de acht beoordelingscriteria door het berekenen van Cronbachs α.

Ten tweede onderzochten we de interbeoordelaarsbetrouwbaarheid voor de afzonderlijke analytische scores door het berekenen van jury-α's. Met het gebruiken van correlaties als indicatoren voor betrouwbaarheid stappen we af van het idee dat betrouwbare beoordelingen per se eenzelfde gemiddelde score zouden moeten hebben (Murphy & De Shon, 2000). Verder is het goed mogelijk dat beoordelaars portfolio's op verschillende manieren interpreteren. Aangezien subgroepen van beoordelaars vergelijkbare interpretaties kunnen hebben, zijn verschillende interpretaties binnen een groep beoordelaars niet noodzakelijk idiosyncratisch. Dit laat de mogelijkheid open dat aan beoordelaars gekoppelde variantie-effecten niet automatisch beoordelingsfouten zijn. Hiervan kan echter pas sprake zijn als de beoordelingen accuraat zijn uitgevoerd. Een voorwaarde daarvoor is dat de beoordelaars hun beoordelingen baseren op de beoordelingscriteria. We zijn dit nagegaan door de inhoud van de beoordelingen te vergelijken met de uitgeschreven versie van de gebruikte beoordelingscriteria.

Ten derde analyseerden we de mate van overeenstemming tussen de holistische beoordelingen.

Als aanvullende indicatie van de consistentie van de beoordelingen zijn we ten vierde nagegaan of de portfolio's verschillen in hun gemiddelden op de drie typen beoordelingsscores (ongewogen gemiddeld analytisch, gewogen gemiddeld analytisch, en holistisch).

Ten slotte voerden we op drie typen scores een variantieanalyse uit om een schatting te maken van de variantie die is gerelateerd aan de beoordelaars.

6.2 Cognitieve representaties van de beoordelaars

We analyseerden de data van de interviews kwalitatief. We berekenden de frequenties en gemiddelden van coderingen van de hardopdenkprotocollen en de beoordelingsformulieren. We gebruikten variantieanalyse om de verschillen tussen de beoordelaars in kaart te brengen.

6.3 Relaties tussen beoordelingen en cognitieve representaties

De veronderstelling dat beoordelingen worden beïnvloed door cognitieve representaties van beoordelaars impliceert dat beoordelaars die verschillen in hun representatie, ook kunnen verschillen in hun beoordelingen. We gebruikten multipele regressieanalyse om na te gaan hoe goed de scores op de onderscheiden categorieën in de hardopdenkprotocollen en de beoordelingsformulieren de drie typen beoordelingen kunnen verklaren. Deze analyse is niet uitgevoerd voor de categorie *het eigen*

beoordelingsproces van de beoordelaar, omdat deze in de voorstudie onvoldoende kon worden onderscheiden in de beoordelingsformulieren (Tabel 4). De geldigheidsvoorwaarden voor multipele lineaire regressieanalyses, zoals normale verdeling, constante variantie, lineariteit en geen grote multicollineariteit, zijn door ons op de daarvoor gebruikelijke manieren gecontroleerd. Aan deze voorwaarden bleek in voldoende mate te worden voldaan.

7 Resultaten van de hoofdstudie

7.1 De betrouwbaarheid van de beoordelingen

De acht beoordelingscriteria vormden een betrouwbare schaal: de Cronbachs α was 0.76. De jury-α's voor de analytische beoordelingen waren voldoende voor 12 beoordelaarsparen, variërend van 0.39 tot 0.76. De jury-α's waren laag of zelfs negatief voor zes paren, -0.80 bij één paar, en variërend van -0.11 tot 0.22 bij vijf paren. We vergeleken de inhoud van de beoordelingsformulieren met de uitgeschreven beoordelingscriteria. Beide kwamen voldoende overeen.

In 35% van de paarsgewijze beoordelingen kwamen de holistische scores volledig overeen. Een verschil van een half punt (op een vijfpuntsschaal) kwam voor bij 12% van de beoordelingen, een verschil van een punt kwam voor bij 47% van de beoordelingen, en bij één portfolio (6%) was er een verschil van anderhalf punt.

De overeenstemming tussen de drie typen beoordelingsscores is nagegaan door het berekenen van Pearson-correlaties en gepaarde *t*-toetsen. De scores correleren sterk (zie Tabel 5). Verder is het gemiddelde van de holistische beoordelingen (3.12; *SD* = .78) significant lager dan het gemiddelde van de gewogen gemiddelde analytische beoordelingen (3.60; *SD* = .96). Het gemiddelde van de holistische beoordelingen correspondeert sterk met het gemiddelde van de (ongewogen) gemiddelde analytische scores (3.11; *SD* = .59).

Uit de variantieanalyse blijkt dat er amper sprake is van een beoordelaarseffect (zie Tabel 6).

7.2 Cognitieve representaties van beoordelaars

Open interviews

In de open interviews vertelden de beoordelaars dat ze de beoordelingsprocedure bruikbaar vonden en dat ze deze bij het beoordelen van de portfolio's getrouw volgden. Het beoordelen kostte gemiddeld vier uur per portfolio.

Bij het beoordelen richtten vijf beoordelaars zich voornamelijk op het vormen van een coherent beeld van de betreffende docent op basis van zijn of haar portfolio. Vervolgens onderbouwden ze dat beeld met concrete voorbeelden uit het portfolio. Met name de video-opnames waren ondersteunend bij dat proces. De beoordelaars gingen (uit zichzelf) op twee manieren te werk. Vier beoordelaars bekeken eerst de video-opnames en maakten daarbij aantekeningen over relevant docentgedrag. Vervolgens bestudeerden ze de rest van het portfolio. Ten slotte vulden ze de beoordelingsformulieren in. Daarbij bekeken ze nogmaals het gehele portfolio, maar nu per criterium in plaats van per portfolio-onderdeel. In een alternatief proces begonnen twee beoordelaars direct met het invullen van de beoordelingsformulieren. Zij bestudeerden de portfolio's en de video's slechts één keer. Er bleken geen significante verschillen tus-

Tabel 5

Pearson-correlaties en 'paired samples' t-toetsen voor holistische (h), gewogen gemiddeld analytische (wa), en ongewogen gemiddeld analytische (ma) scores (18 portfolio's)

	Eerste gemiddelde	Tweede gemiddelde	*r* (*p*)	Verschil	*SD*	*df*	*t* (*p*)
H – wa	3.12 (h)	3.60 (wa)	.85 (.00)	-.44	.50	35	-5.28 (.00)
H – ma	3.12 (h)	3.11 (ma)	.85 (.00)	.04	.42	35	.59 (.56)
Wa – ma	3.60 (wa)	3.11 (ma)	.99 (.00)	.49	.38	35	7.59 (.00)

Tabel 6

Enkelvoudige variantieanalyse tussen docenten en tussen beoordelaars, op holistische (h), gewogen gemiddeld analytische (wa), en ongewogen gemiddeld analytische (ma) scores (18 portfolio's)

Effecten	*df*	Sum of squares	Mean square	*F (p)*
			Docenten	
h	17	12.90	.76	1.54 (.17)
wa	17	19.64	1.16	1.67 (.14)
ma	17	7.40	.44	1.62 (.16)
			Beoordelaars	
h	5	.97	.19	.29 (.92)
wa	5	2.04	.41	.41 (.84)
ma	5	1.07	.21	.58 (.72)

sen de uiteindelijk beoordelingen van de beide groepen docenten uit de *t*-toetsen.

De beoordelaars merkten dat eerder beoordeeld portfoliomateriaal en eerder beoordeelde criteria van invloed waren op hun beoordeling van later te beoordelen portfolio's en latere criteria. Beoordelaar 1 was zich bewust van dit proces: "Ik zou me bij de beoordeling van criterium 3 (ORGA) niet moeten baseren op de zelfbeschrijving, maar die informatie blijft toch in mijn achterhoofd zitten en beïnvloedt mijn manier van beoordelen, ook al zou dat eigenlijk niet moeten." De meeste beoordelaars gaven aan hierop te controleren, ze corrigeerden hun beoordelingen naderhand. Beoordelaar 4: "Bij het beoordelen heb ik de neiging om mijn intuïtie te volgen. Omdat ik weet dat dat niet goed is, zet ik mezelf ertoe de getrainde procedure zo precies mogelijk te volgen." Beoordelaar 5 ging in op het gevaar om vooral op die fragmenten in het portfolio te letten die de eigen preconcepties bevestigen. "Ik corrigeer mezelf hier telkens op door tegen mezelf te zeggen: 'Let op, nu doe je het weer'." Beoordelaar 3: "Een paar dagen na het beoordelen van een portfolio lees ik opnieuw mijn beoordeling om te checken of mijn beoordeling wel recht doet aan het portfolio."

De opgebouwde beelden werden ook beïnvloed door de eigen ervaringen van de beoordelaars. Vier beoordelaars bevestigden dat zij hun beoordelingen deels baseerden op hun eigen kennis over docenten. Beoordelaar 3: "Je weet wat je kunt verwachten van docenten. Je eigen ervaringen kleuren je beoordelingen."

Hardopdenkprotocollen en beoordelingsformulieren

In de hardopdenkprotocollen maakten de beoordelaars gemiddeld meer opmerkingen dan in de beoordelingsformulieren. Over het algemeen waren de opmerkingen in de hardopdenkprotocollen minder concreet dan de opmerkingen in de beoordelingsformulieren. Verder verschilden de beoordelaars in het aantal opmerkingen dat ze maakten (zie Tabel 7).

De beoordelaars verschilden significant van elkaar in hun representaties op de dimensies *concreet-abstract* en *positief-negatief* in zowel de hardopdenkprotocollen als de beoordelingsformulieren. De resultaten op de dimensie *concreet-abstract* zijn: $F = 0.06$, $df = 5$, $p < 0.001$ (hardopdenkprotocollen); $F = 4.19$, $df = 5$, $p < 0.001$ (beoordelingsformulieren). Op de dimensie *positief-negatief* zijn de resultaten $F = 0.04$, $df = 5$, $p < 0.001$ (hardopdenk protocollen); $F = 6.24$, $df = 5$, $p < 0.001$ (beoordelingsformulieren). De verschillen bij de andere onderscheiden categorieën waren niet significant.

7.3 Relaties tussen beoordelingen en cognitieve representaties

In hoeverre kunnen de beoordelingen van beoordelaars worden verklaard vanuit hun cognitieve representaties? We gebruikten multipele regressieanalyse voor het beantwoorden

Tabel 7

Gemiddelden en aantallen opmerkingen per beoordelaar in de hardopdenkprotocollen en beoordelingsformulieren

		Concreet-abstract			Positief-negatief			Situatie docent	Eigen beoordeling	Beoordelingsprocedure
Bronnen per beoordelaar	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*N*	*N*	*N*
Beoordelaar 1										
Hardopdenkprotocol	2	1.74	.70	530	2.69	.67	62	42	14	108
Beoordelingsformulier	6	1.54	.57	357	1.98	.98	56	10	49	7
Beoordelaar 2										
Hardopdenkprotocol	2	1.96	.60	138	2.31	.75	20	1	0	13
Beoordelingsformulier	4	1.51	.54	219	2.13	.95	24	6	30	4
Beoordelaar 3										
Hardopdenkprotocol	2	1.89	.66	1079	2.34	.84	128	50	29	104
Beoordelingsformulier	9	1.44	.53	1456	1.77	.94	176	17	72	19
Beoordelaar 4										
Hardopdenkprotocol	2	1.91	.68	554	1.86	.89	94	26	33	80
Beoordelingsformulier	5	1.55	.56	302	2.32	.88	78	6	11	7
Beoordelaar 5										
Hardopdenkprotocol	2	1.95	.60	594	1.97	.90	68	37	27	87
Beoordelingsformulier	7	1.45	.51	504	1.74	.97	76	14	26	2
Beoordelaar 6										
Hardopdenkprotocol	2	1.94	.64	613	2.47	.80	83	31	23	79
Beoordelingsformulier	5	1.47	.53	852	2.16	.91	156	9	22	1

van deze vraag. Daarbij namen we de drie typen beoordelingsscores als criteriumvariabelen en de scores op de categorieën van cognitieve representaties als predictoren. Voor de hardopdenkprotocollen leverde dat geen significante resultaten op. In de beoordelingsformulieren bleken echter de cognitieve representaties duidelijk gerelateerd aan de beoordelingen (n = 36, methode 'enter'). De hoeveelheid verklaarde variantie is bij de drie typen beoordelingen vergelijkbaar. Holistische beoordelingen: R^2 = .59, 'adjusted' R^2 = 0.34, F = 2.36, df = 13, p = 0.04. Gewogen gemiddelde analytische beoordelingen: R^2 = 0.63, adjusted R^2 = 0.39, F = 2.68, df = 13, p = 0.03. Ongewogen gemiddelde analytische beoordelingen: R^2 = 0.65, adjusted R^2 = 0.40, F = 2.59, df = 14, p = 0.03.

We volstaan hier met het rapporteren over het model met het hoogste percentage verklaarde variantie (65%) in de beoordelingen (dus de ongewogen gemiddelde analytische beoordelingen). Er is nauwelijks sprake van multicollineariteit onder de predictoren: de tolerantie ligt tussen 0.58 en 0.88 en de Pearson-correlaties liggen tussen -0.22 en +0.31 (voornamelijk rond r = .00). De β's tonen aan dat de gecategoriseerde representaties op de dimensie *concreet-abstract* (β = .35; t = 2.13; p =.05) en op de dimensie *positief-negatief* (β = .71, t = 4.4; p = .00) significant bijdragen aan het verklaren van de gegeven beoordelingen. De andere in deze analyse betrokken categorieën (*de situatie waarin de docent handelt* en *de beoordelingsprocedure*) dragen niet significant bij.

Aangezien de scores op de dimensies *concreet-abstract* en *positief-negatief* signifcant blijken bij te dragen aan het verklaren van de portfoliobeoordelingen, onderzochten we ook de samenhang ervan met de interbeoordelaarsbetrouwbaarheid. Voor elk be-

oordelaarspaar (n = 18) berekenden we eerst de verschilscores tussen de gemiddelden van de eerste en de tweede beoordelaar op beide dimensies. Vervolgens correleerden we de verschilscores per beoordelaarspaar met hun jury-α's. De resultaten tonen aan dat hoe meer de in de beoordelingsformulieren gemaakte opmerkingen van de beoordelaars met elkaar corresponderen op de dimensies *concreet-abstract* en *positief-negatief*, hoe hoger de interbeoordelaarsbetrouwbaarheid is van de portfoliobeoordelingen (r = .44, p = .09 op de dimensie *concreet-abstract*; r = .53, p = .04 op de dimensie *positief-negatief*).

8 Conclusie en discussie

Ons onderzoek was gericht op de volgende vragen: Wat is de betrouwbaarheid van portfoliobeoordelingen? Welke cognitieve representaties gebruiken beoordelaars bij het beoordelen? In hoeverre zijn de cognitieve representaties gerelateerd aan de gegeven beoordelingen en aan de betrouwbaarheid van de beoordelingen?

We gebruikten de Correspondent Inference Theory (Jones et al., 1965) en de Associated Systems Theory (AST) (Carlston, 1992, 1994) om cognitieve representaties van beoordelaars te onderzoeken. We gebruikten een mix aan kwantitatieve en kwalitatieve methoden om de betrouwbaarheid van beoordelingen in kaart te brengen, de cognitieve representaties van de beoordelaars te beschrijven, en vervolgens beide aan elkaar te relateren.

Zes getrainde beoordelaars beoordeelden paarsgewijs 18 docentportfolio's zowel analytisch als holistisch. De jury-α's van de analytische beoordelingen varieerden van 0.39 tot 0.76 voor 12 van de 18 paren. Variantieanalyse toonde aan dat er nauwelijks sprake was van beoordelaarseffecten. In hun holistische beoordelingen, die sterk waren gerelateerd aan de ongewogen gemiddelde analytische beoordelingen, verschilden de beoordelaars bij slechts één portfolio meer dan 1 punt (op de vijfpuntsschaal).

De beoordelaars gebruikten cognitieve representaties die liggen op de dimensies *concreet-abstract* (bijvoorbeeld visuele manifestaties en het toekennen van karaktereigenschappen) en *positief-negatief* (bijvoorbeeld negatief en positief commentaar), de situatie waarin docenten handelen, het eigen beoordelingsproces en de beoordelingsprocedure.

De beoordelaars verschilden significant van elkaar in hun door ons gecategoriseerde cognitieve representaties op de dimensies *concreet-abstract* en *positief-negatief* in zowel de hardopdenkprotocollen als de beoordelingsformulieren. De (gemiddeld minder abstracte) representaties in de beoordelingsformulieren hingen significant samen met de gegeven beoordelingen. De verschillen in representaties tussen de beoordelaars hingen significant samen met de interbeoordelaarsbetrouwbaarheid.

Wat is nu de waarde van deze resultaten? Volgens verschillende onderzoekers zouden bij portfoliobeoordelingen andere eisen aan de betrouwbaarheid moeten worden gesteld dan bij meer gestandaardiseerde beoordelingsvormen. Bij deze laatste liggen interbeoordelaarsbetrouwbaarheidscoëfficiënten regelmatig boven 0.90 (Nunally, 1978). Gezien het meer open, complexe en contextgebonden karakter van portfolio's zouden ze bij portfoliobeoordelingen lager mogen zijn. Koretz, Klein, McCaffrey, & Stecher (1992) stellen bijvoorbeeld dat interbeoordelaarsbetrouwbaarheidscoëfficiënten van 0.80 voor performance assessments tamelijk hoog zijn. Gentile (1992) rapporteert dat voor portfoliobeoordelingen coëfficiënten boven 0.80 hoog zijn en boven 0.65 voldoende. In ons onderzoek voldoen we slechts gedeeltelijk aan deze criteria. Verder onderzoek zal moeten aantonen wat realistische eisen zijn in verband met de betrouwbaarheid van portfoliobeoordelingen.

Ten tweede kwamen de holistische beoordelingen van de portfolio's sterk overeen met de ongewogen gemiddelde analytische beoordelingen. Het is de vraag in hoeverre dit wenselijk is, aangezien het ene beoordelingscriterium mogelijk belangrijker is dan het andere en daarom meer gewicht zou moeten krijgen. In eerder onderzoek (Van der Schaaf et al., 2003) anticipeerden we hierop door te onderzoeken welke gewichten de beoordelingscriteria volgens de beoordelaars zouden moeten krijgen. Uit het feit dat de beoorde-

laars in de onderhavige studie waren getraind in het toepassen van deze weging en ze daar bij hun holistische beoordeling kennelijk toch van afweken, blijkt dat beoordelaars een sterke neiging hebben om hun beoordelingen te "middelen".

Ten derde blijkt uit de interviews dat beoordelaars ernaar streven een coherent beeld van de docent op te bouwen. Volgens de beoordelaars werd hun opgebouwde beeld beïnvloed door eerder beoordeeld portfoliomateriaal van de docent, eerder beoordeelde portfolio's van andere docenten, en eigen ervaringen. Deze resultaten zijn vergelijkbaar met die van andere studies in verschillende domeinen, zoals bijvoorbeeld het opbouwen door ervaren lezers van een coherent beeld van een tekst (vgl. Zwaan & Brown, 1996).

Ten vierde is een voorwaarde voor valide beoordelingen dat beoordelaars hun beoordelingen baseren op de portfolio-inhoud. Concrete representaties bevorderen dit. Aan de andere kant zijn meer abstracte representaties nodig om de beoordelingen te kunnen generaliseren naar andere situaties dan in de portfolio's getoond. Idealiter wisselen beoordelaars concrete en abstracte representaties af. Verder wordt de kwaliteit van de beoordelingen verbeterd door representaties primair betrekking te laten hebben op het portfolio van de docent (in plaats van op subjectieve percepties van de beoordelaar).

Ten vijfde gebruikten we multipele regressieanalyse om een inschatting te maken van de invloed van de cognitieve representaties. De gecategoriseerde representaties verklaarden 65% van de variantie in de portfoliobeoordelingen. Hieraan werd statistisch significant bijgedragen door de representaties op de dimensies *concreet-abstract* en *positief-negatief*. Verder bleek dat hoe groter de overeenstemming tussen de beoordelaars op deze twee dimensies, hoe hoger de interbeoordelaarsbetrouwbaarheid van de gegeven beoordelingen is. Het is dus aannemelijk dat deze twee dimensies van cognitieve representaties van invloed zijn op de betrouwbaarheid van portfoliobeoordelingen. Dit ondersteunt de waarde van de gebruikte modellen (Carlston, 1992, 1994; Jones et al., 1965) voor het in kaart brengen van beoordelingsactiviteiten van beoordelaars.

Gezien de persoons- en contextgebonden aard van portfolio's, is de beoordeling ervan moeilijk en complex, ook na intensieve training. Dat beïnvloedt mogelijk de betrouwbaarheid. Ook al toont ons onderzoek aan dat bepaalde cognitieve representaties bijdragen aan het betrouwbaar beoordelen van portfolio's, het laat ook zien dat beoordelaars op andere cognitieve representaties van elkaar kunnen verschillen, ongeacht hun overeenstemming in gegeven beoordelingen. Aangezien beoordelaars bij het beoordelen van portfolio's verschillende onderdelen als relevant zullen selecteren, deze verschillend interpreteren, en de gemaakte interpretaties verschillend extrapoleren naar de door hen geïnterpreteerde beoordelingscriteria, is het de vraag in hoeverre consensusmethoden voor het aantonen van de betrouwbaarheid van beoordelingen de voorkeur verdienen. Als alternatieve procedure stellen Delandshere en Petrosky (1994) voor, uit te gaan van confirmatie (vergelijkbaar met het krijgen van een 'second opinion' van een arts), in plaats van replicatie. Linn (1994) illustreert hoe comités bij het onderzoeken van kwalificaties van kandidaten tot geïntegreerde beslissingen komen. Hij beweert dat zo'n confirmerende benadering ook bruikbaar kan zijn voor portfoliobeoordelingen. Nader onderzoek zal moeten uitwijzen wat de bruikbaarheid en consequenties van confirmerende benaderingen zijn, bijvoorbeeld door quasi-experimentele studies waarin de beide benaderingen (consensus versus confirmatie) in verschillende settings worden toegepast en op bruikbaarheid en consequenties worden vergeleken.

De consensusbenadering berust op de betekenis van cijfermatige consistentie tussen beoordelaars. Voor summatieve beoordelingen, bijvoorbeeld gericht op certificering of functiewaardering, bevatten portfolio's doorgaans een verzameling van het beste werk van de beoordeelde en is een hoge interbeoordelaarsbetrouwbaarheid nodig om tot adequate en eerlijke beslissingen te kunnen komen. Voor formatieve doeleinden kunnen echter vraagtekens worden gezet bij de noodzaak en het nut van beoordelaars die cijfermatig identiek beoordelen. In dat geval lijkt het belangrijker dat de beoordelaars het

eens zijn over de consequenties van de gegeven beoordelingen en de te geven feedback dan dat ze het eens zijn over de te geven score.

Uit ons onderzoek blijkt ten slotte dat externe beoordelaars situationele informatie uit de portfolio's bij hun beoordeling betrekken (bijvoorbeeld over kenmerken van leerlingen en het schoolbeleid). Het valt te verwachten dat externe beoordelaars deze informatie deels kleuren vanuit hun eigen onderwijservaringen. Interne beoordelaars, die in dezelfde onderwijsorganisatie werken als de beoordeelde docenten, zouden daarom meer valide en consistent kunnen beoordelen. Ook wordt vaak gesuggereerd dat beoordelaars met vergelijkbare achtergronden sneller tot overeenstemming komen (Pula et al., 1993). Het lijkt daarom bij portfoliobeoordeling gepast om te werken met interne beoordelaars (Linn, 1994). Een nadeel van interne beoordelaars is echter de waarschijnlijk grotere invloed van persoonlijke interacties en indrukken, en de relatie met de te beoordelen docent. Een mogelijk gevolg hiervan is dat het werken met interne beoordelaars tot meer idiosyncratische beoordelingen leidt, en dat heeft weer een negatief effect op de validiteit en de betrouwbaarheid van de beoordelingen. Het valt dan ook te verwachten dat beoordelingen van interne beoordelaars meer "zelfgericht" zijn dan die van externe beoordelaars (Carlston, 1994). Nauwkeurige interne beoordeling vergt daarom beoordelaarstraining waarin voldoende aandacht wordt besteed aan "doelgericht" beoordelen. Onderzoek moet uitwijzen wat de mogelijkheden voor dergelijke trainingen zijn.

Noten

1 Dit onderzoek is gefinancierd door NWO/PROO, aanvraagnummer 490-23-081.

Literatuur

Andrews, T. E., & Barnes, S. (1990). Assessment of teaching. In W.R. Houston (Ed.), *Handbook of research on teacher education* (pp. 569-598). New York: Macmillan.

Bird, T. (1990). The schoolteacher's portfolio. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers* (pp. 241-256). Newbury Park, CA: Sage Publications.

Burns, C. W. (1999). Teaching portfolios and the evaluation of teaching in higher education: Confident claims, questionable research support. *Studies in Educational Evaluation, 25*, 131-142.

Carlston, D. (1992). Impression formation and the modular mind: The associated systems theory. In L. L. Martin & A. Tesser (Eds.), *The construction of social judgments* (pp. 301-341). Hillsdale, NJ: Erlbaum.

Carlston, D. (1994). Associated systems theory: A systematic approach to cognitive representations of persons. *Advances in Social Cognition, 7*, 1-78.

Conway, M. A. (1990). Associations between autobiographical memories and concepts. *Journal of Experimental Psychology: Learning, Memory and Cognition, 16*, 799-812.

Day, D. V., & Sulsky, L. M. (1995). Effects of frame-of-reference training and information configuration on memory organization and rating accuracy. *Journal of Applied Psychology, 80*, 158-167.

Delandshere, G., & Petrosky, A. (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.

DeNisi, A. S., Cafferty, T. P., & Meglino, B. M. (1984). A cognitive view of the performance appraisal process: A model and research propositions. *Organizational Behavior and Human Performance, 33*, 360-396.

Dunbar, S. B., Koretz, D., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*, 289-304.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*, 215-251.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Feldman, J. M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal.

Journal of Applied Psychology, 66, 127-148.

Fiske, S. T. (1992). Thinking is for doing: Portraits of social cognition from daguerreotype to laserphoto. *Journal of Personality and Social Psychology, 63*, 877-889.

Gentile, C. (1992). *Exploring new methods for collecting students' school-based writing: NAEP's 1990 Portfolio Study.* Washington, DC: Office of Educational Research and Improvement.

Gilbert, D. T. (1989). Thinking lightly about others: Automatic components of the social inference process. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 189-211). New York: Guilford.

Huot, B. A. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment. Theoretical and empirical foundations* (pp. 206-232). Cresskill, New Jersey: Hampton Press, Inc.

Jones, E. E., & Davis, K. E. (1965). From acts to dispositions: the attribution process in person perception. In L. Berkowitz (Ed.), *Advances in experimental social psychology 2* (pp. 219-266). New York: Academic Press.

Johnson, R. L., McDaniel, F., & Willeke, M. J. (2000). Using portfolios in program evaluation: an investigation of interrater reliability. *The American Journal of Evaluation, 21*, 65-80.

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

Kelly, G. A. (1955). *The psychology of personal constructs*. New York: Norton.

Koretz, D., Klein, S., McCaffrey, D., & Stecher, B. (1992). *The reliability of scores from the 1992 Vermont portfolio assessment program*. Washington, DC: RAND Institute on Education & Teaching.

Krull, D. S. (2001). On partitioning the fundamental attribution error: Dispositionalism and the correspondence bias. In G. B. Moskowitz (Ed.). *Cognitive social psychology. The Princeton symposium on the legacy and future of social cognition* (pp. 211-227). Mahwah NJ: Lawrence Erlbaum.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*, 72-107.

LeMahieu, P., Gitomer, D., & Eresh, J. (1995). Portfolios in large-scale assessment: difficult but not impossible. *Educational Measurement: Issues and Practice, 14*, 11-16, 25-28.

Lievens, F. (2001). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology*, *86*, 255-264.

Linn, R. L. (1994). Performance assessment. Policy promises and technical measurement standards. *Educational Researcher, 23*, 4-14.

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

Long, D. L., & Bourgh, T. (1996). Thinking aloud: Telling a story about a story. Commentary. *Discourse Processes, 21*, 329-339.

Lyons, N. (Ed.). (1998). *With portfolio in hand. Validating the new teacher professionalism.* New York: Teachers College Press.

Martindale, C. (1991). *Cognitive psychology: A neural-network approach*. Pacific Grove, CA: Brooks/Cole.

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

Murphy, K. R., & De Shon, R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology, 53*, 873-900.

Norman, D. A. (1985). Human information processing: the conventional view. In A. M. Aitkenhead & J. M. Slack (Eds.), *Issues in cognitive modelling* (pp. 309-336). Hillsdale, New York: Lawrence Erlbaum Associates.

Nunally, J. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill: New York.

Pula, J. J., & Huot, B. A. (1993). A model of background influences on holistic raters. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment. Theoretical and empirical foundations* (pp. 237-265). Cresshill, NJ: Hampton Press.

Popping, R. (1983). *Overeenstemmingsmaten voor nominale data*. Dissertatie, Rijksuniversiteit Groningen.

Reckase, M. D. (1995). Portfolio assessment: A theoretical estimate of score reliability. *Educational Measurement: Issues and Practice, 14*, 12-14, 31.

Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory & Cognition, 17*, 759-769.

Schaaf, M. F. van der, Stokking, K. M., & Verloop, N. (2003). Developing performance standards for teacher assessment by policy capturing. *Assessment & Evaluation in Higher Education, 28*, 395-410.

Schaaf, M. F. van der, Stokking, K. M., & Verloop, N. (ter publicatie aangeboden). *Developing teaching content standards using a delphi method.*

Schleicher, D. J., & Day, D. V. (1998). A cognitive evaluation of frame-of-reference rater training: Content and process issues. *Organizational Behavior and Human Decision Processes, 73*, 76-101.

Shapley, K. S., & Bush, M. J. (1999). Developing a valid and reliable portfolio assessment in the primary grades: Building on practical experience. *Applied Measurement in Education, 12*, 111-132.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*, 215-232.

Stokking, K., Schaaf, M. van der, Jaspers, J., & Erkens, G. (2004). Teachers' assessment of students' research skills. *British Journal of Educational Research*, *30*, 93-115.

Sulsky, L. M., & Day, D. V. (1992). Frame-of-reference training and cognitive categorization: An empirical investigation of rater memory issues. *Journal of Applied Psychology, 77*, 501-510.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization and memory* (pp. 381-403). New York: Academic Press.

Tulving, E. (1983). *Elements of episodic memory*. Oxford: Clarendon Press.

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: a quantitative review. *Journal of Occupational and Organizational Psychology, 67*, 189-205.

Zwaan, R. A., & Brown, C. M. (1996). The influence of language proficiency and comprehension skill on situation-model construction. *Discourse Processes, 21*, 289-327.

Manuscript aanvaard: 1 november 2004

Auteurs

Marieke van der Schaaf is als onderzoeker en docent onderwijskunde verbonden aan de Capaciteitsgroep Onderwijskunde van de Universiteit Utrecht.

Karel Stokking is als hoogleraar onderwijskunde verbonden aan de Faculteit Sociale Wetenschappen van de Universiteit Utrecht.

Nico Verloop is als hoogleraar onderwijskunde en directeur werkzaam bij het ICLON van de Universiteit Leiden.

Correspondentieadres: Marieke van der Schaaf, Capaciteitsgroep Onderwijskunde, Universiteit Utrecht, Postbus 80140, 3508 TC Utrecht, e-mail: m.f.vanderschaaf@fss.uu.nl

Abstract

The influence of raters' cognitive representations on the assessment of teacher portfolios

Nowadays, portfolios are frequently used to assess teachers' competences. In portfolio assessment, the issue of rater reliability is a vexing problem. Insight into the representations raters form during the assessment process is crucial to improving the quality of assessment. We used a mixed quantitative and qualitative approach to research cognitive processes underlying raters' reliability. Six raters systematically assessed 18 portfolios. The interrater reliability of 12 portfolios was reasonable to good. Variance analysis showed slight rater effects. We used the Associated Systems Theory (Carlston, 1992, 1994) and the Correspondent Inference Theory (Jones & Davis, 1965) to analyse raters' retrospective verbal protocols and judgment forms. Raters' cognitive representations on the *concrete-abstract remarques* and *positive-negative evaluation* dimensions were significantly related to the judgments given.

Appendix

Cognitieve representaties in hardopdenkprotocollen en beoordelingsformulieren

Dimensie concreet-abstract (gecodeerd per regel)
1 visuele manifestaties (bv. “Ze zegt …”);
2 categorisaties (bv. “Dat betekent …”);
3 toekennen van karaktereigenschap (bijvoorbeeld “Ze is een georganiseerd persoon.”).

Dimensie positief-negatief (gecodeerd per regel)
1 positief (alle vormen van positief commentaar);
2 neutraal (neutrale of gebalanceerde feedback, bijvoorbeeld “Enerzijds [gevolgd door negatief commentaar], anderzijds [gevolgd door positief commentaar]”);
3 negatief (alle vormen van negatief commentaar);
4 tips en suggesties (bijvoorbeeld “Hij zou beter … kunnen doen”).

De situatie waarin docenten handelen
1 zich vanuit zijn eigen ervaringen en achtergrond in de situatie van een docent verplaatsen*;
2 opmerkingen met betrekking tot het gebruikte materiaal en de methode.

Tijdens het beoordelen rekening houden met:
1 kenmerken van de leerlingen;
2 de hoeveelheid tijd die de docent tot zijn beschikking heeft;
3 ervaring van de docent in het aanleren van onderzoeksvaardigheden bij leerlingen;
4 de mate van samenwerking tussen de docent en zijn collega’s*;
5 het schoolbeleid;
6 eisen die je in het algemeen aan docenten kunt stellen;
7 de mogelijke invloed van de aanwezigheid van de onderzoeker in de onderwijscontext van de docent*.

Het eigen beoordelingsproces van de beoordelaar
1 verwijzingen naar de beoordelaarshandleiding tijdens het beoordelen (bijvoorbeeld “In de handleiding staat …”);
2 opmerkingen over gebrek aan informatie in het portfolio om tot een adequate beoordeling te kunnen komen (bijvoorbeeld “Ze zegt dat ze haar leerlingen feedback geeft, maar in het portfolio vind ik daar geen enkele aanwijzing voor”);
3 toelichten van het eigen beoordelingsproces (bijvoorbeeld “Ik beoordeel dit op deze manier, omdat …”);
4 tekortkomingen en onduidelijkheden tijdens het beoordelen (bijvoorbeeld “Het lastige is …”);
5 het eigen beoordelingsproces bevragen (al dan niet naar de onderzoeker) (bijvoorbeeld “Doe ik het nu goed?”)*;
6 het vaststellen en corrigeren van fouten (bijvoorbeeld “Wat ik nu aan het doen ben, is niet goed”);
7 verwijzingen naar eerder beoordeeld portfoliomateriaal;
8 verwijzingen naar eerder beoordeelde portfolio’s;
9 andere metacognitieve uitingen.

De beoordelingsprocedure
Opmerkingen (positief en negatief) over:
1 beoordelingscriteria en standaarden;
2 portfoliomateriaal;
3 beoordelingsprocedure;
4 weging van de criteria;
5 anders.

* Deze categorie komt niet voor in de beoordelingsformulieren