De betrouwbaarheid en generaliseerbaarheid van competentiebeoordelingen op basis van een videodossier[1]

M. E. J. Bakker, P. Sanders, D. Beijaard, E. Roelofs, D. Tigelaar en N. Verloop

Samenvatting

Er komt steeds meer aandacht voor het ontwikkelen van procedures voor het beoordelen van (docent)competenties. Het waarborgen van de betrouwbaarheid en de validiteit van deze beoordelingsprocedures is hierbij een belangrijk punt. Vanuit de literatuur worden ontwerpprincipes aangedragen die de betrouwbaarheid en de validiteit van competentiebeoordelingen zouden kunnen bevorderen, zoals het verhogen van het aantal beoordelaars en taken, het standaardiseren van taken en het gebruiken van taken die heel direct de beoogde competenties meten. Veelal ontbreekt echter de empirische evidentie voor de werkzaamheid van deze principes. In dit onderzoek is nagegaan in hoeverre deze ontwerpprincipes daadwerkelijk leiden tot betrouwbare en valide competentiebeoordelingen. Voorafgaand aan het onderzoek zijn op basis van ontwerpprincipes videodossiers ontworpen die kunnen worden ingezet bij het beoordelen van de coachcompetentie van docenten in het mbo. Een videodossier bestaat uit verschillende videofragmenten van een coachende docent in kritische situaties in de klas. Om de coachcompetentie van de docenten valide te kunnen beoordelen, zijn de videodossiers aangevuld met bronnen waarin contextinformatie is opgenomen. Na het ontwerpen van de videodossiers is bepaald in hoeverre de gebruikte ontwerpprincipes bijdragen aan betrouwbare en valide competentiebeoordelingen. Het onderzoek tracht antwoorden te geven op de volgende onderzoeksvragen: a) in hoeverre wordt de coachcompetentie van docenten in het mbo op basis van een videodossier betrouwbaar gescoord door beoordelaars? en b) in hoeverre zijn de beoordelingen op afzonderlijke videofragmenten van de coachperformance van docenten generaliseerbaar naar het beoogde universum van videofragmenten? Hiertoe zijn vier videodossiers voorgelegd aan twaalf beoordelaars. Scoreformulieren met toegekende scores zijn verzameld. Er is een acceptabele tot hoge overeenstemming tussen beoordelaars gevonden in toegekende scores aan afzonderlijke videofragmenten en zelfs een hoge overeenstemming in toegekende overallbeoordelingen. Daarnaast is er met uitzondering van één beoordelingsschaal (coaching op leerhouding) een acceptabele tot hoge overeenstemming gevonden tussen de toegekende scores aan een videofragment en de gemiddelde toegekende scores aan de andere videofragmenten binnen een schaal. De toegepaste ontwerpprincipes blijken samen te gaan met positieve resultaten op zowel het gebied van het scoren door beoordelaars als op het gebied van de generaliseerbaarheid van beoordelingen over videofragmenten.

1 Inleiding

De ontwikkeling van instrumenten voor het beoordelen van docentcompetenties staat volop in de belangstelling. Ten behoeve van opleiding en verdere professionalisering van docenten worden steeds vaker instrumenten ontworpen die zowel inzicht verschaffen in de ontwikkeling van competenties als ondersteunend zijn voor de verdere professionele ontwikkeling. Geleidelijk is een kennisbasis ontstaan over de wijze waarop docentcompetenties gemeten kunnen worden. Om met beoordelingsinstrumenten verschillende beoordelingsdoelen te dienen, dient een mix van bewijsbronnen gebruikt te worden waarbij het bewijs zoveel mogelijk in authentieke taaksituaties moet worden verzameld (Darling-Hammond & Snyder, 2000; Dwyer, 1998; Gipps, 1994; Haertel, 1991).

Parallel aan het ontwerpen van instrumenten ten behoeve van competentiebeoordelingen is men gaan nadenken over het waarborgen en het evalueren van de kwaliteit van

dergelijke beoordelingsprocedures. Zoals bij elke beoordeling, zowel summatief als formatief, is het belangrijk om te weten in hoeverre de betekenis die aan de uitkomsten van de beoordeling gegeven wordt, gepast is (Messick, 1989). Is er wel gemeten wat beoogd wordt te meten en kunnen er op basis van de beoordeling gepaste beslissingen genomen worden ten aanzien van selectie, certificering of professionele ontwikkeling?

Een belangrijk aandachtspunt in de evaluatie van de kwaliteit van competentiebeoordelingen is dat deze uit meerdere facetten bestaan (Kane, 2004). Bij een competentiebeoordeling zijn niet alleen respondenten betrokken die vragen of opdrachten voorgelegd krijgen, maar ook beoordelaars die de *performance* van een kandidaat moeten interpreteren en beoordelen. Voor een onderzoek naar de validiteit van een beoordelingsprocedure betekent dit dat zowel de taken als het scoren door beoordelaars aan een onderzoek zullen moeten worden onderworpen.

Kane (2006) heeft een procedure ontwikkeld aan de hand waarvan de validiteit van assessments kan worden geëvalueerd. Hij stelt in zijn *validity argument*-methode dat de validiteit van een assessment kan worden onderzocht door de gevolgtrekkingen in verschillende fasen van een argumentatieketen te onderzoeken. Kane onderscheidt drie fasen in een zogeheten keten van gevolgtrekkingen: 1) het betrouwbaar en valide scoren van *performance* op assessmenttaken (door beoordelaars), 2) het generaliseren van geobserveerde scores op assessmenttaken naar een breder universum van taken en 3) extrapolatie van resultaten naar het praktijkdomein. In een validiteitsonderzoek wordt volgens deze methode de houdbaarheid van gedane gevolgtrekkingen in elk van deze fasen onderzocht.

Tot voor kort werd bij *performance*-assessments voornamelijk gekeken naar de interbeoordelaarsbetrouwbaarheid als maat voor betrouwbaar scoren (Dunbar, Koretz, & Hoover, 1991). Het scoren van *performance* blijkt echter niet eenvoudig te zijn (Gipps, 1994; Moss, 1994). De voornaamste oorzaak is dat bij competentiebeoordelingen gebruik wordt gemaakt van open en complexe taken, die vaak in wisselende (authentieke) contexten worden uitgevoerd. Respondenten kunnen op vele verschillende manieren reageren op een taak en het is voor beoordelaars niet gemakkelijk om de zeer uiteenlopende informatie in de wisselende contexten consistent te beoordelen. Vooral selectieve waarneming, vooroordelen en persoonlijke overtuigingen van de beoordelaar zijn serieuze bedreigingen voor de betrouwbaarheid en de validiteit van het scoren (Gipps, 1994; Moss, 1994).

De laatste jaren is er steeds meer aandacht gekomen voor de mate waarin taken die zijn opgenomen in beoordelingsprocedures generaliseerbaar zijn naar een breder domein van taken en in hoeverre de steekproef van taken een representatieve afspiegeling vormt van het construct dat men wil meten (en dus te extrapoleren is naar *performance* buiten de beoordelingsituatie). Het probleem bij competentiebeoordelingen is dat de open en complexe taken veel meer tijd kosten dan bijvoorbeeld het beantwoorden van vragen uit een toets of vragenlijst. Er kan daardoor slechts een beperkt aantal taken opgenomen worden in de beoordelingsprocedure, waardoor het moeilijk is om te generaliseren naar een breder domein van taken en om te extrapoleren naar *performance* buiten de beoordelingssituatie (Brennan, 2000; Dunbar, Koretz, & Hoover, 1991; Linn, Baker, & Dunbar, 1991; Linn & Burton, 1994; Miller & Linn, 2000; Ruiz-Primo, Baxter, & Shavelson, 1993; Shavelson, Baxter, & Gao, 1993).

Er zijn verschillende ontwerpprincipes waarmee de houdbaarheid van gevolgtrekkingen in de fasen van scoren, generaliseren en extrapoleren van competentiebeoordelingen kan worden gewaarborgd. Voorbeelden zijn het verhogen van het aantal beoordelaars en het aantal taken, het standaardiseren van taken en het gebruiken van authentieke taken. In paragraaf 2 worden deze algemene ontwerpprincipes besproken. Het doel van dit onderzoek is na te gaan in hoeverre deze ontwerpprincipes daadwerkelijk bijdragen aan valide competentiebeoordelingen. Vooraf gaand aan het onderzoek is een beoordelingsprocedure ontwikkeld die is gebaseerd op de ontwerpprincipes voor betrouwbaar en valide scoren, generaliseren en extrapoleren. In paragraaf 3 wordt uitgebreid ingegaan op de manier waarop de algemene ontwerpprinci-

pes uit paragraaf 2 zijn toegepast bij het ontwikkelen van de beoordelingsprocedure in dit onderzoek. De ontwikkelde procedure beoogt de coachcompetentie te meten van docenten in het mbo. Sinds de invoering van zelfstandig leren in het mbo wordt van docenten verwacht dat zij leerlingen kunnen coachen bij het zelfstandig werken en leren in het kader van langlopende opdrachten (Moerkamp, De Bruijn, Van der Kuip, Onstenk, & Voncken, 2000; Onstenk, 2000). De docenten worden beoordeeld op basis van een videodossier. Een videodossier is een gedocumenteerde verzameling van videofragmenten die het handelen toont van docenten in doelbewust gekozen kritische lessituaties. Om de *performance* in de videofragmenten valide te kunnen beoordelen, zijn verschillende bronnen met contextinformatie toegevoegd. De samenstelling en beoordeling van een videodossier zal nader worden beschreven in paragraaf 3. Na het ontwikkelen van vier videodossiers zijn deze voorgelegd aan twaalf beoordelaars en is er een validiteitsonderzoek uitgevoerd.

2 Betrouwbaar en valide scoren, generaliseren en extrapoleren

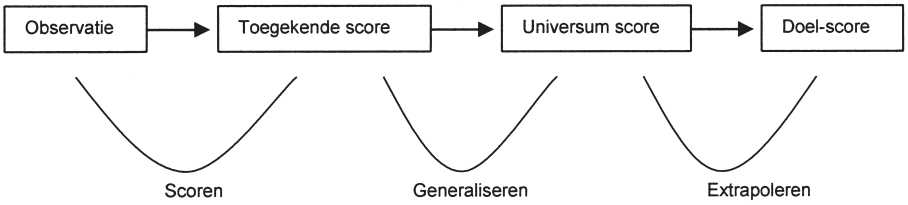
Onder de betrouwbaarheid van een beoordelingsprocedure verstaan we de mate van herhaalbaarheid van een beoordeling. De herhaalbaarheid heeft betrekking op de vraag in hoeverre beoordelingen variëren, wanneer de beoordeling onder gelijkblijvende condities wordt herhaald.

Het begrip validiteit heeft in de afgelopen decennia een ontwikkeling doorgemaakt. In het verleden werden drie verschillende perspectieven op validiteit onderscheiden: criteriumvaliditeit, inhoudsvaliditeit en constructvaliditeit. Criteriumvaliditeit betreft de samenhang van de beoordeling met een extern criterium dat wordt gezien als een directe meting van de eigenschap die men beoogt te meten. Inhoudsvaliditeit richt zich op de mate waarin een beoordeling representatief is voor het te meten domein. Constructvaliditeit gaat over de vraag in welke mate het construct (of eigenschap) gemeten wordt dat beoogd wordt te meten. Deze klassieke driedeling is tegenwoordig minder gangbaar. Constructvaliditeit wordt steeds meer gezien als een overkoepelend begrip dat ook de criterium- en inhoudsvaliditeit omvat (Messick, 1989). Validiteit wordt dan gezien als "de mate waarin empirisch bewijs en theoretische redeneringen de *adequaatheid* en *gepastheid* van gevolgtrekkingen en acties op basis van de beoordeling ondersteunen" (Messick, 1989). Tegen deze opvatting van constructvaliditeit worden ook bezwaren ingebracht die betrekking hebben op de brede opvatting van validiteit, die impliceert dat bijna alle testgerelateerde zaken als relevant worden gezien voor de validiteit (Borsboom & Mellenbergh, 2004). Desondanks is de visie van Messick op validiteit vanaf eind tachtiger jaren vrij algemeen gangbaar geworden.

De validiteit van een beoordelingsprocedure kan systematisch onderzocht worden door de gevolgtrekkingen waarop de beoordeling is gebaseerd expliciet te maken en te onderzoeken volgens de methode van Kane (2004, 2006). De drie gevolgtrekkingen die centraal staan bij competentiebeoordelingen zijn: scoren, generaliseren en extrapoleren. Deze gevolgtrekkingen zijn schematisch weergegeven in Figuur 1.

2.1 Betrouwbaar en valide scoren

De eerste gevolgtrekking in de keten heeft betrekking op het scoren van de *performance* van kandidaten door beoordelaars: is de toegekende score voldoende betrouwbaar en valide? Vooral persoonlijke invloeden op de score vormen een serieuze bedreiging voor deze gevolgtrekking. De invloeden komen tot uiting in het selectief waarnemen, het hanteren van vooroordelen en het mee laten wegen van persoonlijke overtuigingen (Gipps, 1994; Moss 1994). Er zijn verschillende factoren die de betrouwbaarheid en de validiteit van scoretoekenning beïnvloeden. Ten eerste zullen de beoordelingen meer betrouwbaar en valide zijn wanneer er gepaste criteria, standaarden en scoringsvoorschriften worden gebruikt door de beoordelaars en wanneer deze consistent worden toegepast. Het trainen van beoordelaars in het toepassen van criteria en scoringsvoorschriften blijkt een positieve invloed te hebben op de mate waarin beoordelaars consistent scoren (Day & Sulsky, 1995;



Figuur 1. Keten van gevolgtrekkingen waarop een performance assessment is gebaseerd.

Stamoulis & Hauenstein, 1993). Ten tweede heeft het aantal beoordelaars invloed op vooral de betrouwbaarheid van de verkregen scores (Kane, 2006). Naarmate de *performance* door meer personen wordt beoordeeld, hebben persoonlijke invloeden van individuele beoordelaars minder invloed, waardoor de toegekende score nauwkeuriger wordt. Ten derde is gebleken dat beoordelaars beter in staat zijn om de *performances* consistent en overeenstemmend te scoren, wanneer alle te beoordelen personen dezelfde taken krijgen voorgelegd tijdens een beoordelingsprocedure (Crooks, Kane, & Cohen, 1996).

2.2 Generaliseren over assessmenttaken

Bij het bepalen van de houdbaarheid van de tweede gevolgtrekking staat de volgende vraag centraal: representeert de score die de respondent behaald heeft op de verschillende taken de score die de respondent behaald zou hebben als deze alle mogelijke taken zou hebben gemaakt die in aanmerking komen om de specifieke eigenschap te meten? Het gaat hierbij dus om de vraag of de kandidaat een andere score zou hebben behaald als deze andere taken zou hebben uitgevoerd. Dit heeft te maken met de vraag of de steekproef van taken die tijdens de beoordelingsprocedure wordt voorgelegd aan respondenten een verantwoorde representatie is van alle taken in het universum. Het universum van taken verwijst in dit verband naar de verzameling van alle denkbare taken die acceptabel of geschikt zouden zijn voor het meten van de specifieke eigenschap (Sanders, 1998). Dit aspect blijkt vooral bij competentiebeoordelingen problematisch te zijn. Respondenten blijken een aanzienlijke variatie te vertonen in *performance* op verschillende taken, zelfs op taken die uit eenzelfde domein komen. Een maatregel om dit probleem te ondervangen is het standaardiseren van de taken. Met het standaardiseren van de taken beoogt men taken te creëren die een beroep doen op steeds dezelfde eigenschap, waardoor de interne consistentie tussen de taken groter wordt. Het is eenvoudiger om voor gestandaardiseerde taken (gedetailleerde) scoringsvoorschriften te ontwikkelen en het is voor beoordelaars eenvoudiger om de *performance* consistent te scoren (Brennan, 2000; Kane, 2006). De standaardisatie van taken zal uiteindelijk moeten leiden tot meer consistentie tussen de beoordelingen op de verschillende taken, zodat de beoordelingen beter gegeneraliseerd kunnen worden naar het universum van taken.

2.3 Extrapoleren naar *performances* buiten de beoordelingssituatie

Bij de bepaling van de houdbaarheid van de laatste gevolgtrekking gaat het om de vraag in hoeverre de taken het beoogde doel- of competentiedomein meten. Het opnemen van assessmenttaken in de beoordelingsprocedure die heel direct het beoogde domein van gedrag en kennis meten dat je wilt meten, is een maatregel die je kunt nemen om verantwoord te extrapoleren. Deze taken worden ook wel *high-fidelity*-taken genoemd (Kane, 2006). Dit zijn echter vaak lange, complexe taken die moeilijk te scoren zijn. Daarnaast zijn deze lange, complexe taken erg tijdrovend met als gevolg dat er vanwege de praktische uitvoerbaarheid van de beoordelingsprocedure slechts weinig taken worden opgenomen. Door het soms noodgedwongen gebruik van kleine aantallen taken in een beoordelingsprocedure is het moeilijk om een representatieve steekproef van taken te realiseren teneinde te kunnen extrapoleren naar *performance* buiten de beoordelingssituatie. Juist een groot aantal taken blijkt een aanzienlijke

positieve invloed te hebben op verantwoord extrapoleren (Dunbar, Koretz, & Hoover, 1991; Ruiz-Primo, Baxter, & Shavelson, 1993). Dit blijkt een lastig probleem waar nog geen goede oplossing voor is.

In dit onderzoek zullen twee van de drie gevolgtrekkingen uit het model van Kane (2006) worden onderzocht. Er wordt getracht een antwoord te geven op de volgende vragen:

1) In hoeverre wordt de coachcompetentie van docenten in het MBO op basis van een videodossier betrouwbaar gescoord door beoordelaars?, en
2) In hoeverre zijn de beoordelingen op afzonderlijke videofragmenten van de *coachperformance* van docenten generaliseerbaar naar het beoogde universum van videofragmenten?

Bij de eerste onderzoeksvraag beperken we ons tot de vraag in hoeverre beoordelaars de videodossiers betrouwbaar gescoord hebben. Bij deze gevolgtrekking is het tevens belangrijk om te bekijken in hoeverre beoordelaars valide zijn in het scoren, dat wil zeggen in hoeverre beoordelaars criteria toepassen die ze verondersteld worden toe te passen. Het onderzoeken van de validiteit van het scoringsproces van beoordelaars is echter een complex proces, waarbij uitgebreide kwalitatieve analyses betrokken zijn met betrekking tot argumenten en overwegingen die een rol spelen in het beslisproces van beoordelaars. Dit aspect zal uitgebreid aan de orde komen in een ander (vervolg)onderzoek. Voor het beantwoorden van de tweede onderzoeksvraag wordt doorgaans een generaliseerbaarheidsstudie uitgevoerd. Echter, omdat de constructie van de videoportfolio's volgens de ontwerpprincipes een complex en tijdrovend proces bleek, was het niet mogelijk om een aanzienlijke steekproef van videoportfolio's te realiseren die nodig is om de generaliseerbaarheid van scores vast te stellen op basis van een generaliseerbaarheidsstudie. Vandaar dat de generaliseerbaarheid over assessmenttaken op een andere wijze is onderzocht. In dit onderzoek is niet onderzocht in hoeverre de beoordelingen van de *performance* in het videodossier te extrapoleren zijn naar beoordelingen op basis van *coachperformance* in de praktijk (de derde gevolgtrekking uit het model van Kane). Het beantwoorden van deze vraag vergt namelijk een *job*-analyse waarbij in kaart moet worden gebracht welke coachsituaties zich voordoen in de praktijk, wanneer specifieke coachsituaties zich voordoen en hoe vaak deze specifieke situaties zich voordoen. Aangezien een *job*-analyse voor de coachcompetentie van docenten in het mbo nog niet voorhanden is, is ervoor gekozen om het aspect van extrapolatie niet mee te nemen in dit onderzoek.

3 Methode

3.1 Het ontwerp van de beoordelingsprocedure

Beoordelingskader

Op basis van een literatuurstudie op het gebied van het begeleiden van zelfstandig leren (Boekaerts, 1999; Boekaerts & Simons, 1995; Bolhuis, 2003; Butler & Winne, 1995) en observaties in de praktijk is coachen gedefinieerd als het ondersteunen van leeractiviteiten die de leerlingen (nog) niet zelfstandig uitvoeren. Belangrijke interventies om de leerlingen te ondersteunen in het uitvoeren van leeractiviteiten zijn het stellen van vragen en het geven van feedback (Boekaerts & Simons, 1995; Butler & Winne, 1995). Deze coachinterventies kunnen worden ingezet om vier verschillende soorten leeractiviteiten te ondersteunen: leeractiviteiten die betrekking hebben op het verwerven en verwerken van domeinspecifieke kennis en vaardigheden (cognitieve leeractiviteiten), leeractiviteiten die betrekking hebben op het sturen van het eigen leren op basis van langlopende opdrachten (metacognitieve leeractiviteiten), leeractiviteiten die betrekking hebben op het op peil houden van de leerhouding (affectieve leeractiviteiten) en leeractiviteiten die betrekking hebben op het samenwerken met andere leerlingen. De eerste drie leeractiviteiten zijn gebaseerd op Shuell (1993) en Vermunt en Verloop (1999). De vierde leeractiviteit is gebaseerd op inzichten van Johnson en Johnson (1994) en Slavin (1990).

Vervolgens is een beoordelingsschaal ont-

wikkeld waarop het niveau van competentie in coachen kan worden uitgedrukt. Om niveaus van competentie te kunnen operationaliseren is in dit onderzoek aangesloten bij de definitie van competent handelen van Roelofs en Sanders (2007). Zij beschrijven competent handelen als het vermogen om op basis van een persoonlijke kennisbasis verantwoorde beslissingen te nemen bij het uitvoeren van taken in een specifieke context, resulterend in werkgedrag dat bijdraagt aan vooraf als wenselijk beschouwde resultaten. Het begrip competent coachen is in dit onderzoek vervolgens gedefinieerd als de mate waarin de docent constructief coacht. Van constructief coachen is sprake wanneer de docent interventies gebruikt die de leerlingen de kans bieden en hen uitdagen om hun leeractiviteiten, van het type zoals hierboven aangeduid, te verbeteren (Vermunt & Verloop, 1999). De aanduiding constructief is gebaseerd op het principe dat er precies voldoende ondersteuning moet worden aangeboden door de docent, zodat de leerlingen net een stap verder kunnen komen dan ze zelfstandig hadden kunnen bereiken (Vygotsky, 1978). Voor het coachen betekent dit dat wanneer de leerling beter wordt in het uitvoeren van een leeractiviteit, de ondersteuning van de docent afneemt totdat de leerling de leeractiviteit volledig zelfstandig kan uitvoeren. Dit wordt in de literatuur ook wel *fading* genoemd (Collins, Brown, & Newman, 1989). In Tabel 1 is de beoordelingsschaal met vier competentieniveaus opgenomen.

Videodossiers

Het ontwikkelde beoordelingskader is gebruikt om videodossiers te scoren. Een videodossier bestaat uit een mix van informatiebronnen om een zo volledig mogelijk beeld te geven van, in dit geval, de coachcompetentie van een docent. De belangrijkste informatiebron in het dossier zijn de videofragmenten die verschillende kritische situaties tonen waarin een docent zijn of haar coachperformance laat zien. Verder zijn vier bronnen met contextinformatie toegevoegd: een samenvatting van wat er tijdens het videofragment te zien is en wat vooraf ging aan het videofragment, achtergrondinformatie over de leerlingen die tijdens het videofragment te zien zijn (leeftijd, vooropleiding, begeleidingsbehoefte enz.), een beschrijving van het lesmateriaal dat tijdens het videofragment gebruikt wordt en een interview met de docent en met de leerling(en) waarin gereflecteerd wordt op de coachsituatie. Het

Tabel 1

Beoordelingsschaal voor constructief coachen

Niveau	Omschrijving
Niveau 4 Groeispurt	De coach hanteert (een) interventie(s) die de deelnemers stimuleert bij het uitvoeren van de leeractiviteit het beste uit zichzelf te halen. En/of: Hij/zij benut zo goed als alle kansen kans om niveauverhoging gericht te stimuleren.
Niveau 3 Groei	De coach hanteert (een) interventie(s) die de deelnemers stimuleert de leeractiviteit op een hoger niveau uit te voeren. En/of: Hij/zij laat nog enkele kansen liggen, maar dit weerhoudt de deelnemers er niet van om de leeractiviteit op een hoger niveau uit te voeren.
Niveau 2 Incidentele groei	De coach hanteert (een) interventie(s) die de deelnemers incidenteel stimuleert om de leeractiviteit op een hoger niveau uit te voeren; om meer uit zichzelf te halen. En/of: Hij/zij laat nog veel kansen liggen om niveauverhoging gericht te stimuleren.
Niveau 1 Stilstand	De coach hanteert (een) interventie(s) die de deelnemers niet stimuleert de leeractiviteit op een hoger niveau uit te voeren. En/of: Hij/zij laat bijna iedere kans liggen (of herkent die niet) om niveauverhoging gericht te stimuleren.

interview met de docent bestaat uit vragen die betrekking hebben op de aanleiding van het coachen, het doel dat de docent voor ogen had, de aanpak die de docent hanteerde en de mate waarin de docent tevreden was over zijn of haar *coachperformance*. Aan de leerlingen werd gevraagd in hoeverre de ondersteuning van de docent hen verder heeft geholpen met een specifiek onderwerp of probleem en of de ondersteuning op het juiste moment kwam.

Scoringsprocedure

De videodossiers zijn aan getrainde beoordelaars voorgelegd die de dossiers hebben gescoord en beoordeeld volgens een specifiek stappenplan (zie Tabel 2).

Tabel 2

Scoringsprocedure voor het beoordelen van videodossiers

Stap 1 Het verzamelen van bewijs uit een video fragment

Bekijk de volgende informatiebronnen in het dossier:
- Achtergrondinformatie van de leerlingen
- Samenvatting van de video fragmenten
- Interview met de docent

Bekijk de videofragmenten en beantwoord de volgende vragen:
- Welke coachinterventies dragen met name wel/niet bij aan de groei van de leerlingen in het uitvoeren van leeractiviteiten?
- Zijn er belangrijke kansen die de docent heeft laten liggen om de groei van leerlingen in het uitvoeren van leeractiviteiten te stimuleren?
- Zoek naar zowel positief bewijs als naar negatief bewijs. Negatief bewijs bestaat uit (a) interventies waarmee de docent de leerlingen geen kans biedt om verder te groeien in het uitvoeren van leeractiviteiten en (b) uit gemiste kansen om de groei van leerlingen in het uitvoeren van leeractiviteiten te stimuleren
- Maak aantekeningen op het scoreformulier
- Bepaal welke interventies relevant bewijs zijn

Stap 2 Het toekennen van een score aan de coachperformance in een videofragment

Bekijk al het verzamelde bewijs:
- Welke bewijzen zijn belangrijk en welke bewijzen zijn minder belangrijk?
- Hoe kunnen de bewijzen tegen elkaar afgewogen worden?
- Welk patroon is in de bewijzen te ontdekken? Wijzen de bewijzen in een bepaalde richting of juist niet?
- Bepaal na het toekennen van het eindoordeel of er bewijs over het hoofd gezien is en pas het eindoordeel eventueel aan
- Bepaal in hoeverre het bewijs overeenkomt met de beschrijving op niveau 1, 2, 3 of 4. Gebruik hierbij de beoordelingsschaal
- Schrijf een samenvatting waarin je uitlegt waarom de toegekende score een 1, 2, 3 of 4 moet zijn. Verwijs daarbij naar relevant bewijs en relevante argumenten

Stap 3 Het toekennen van een overallscore aan de coachperformance over meerdere videofragmenten

- Bepaal in hoeverre de performances over alle videofragmenten heen overeenkomt met de beschrijving op niveau 1, 2, 3 of 4. Gebruik hierbij de beoordelingsschaal
- De toegekende score hoeft niet gelijk te zijn aan de gemiddelde score over de videofragmenten. Je krijgt de kans om de coachperformance in de videofragmenten zwaarder of lichter mee te laten tellen. Op deze manier kun je corrigeren voor complexiteit van bepaalde coachsituaties en voor (extreem) goede of slechte coaching in specifieke situaties
- Hoe kunnen de beoordelingen van de afzonderlijke coachperformances tegen elkaar afgewogen worden?
- Welk patroon is er in de coachperformances te ontdekken? Wijzen de prestaties van de docent duidelijk in een bepaalde richting of wisselen de prestaties?
- Bepaal na het toekennen van het eindoordeel of alle situaties in beschouwing zijn genomen en pas het eindoordeel eventueel aan
- Schrijf een samenvatting waarin je uitlegt waarom de toegekende score een 1, 2, 3 of 4 moet zijn. Verwijs daarbij naar relevant bewijs en relevante argumenten

Stap 4 Discussie met een collega beoordelaar

- Nadat je een videodossier individueel hebt beoordeeld, bespreek je de beoordeling met een collega-beoordelaar
- Vergelijk niet alleen de toegekende scores, maar ook de bewijzen en argumenten
- Spits de discussie toe op verschillen in toegekende scores en onderbouwingen van deze scores
- Na het overleg heb je de kans om bij de originele beoordeling te blijven of om de beoordeling aan te passen

3.2 Maatregelen om betrouwbaar en valide te kunnen scoren

Beoordelingskader
In het ontwerp van de beoordelingsprocedure zijn verschillende maatregelen opgenomen om het scoren van de *performance* zo betrouwbaar en valide mogelijk te maken. Om persoonlijke invloeden van beoordelaars, zoals selectief waarnemen en beoordelen op basis van persoonlijke overtuigingen en constructen (DeNisi, Cafferty, & Meglino, 1984; Feldman, 1981; Landy & Farr, 1980; Van der Schaaf, Stokking, & Verloop, 2005) zoveel mogelijk te vermijden, is een beoordelingskader ontwikkeld met gedetailleerde scoringsvoorschriften. Bovendien zijn de beoordelaars getraind in het hanteren van het beoordelingskader.

Theorie en praktijk
Bij de ontwikkeling van het beoordelingskader is begonnen met een literatuuronderzoek. Op basis van dit literatuuronderzoek zijn coachinterventies uitgewerkt waarvan aangenomen kon worden dat deze verschillende leeractiviteiten ondersteunen. Deze uitwerking is voorgelegd aan mbo-docenten en kaderfunctionarissen binnen het mbo. Ook zijn lesbezoeken verricht waarbij is gekeken naar het voorkomen van de verschillende interventies. Op grond hiervan en op grond van gesprekken met betrokkenen is het kader aangepast aan de specifieke context van het mbo. Door het praktijkgerichte perspectief mede te verwerken in het beoordelingskader is getracht om gepaste beoordelingscriteria te ontwikkelen, hetgeen ten goede zou moeten komen aan de validiteit van de beoordelingen.

Concrete voorbeelden
Tijdens de ontwikkeling van het beoordelingskader zijn voorbeelden verzameld van concrete coachinterventies die docenten in de praktijk gebruiken. Deze voorbeeldinterventies moeten de beoordelaars helpen bij het herkennen van coachactiviteiten (Frederiksen, Sipusic, Sherin, & Wolfe, 1998). Door de voorbeelden zouden beoordelaars duidelijk voor ogen moeten hebben welke docentactiviteiten gekwalificeerd kunnen worden als coachactiviteiten en dus beoordeeld zouden moeten worden. Het opnemen van voorbeelden van concrete coachinterventies in het beoordelingskader zou moeten bijdragen aan een hoge interbeoordelaarsovereenstemming.

Competentieniveaus
Ten behoeve van het scoren van de *coachperformance* is een schaal met vier competentieniveaus beschreven. Ieder competentieniveau is voorzien van een beschrijving van het bijbehorende handelen van de docent en de gevolgen voor de leerlingen. De beschrijvingen zijn bedoeld als hulpmiddel voor de beoordelaars, zodat zij makkelijker kunnen afwegen of de *coachperformance* van een docent moet worden gewaardeerd met een score een, twee, drie of vier. Deze aanpak zou ervoor moeten zorgen dat beoordelaars beter in staat zijn om de *performance* in wisselende contexten consistent te beoordelen, wat uiteindelijk ten goede moet komen aan een hoge beoordelaarsovereenstemming.

Scoringsvoorschriften
Een wezenlijk onderdeel van het beoordelingskader vormen de scoringsvoorschriften. Gekozen is voor een scoringsprocedure waarbij beoordelaars specifieke criteria en scoringsvoorschriften moeten hanteren om eerst relevante aspecten van de *performance* te scoren en deze scores daarna te combineren tot een eindoordeel. Door een score zorgvuldig op te bouwen op basis van specifieke criteria en scoringsvoorschriften worden persoonlijke opvattingen en overtuigingen van beoordelaars zoveel mogelijk uitgesloten. Op deze manier worden beoordelingen meer betrouwbaar en valide (Klein & Stecher, 1998). De scoringsprocedure is uitgewerkt conform drie principes ontleend aan Moss, Schutz en Collins (1998): a) beoordelaars moeten de beoordeling baseren op al het beschikbare bewijs, b) beoordelaars moeten actief zoeken naar tegenbewijzen en(c) valide interpretaties ontstaan in discussie met andere beoordelaars, waarbij beoordelaars elkaars interpretaties kritisch onderzoeken. In de scoringsprocedure (Tabel 2) worden de beoordelaars in verschillende stappen aangemoedigd al het verzamelde bewijs bij de beoordeling te betrekken en actief te zoeken naar zowel posi-

tief als negatief bewijs. Principe c) komt tot uitdrukking in de laatste fase van de scoringsprocedure die bestaat uit een discussie met een collegabeoordelaar.

Training voor beoordelaars

Het trainen van beoordelaars in het scoren van *performance* blijkt een positief effect te hebben op de betrouwbaarheid en de validiteit van toegekende scores door beoordelaars (Day & Sulsky, 1995; Stamoulis & Hauenstein, 1993). Daarom is een training van vier dagdelen ontwikkeld waarin het scoren en het beoordelen van videodossiers centraal staat. In de scoringsprocedure is het belangrijk dat beoordelaars dezelfde opvatting hebben over wat coachen is. Tevens moeten ze in staat zijn om specifieke *performances* te waarderen op een schaal met vier competentieniveaus. Om beoordelaars hierin te trainen zijn elementen van een *frame of reference*-training gebruikt (Woerh & Huffcutt, 1994). Daarnaast zijn ook elementen uit een *rating error*-training gebruikt om beoordelaars bewust te maken van beoordelingsfouten en hoe ze deze fouten kunnen vermijden (Woerh & Huttcuff, 1994).

Tijdens de training werden kritische situaties bekeken en besproken, die niet waren opgenomen in de videodossiers. De scoringsprocedure werd stap voor stap geoefend en beoordelaars kregen feedback op:

- het identificeren, selecteren en citeren van bewijs uit een videofragment van een kritische situatie;
- het waarderen van bewijs en het redeneren over bewijs in termen van gevolgen voor de leerlingen;
- het toekennen van scores aan videofragmenten van kritische situaties op basis van de beoordelingsschaal;
- het waarderen van bewijzen in verschillende videofragmenten van kritische situaties en redeneren over bewijs in termen van gevolgen voor leerlingen uit verschillende kritische situaties;
- het toekennen van overallscores aan het videodossier op basis van de beoordelingsschaal, en
- het schrijven van een samenvatting waarin beoordelaars aangeven op basis van welke bewijzen en argumenten ze een bepaalde score hebben toegekend.

Tijdens de training is veel aandacht besteed aan het wegen van bewijs als voorbereidende activiteit op het toekennen van een score. Het betreft zowel het wegen van bewijzen binnen één enkel videofragment van een kritische situatie die vervolgens wordt gescoord als het wegen van de scores op de verschillende videofragmenten voor het toekennen van een overallscore.

Organisatie en ordening van bewijzen

Om optimale beoordeelbaarheid te garanderen zijn drie maatregelen getroffen. Allereerst zijn de video-opnames van coachbijeenkomsten uitgevoerd door een professioneel productiebedrijf dat werkte met drie camera's en drie microfoons die de activiteiten van de docent en de deelnemers synchroon registreerden. Uitgangspunt hierbij was dat alle interactie tussen docent en leerling(en) duidelijk waarneembaar zou moeten zijn voor beoordelaars, zodat er geen bewijs verloren zou gaan. In de uiteindelijke videodossiers zijn alle kritische situaties synchronisaties van drie beeldlijnen: de docent, detailopnames van de leerlingen en, indien van toepassing, het leermateriaal. Bovendien is rondom de beeldbewijzen informatie opgenomen over de context waarin het coachen plaatsvond. De contextinformatie werd overzichtelijk geordend en beoordelaars konden deze aanvullende contextinformatie eenvoudig raadplegen. Beoordelaars blijken deze contextinformatie nodig te hebben om het niveau van de performance te kunnen schatten (Heller, Sheingold, & Myford, 1998; Schutz & Moss, 2004). Tot slot is het bewijs per docent visueel geordend onder vier coachingsbijeenkomsten en bijbehorende videofragmenten van kritische situaties, waardoor beoordelaars bewijs voor verschillende aspecten van competent coachen in samenhang konden evalueren.

3.3 Maatregel om verantwoord te kunnen generaliseren

Om (beter) over videofragmenten te kunnen generaliseren, zijn de videofragmenten gestandaardiseerd. De standaardisatie is gebeurd aan de hand van het selecteren van videofragmenten van specifieke kritische coach-

situaties. De selectie van kritische coachsituaties is gebeurd aan de hand van de volgende definitie: een kritische coachsituatie is een situatie waarin leerlingen ondersteuning behoeven in het uitvoeren van een leeractiviteit die ze moeten ondernemen om een langlopende opdracht te volbrengen. Het is een situatie waarvan verondersteld wordt dat deze bewijzen bevat voor de coachcompetentie van de docent.

3.4 Maatregelen om verantwoord te kunnen extrapoleren

Zoals reeds eerder aangegeven, was het in dit onderzoek niet mogelijk om de maatregelen voor verantwoord extrapoleren te evalueren. Desondanks worden de maatregelen die in dit verband zijn toegepast in de beoordelingsprocedure, hieronder uiteengezet. Bij de ontwikkeling van de beoordelingsprocedure is zoveel mogelijk uitgegaan van *high fidelity*-taken, die heel direct de coachcompetentie van docenten meten. Een manier om de coachcompetentie heel direct te meten is het filmen van docenten in een echte lessituatie. Uit deze opnames zijn kritische coachsituaties geselecteerd die in het videodossier zijn opgenomen. Om valide te kunnen extrapoleren naar de coachcompetentie van de docenten in de praktijk is het van belang dat er een steekproef van situaties wordt samengesteld die zo goed mogelijk de verschillende coachsituaties in de praktijk representeert. Om voldoende variantie in coachsituaties te krijgen, zijn de kritische situaties geselecteerd aan de hand van de volgende criteria: spreiding over de vier weken waarin de leerlingen aan de opdracht werkten en spreiding in de verschillende aan te sturen leeractiviteiten. Een tweede factor die bijdraagt aan het valide extrapoleren naar situaties buiten de assessmentsituatie is het aantal kritische situaties dat wordt opgenomen in het videodossier. Naarmate er meer situaties worden opgenomen in het dossier, zal er beter geextrapoleerd kunnen worden. Hierbij spelen ook praktische overwegingen een rol. Beoordelaars kunnen bijvoorbeeld slechts een beperkt aantal kritische situaties beoordelen binnen een redelijk tijdsbestek. Een belangrijke afweging is dus bij welk aantal kritische situaties een redelijke variatie in situaties gerealiseerd kan worden en die tevens binnen een redelijk tijdsbestek beoordeeld kan worden. In dit onderzoek is besloten om tien kritische situaties op te nemen in het dossier.

3.5 Respondenten

Van vier docenten, werkzaam in het mbp, zijn vier videodossiers gemaakt. De vier docenten (drie mannen en één vrouw) coachten leerlingen uit het eerste jaar van de bol-opleiding bouwkunde en infratechniek (niveau vier). Deze docenten hadden tussen de één en twee jaar ervaring in het coachen van leerlingen. Twee van deze coaches waren zogeheten prestatiebegeleiders, docenten die leerlingen voornamelijk coachen op vakinhoud, reguleren en leerhouding. De andere twee coaches waren zogeheten stamgroepcoaches, docenten die de leerlingen voornamelijk coachen op reguleren, samenwerken en leerhouding. In de ontwikkelde videodossiers is hiermee rekening gehouden. Daarom zijn kritische situaties opgenomen die betrekking hebben op de leeractiviteiten die de docenten begeleiden, uitgaande van hun functieprofiel als prestatiebegeleider of stamgroepcoach.

De videodossiers zijn beoordeeld door twaalf beoordelaars uit hetzelfde vakgebied met ongeveer vergelijkbare coachervaring op het gebied van eerstejaars leerlingen van de bol-opleiding. Zes van de twaalf beoordelaars waren collega's van de docenten van wie een videodossier is gemaakt. De andere zes beoordelaars kwamen van een ander roc.

3.6 Dataverzameling

Na afloop van de vier trainingsessies hebben de beoordelaars de vier videodossiers zelfstandig beoordeeld. Aan ieder videofragment waarin de coachperformance van een docent in een kritische situatie getoond wordt, werd een score toegekend die overeenkomt met één van de vier onderscheiden competentieniveaus. Vervolgens hebben beoordelaars een overallscore bepaald. Ook de overallscore werd uitgedrukt op de vierniveaubeoordelingsschaal (zie Tabel 1). Bij de toekenning van een overallscore werd van de beoordelaar gevraagd om zich een totaalbeeld te vormen van de *coachperformance* over de verschillende videofragmenten heen. In totaal werden er drie overallscores toegekend voor con-

structief coachen op a) vakinhoud, b) reguleren en c) leerhouding als het ging om een prestatiebegeleider en werden er drie overallscores toegekend op a) reguleren, b) leerhouding en c) samenwerken als het ging om een stamgroepcoach. Na de zelfstandige beoordeling van de videodossiers door individuele beoordelaars hebben de beoordelaars in paren hun beoordelingen besproken en eventueel aangepast. Scoreformulieren met toegekende scores zijn verzameld.

3.7 Analyse: het scoren door beoordelaars

Om te onderzoeken in hoeverre de coachcompetentie van docenten in het mbo op basis van een videodossier betrouwbaar kon worden gescoord, zijn verschillende analyses uitgevoerd. Ten eerste is onderzocht of persoonlijke antwoordtendensen van beoordelaars, zoals strengheid en mildheid, van invloed zijn op de toegekende scores. Deze analyses zijn gedaan om allereerst zicht te krijgen op beoordelaars die extreme scores toekennen. Bij deze analyse is de gemiddelde toegekende score per beoordelaar bepaald over tien videofragmenten voor de afzonderlijke vier docenten. Deze analyse is ook gedaan voor de gemiddelde toegekende overallscores per beoordelaar voor de vier docenten. Zo blijken beoordelaars meer of minder mild dan wel streng te beoordelen. Daarnaast zijn beoordelaars bij sommige docenten strenger of milder geweest en is er sprake van een interactie-effect tussen beoordelaars en docenten.

Ten tweede is de overeenstemming onderzocht tussen beoordelaars in toegekende scores. Deze analyses zijn twee keer uitgevoerd: een keer voor de dataset waarbij alle beoordelaars zijn betrokken en een keer voor een dataset waar de beoordelaars met de hoogste en met de laagste toegekende scores buiten beschouwing zijn gelaten. Om een indicatie te krijgen van de mate van overeenstemming in de beoordelingen van videodossiers is nagegaan in hoeveel gevallen beoordelaars tot dezelfde beoordeling komen. Bij deze analyse is bepaald in hoeveel gevallen meer dan de helft van de beoordelaars dezelfde (overall)score toekent, uitgesplitst naar de docenten één, twee, drie en vier. Naast de bepaling van het aantal gevallen waarin meer dan de helft van de beoordelaars dezelfde (overall)score toekent, is de beoordelaarsovereenstemming in toegekende (overall)scores berekend. Als maat voor de beoordelaarsovereenstemming is de Gower-coëfficiënt gebruikt. Vergeleken met andere maten voor beoordelaarsovereenstemming is de Gower-coëfficiënt niet gevoelig voor gebrek aan variantie. Vanwege het gebrek aan variantie in toegekende scores zouden andere maten ten onrechte op gebrek aan beoordelaarsovereenstemming gewezen hebben.

De Gower-coëfficiënt is gebaseerd op de absolute verschillen tussen beoordelaars, dat wil zeggen het aantal gevallen waarin beoordelaars de *performance(s)* op hetzelfde competentieniveau plaatsen en de mate waarin de beoordelingen van de beoordelaars uit elkaar liggen op de beoordelingsschaal wanneer zij de *performance(s)* niet op hetzelfde niveau plaatsen. De Gower-coëfficiënt is gedefinieerd met de volgende formule:

$$G_{xy} = 1 - \{\Sigma \mid X_i - Y_i \mid / nR\}$$

In de formule worden X_i en Y_i (met i = 1, 2,…., n) voorgesteld als de scores toegekend door twee beoordelaars. De n geeft het aantal beoordeelde objecten aan en de R het bereik van de beoordelingsschaal (Zegers, 1989). De Gower-coëfficiënt ligt tussen de 0 (geen overeenstemming tussen beoordelaars) en de 1 (perfecte overeenstemming tussen de beoordelaars). Een Gower-coëfficiënt lager dan 0,65 representeert een lage overeenstemming, een Gower-coëfficiënt tussen de 0,65 en de 0,80 representeert een acceptabele overeenstemming en een Gower-coëfficiënt hoger dan 0,80 representeert een hoge overeenstemming. Zoals de formule aangeeft, wordt de Gower-coëfficiënt tussen twee beoordelaars berekend. In het kader van deze studie is er voor ieder mogelijk beoordelaarspaar een Gower-coëfficiënt berekend.

Ten derde is onderzocht in hoeverre beoordelingen generaliseerbaar zijn over beoordelaars en hoeveel beoordelaars er minimaal ingezet zouden moeten worden om een acceptabele overeenstemming te realiseren. Dit is een belangrijk punt, omdat bij dit onderzoek een hoog aantal, namelijk twaalf, be-

oordelaars betrokken zijn. In de praktijk zal het niet mogelijk zijn om dit aantal beoordelaars in te zetten, gezien de tijd en kosten die dit met zich meebrengt. Naarmate er beter gegeneraliseerd kan worden over beoordelaars, betekent dit dat er minder beoordelaars hoeven te worden betrokken bij een beoordelingsprocedure om een acceptabel niveau van betrouwbaarheid te bereiken. Er is berekend in hoeverre de gemiddelde toegekende scores over twee, drie, vier, vijf, zes, zeven, acht en negen beoordelaars de gemiddelde toegekende score over negen tot tien beoordelaars benaderen. Deze analyses zijn wederom uitgevoerd voor toegekende scores, waarbij de extreme beoordelaars buiten beschouwing zijn gelaten en waarbij de toegekende scores van alle beoordelaars betrokken zijn.

3.8 Analyse: generaliseren over videofragmenten

Om te onderzoeken in hoeverre de scores toegekend aan de *coachperformance* in de videofragmenten generaliseerbaar zijn naar het universum van videofragmenten, zijn twee verschillende analyses uitgevoerd. Als eerste is er een algemene analyse uitgevoerd waarin is onderzocht welke videofragmenten wisselende beoordelingen uitlokken. Op basis van deze analyse kan geen directe uitspraak worden gedaan over de generaliseerbaarheid van videofragmenten, maar de analyse levert wel een overzicht op van de videofragmenten die een bedreiging vormen voor de generaliseerbaarheid naar het universum van videofragmenten. Bij deze analyse is voor elk videofragment de standaarddeviatie bepaald voor de toegekende scores over alle twaalf beoordelaars. Naarmate de standaarddeviatie kleiner is, lokken de videofragmenten dezelfde beoordeling uit bij beoordelaars. Naarmate de standaarddeviatie groter is, lokken de videofragmenten wisselende beoordelingen uit bij beoordelaars. Vervolgens is er een rangorde bepaald van videofragmenten op basis van de gevonden standaarddeviaties. Vooral de videofragmenten laag in de rangorde vormen een bedreiging voor de generaliseerbaarheid naar het universum.

Ten tweede is de overeenstemming bepaald tussen scores die zijn toegekend aan de videofragmenten. Die overeenstemming is gebruikt om te bepalen of geobserveerde scores gegeneraliseerd kunnen worden naar de universumscores. In de ontwikkelde beoordelingsprocedure zijn vier soorten videofragmenten opgenomen in de videodossiers: videofragmenten waarin de docent coacht op vakinhoud, reguleren, leerhouding en samenwerken. De videofragmenten die betrekking hebben op dezelfde soort van videofragmenten vormen samen een schaal. Het is de bedoeling dat de videofragmenten die behoren tot dezelfde schaal, generaliseerbaar zijn naar het universum. Naarmate de scores op de videofragmenten beter generaliseerbaar zijn, betekent dit voor de praktijk dat minder videofragmenten van de betreffende schaal in het dossier hoeven te worden opgenomen om een acceptabel betrouwbare beoordeling te realiseren. Voor elk videofragment is de overeenstemming bepaald met de gemiddelde toegekende restscore van de schaal waartoe het videofragment behoort. Een gemiddelde restscore is de gemiddelde score over alle videofragmenten van de schaal exclusief het videofragment waarvoor de overeenstemming wordt bepaald. De overeenstemming is wederom berekend op basis van een Gowercoëfficiënt.

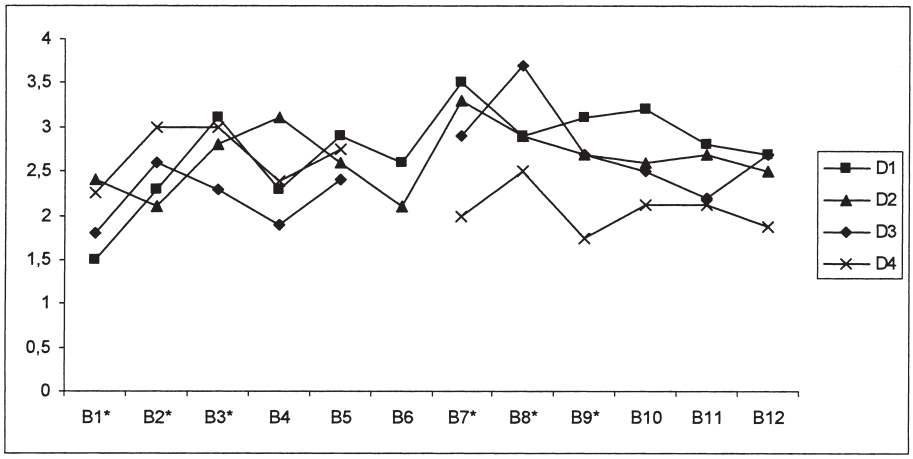
4 Resultaten

4.1 Resultaten: het scoren door beoordelaars

Antwoordtendensen van beoordelaars

In Figuur 2 zijn de gemiddelde toegekende scores afgezet tegen de beoordelaars voor de afzonderlijke docenten.

Figuur 2 laat zien dat de lijnen voor de docent drie en vier onderbroken zijn, omdat beoordelaar zes de docenten drie en vier niet heeft beoordeeld. In Figuur 2 is duidelijk te zien dat de lijnen die de gemiddelde score van de docenten aangeven, niet evenwijdig lopen. Dit betekent dat de verschillende beoordelaars niet bij alle docenten even streng of even mild geweest zijn. De uitkomsten van de analyses voor de gemiddelde overallscores laten hetzelfde beeld zien. De lijnen van de gemiddelde toegekende overallscore



Figuur 2. Gemiddelde toegekende score aan tien videofragmenten door twaalf beoordelaars voor de docenten één, twee, drie en vier.

* Deze beoordelaars zijn collega's van de beoordeelde docenten.

lopen eveneens niet evenwijdig, wat duidt op een interactie-effect tussen beoordelaars en docenten.

Op basis van voorgaande analyses is tevens gebleken dat voornamelijk collega's van de beoordeelde docenten een extreme beoordeling geven. Uit Figuur 2 is af te lezen dat bij de beoordeling van de videofragmenten van docent één, beoordelaar één de strengste beoordelaar was. Voor docent twee zijn dit beoordelaars twee en zes, voor docent drie is dit beoordelaar één en voor docent vier is dit beoordelaar negen. Uit Figuur 2 blijkt ook wie de meest milde beoordelaars zijn voor elke docent. Vervolgens is bekeken welke beoordelaars de meest extreme beoordelingen geven. In 90% van de gevallen werd een extreme beoordeling gegeven door beoordelaars die een collega waren van de beoordeelde docent. Bij de overallscores werd er in 60% van de gevallen een extreme beoordeling gegeven door beoordelaars die een collega waren van de beoordeelde docent. De beoordelaars blijken hun collega's deels extreem streng en deels extreem mild te beoordelen.

Overeenstemming tussen beoordelaars: frequentie

Als eerste is bepaald in hoeveel gevallen meer dan de helft van de beoordelaars dezelfde score heeft toegekend aan de videofragmenten uit het videodossier van de vier docenten. Ten tweede is bepaald in hoeveel gevallen meer dan de helft van de beoordelaars dezelfde overallscore heeft toegekend. Voor docent één is voor zes van de tien videofragmenten in het videodossier door meer dan de helft van de beoordelaars dezelfde score toegekend. Voor docent twee geldt dit voor acht van de tien videofragmenten en voor de docenten drie en vier slechts voor twee van de tien en drie van de acht videofragmenten. Uit deze gegevens blijkt dat beoordelaars het vaker eens zijn over videofragmenten van de docenten één en twee dan over de videofragmenten van docenten drie en vier. Ten aanzien van de toegekende overallscores is hetzelfde beeld te zien. Ook hier blijken beoordelaars het meer eens te zijn over de docenten één en twee dan over de docenten drie en vier.

Overeenstemming tussen beoordelaars: Gower-coëfficiënt

In Tabel 3 zijn de gemiddelde Gower-coëfficiënten voor alle mogelijke beoordelaarsparen opgenomen voor zowel de overeenstemming in scores die zijn toegekend aan de videofragmenten als de overeenstemming in de overallscores. In deze tabel zijn de gemiddelde Gower-coëfficiënten uitgesplitst naar beoordeelde docenten en is ook de range van de gevonden Gower-coëfficiënten tussen de beoordelaarsparen weergegeven.

Tabel 3

Overzicht Gower-coëfficiënten voor toegekende fragment- en overalloordelen

	Toegekende scores aan videofragmenten	Bereik van Gower-coëfficiënten	Toegekende overall scores	Bereik van Gower-coëfficiënten
Alle docenten				
38 videofragmenten/12 beoordelaars	0,74	0,63-0,87	0,80	0,61-0,95
11 overallscores/12 beoordelaars	0,73*	0,56-0,85*	0,78*	0,53-0,95*
Docent 1				
10 videofragmenten/12 beoordelaars	0,80	0,56-00,93	0,79	0,33-1,00
3 overallscores/12 beoordelaars	0,75*	0,33-0,93*	0,75*	0,33-1,00*
Docent 2				
10 videofragmenten/12 beoordelaars	0,80	0,59-0,92	0,93	0,78-1,00
3 overallscores/12 beoordelaars	0,78*	0,54-0,92*	0,85*	0,56-1,00*
Docent 3				
10 videofragmenten/11 beoordelaars	0,71	0,52-0,85	0,76	0,56-1,00
3 overallscores/10 beoordelaars	0,68*	0,37-0,90*	0,68*	0,22-1,00*
Docent 4				
8 videofragmenten/11 beoordelaars	0,76	0,63-0,90	0,82	0,67-1,00
2 overallscores/11 beoordelaars	0,73*	0,57-0,92*	0,82*	0,67-1,00*

* Gower-coëfficiënt wanneer extreem strenge en milde beoordelaars wel zijn meegenomen in de analyse.

De Gower-coëfficiënten voor de toegekende scores aan videofragmenten liggen tussen 0,71 (docent drie) en 0,80 (docenten één en twee) wanneer de extreem strenge en milde beoordelaars uit de analyses zijn gelaten. Wanneer deze beoordelaars wel worden meegenomen in de analyses, ligt de overeenstemming iets lager (tussen 0,68 en 0,78). Deze Gower-coëfficiënten geven aan dat er sprake is van een acceptabele overeenstemming over de toegekende scores aan videofragmenten. De Gower-coëfficiënten voor de toegekende overallscores liggen tussen 0,76 (docent drie) en 0,93 (docent twee), wanneer de extreem strenge en milde beoordelaars uit de analyses zijn gelaten. Deze overeenstemming voor de toegekende overallscores is hoog te noemen. Wanneer de extreme beoordelaars wel worden meegenomen, ligt ook bij de toegekende overallscores de overeenstemming iets lager (tussen 0,68 en 0,85); deze is acceptabel tot hoog te noemen.

Generaliseerbaarheid van scores over beoordelaars

Uit de analyses is gebleken dat de gemiddelde toegekende score van twee beoordelaars een overeenstemming heeft met de gemiddelde toegekende score over tien beoordelaars van 0,88 tot 0,91. Deze resultaten laten zien dat de gemiddelde toegekende score op basis van twee beoordelaars het gemiddelde over twaalf beoordelaars al heel dicht benadert. Wanneer de extreme beoordelaars worden meegenomen in de analyses, is er een overeenstemming gevonden van 0,72 tot 0,90 tussen de gemiddelde toegekende score van twee beoordelaars en de gemiddelde toe-

gekende score over twaalf beoordelaars. Dit betekent dat er nog steeds een acceptabele tot hoge overeenstemming is tussen het gemiddelde op basis van twee en twaalf beoordelaars.

4.2 Resultaten: generaliseren over videofragmenten

Overeenstemming op specifieke videofragmenten

De rangorde van videofragmenten van lage naar hoge standaarddeviatie over de toegekende scores is verdeeld in drie groepen: groep één waarbij de toegekende scores een spreiding hebben over twee van de vier schaalpunten (standaarddeviatie van 0,37 – 0,49), groep twee waarbij de toegekende scores een spreiding hebben over drie van de vier schaalpunten (standaarddeviatie 0,51 – 0,79) en groep drie waarbij de toegekende scores een spreiding hebben over alle vier de schaalpunten (standaarddeviatie van 0,83 – 0,99). Van alle 38 videofragmenten zijn er 8 videofragmenten die in groep één vallen, 17 videofragmenten in groep twee en 13 in groep drie. De videofragmenten die de meest eenduidige reacties uitlokken bij beoordelaars, zijn de videofragmenten in groep één. De videofragmenten die de meest wisselende beoordelingen uitlokken, zijn de videofragmenten in groep drie. De videofragmenten in groep één hebben voornamelijk betrekking op het coachen van de vakinhoud door de docenten één en twee. De videofragmenten uit groep twee zijn voornamelijk videofragmenten die betrekking hebben op het coachen van de leeractiviteit reguleren door de docenten één en twee. Daarnaast zijn in deze groep ook de videofragmenten van docent vier te vinden die coacht op de leeractiviteit samenwerken. De groep videofragmenten die de minst eenduidige beoordelingen uitlokken bij beoordelaars, hebben betrekking op het coachen van docent drie. Daarnaast zijn er vier van de zes videofragmenten in deze groep te vinden die betrekking hebben op het coachen van de leerhouding.

Overeenstemming tussen score op videofragmenten en de gemiddelde restscore

In Tabel 4 wordt voor elk videofragment van elke docent de Gower-coëfficiënt weergegeven. Deze geeft een indicatie van de over-

Tabel 4

Gower-coëfficiënten voor de overeenstemming tussen de gemiddelde toegekende scores aan een videofragment en de gemiddelde toegekende restscores aan de andere videofragmenten uit de schaal

Videofragmenten	Docent 1	Docent 2	Docent 3	Docent 4
Vakinhoud 1	0,81	0,83	-	-
Vakinhoud 2	0,83	0,72	-	-
Vakinhoud 3	0,80	0,83	-	-
Vakinhoud 4	0,78	-	-	-
Reguleren 1	0,82	0,82	0,78	0,70
Reguleren 2	0,85	0,83	0,74	0,80
Reguleren 3	0,89	0,82	0,72	0,77
Reguleren 4	-	0,81	0,64	-
Reguleren 5	-	0,71	-	-
Samenwerken 1	-	-	0,66	0,73
Samenwerken 2	-	-	0,78	0,80
Samenwerken 3	-	-	0,74	0,86
Samenwerken 4	-	-	0,66	0,79
Samenwerken 5	-	-	-	0,78
Leerhouding 1 en 2	0,78	0,67	0,53	-

eenstemming tussen de gemiddelde toegekende scores aan het fragment en de gemiddelde toegekende restscore van de schaal waartoe het videofragment behoort.

Uit Tabel 4 blijkt dat de videofragmenten waarin de docent coacht op vakinhoud over het algemeen een redelijk tot hoge overeenstemming vertonen met de restscore. Dit duidt erop dat de scores op de videofragmenten van deze schaal redelijk te generaliseren zijn naar het universum van videofragmenten. De overeenstemming tussen videofragmenten waarin gecoacht wordt op reguleren en de restscore laat een verdeeld beeld zien. Tabel 4 laat zien dat er voor de videofragmenten van de docenten één en twee een hoge overeenstemming is gevonden met de restscore. Voor deze fragmenten geldt dat de scores toegekend aan videofragmenten waarin de docent coacht op reguleren, redelijk te generaliseren zijn naar het universum van videofragmenten waarin gecoacht wordt op reguleren. De overeenstemming tussen de videofragmenten van docenten drie en vier en de restscore ligt lager. Voor deze videofragmenten is het moeilijker om scores te generaliseren naar het universum van videofragmenten waarin gecoacht wordt op reguleren. Uit Tabel 4 blijkt verder dat de overeenstemming tussen videofragmenten waarin gecoacht wordt op samenwerken en de restscore acceptabel is. Deze videofragmenten zijn net als de videofragmenten voor docenten drie en vier waarin gecoacht wordt op reguleren minder homogeen, waardoor het moeilijker is om de scores te generaliseren naar het universum van videofragmenten waarin gecoacht wordt op samenwerken. In Tabel 4 is te zien dat de overeenstemming tussen de twee videofragmenten waarin gecoacht wordt op leerhouding het meest problematisch is. Voor deze videofragmenten is er een acceptabele tot lage overeenstemming gevonden. Voor deze videofragmenten is het moeilijk om de score te generaliseren naar het universum van videofragmenten.

5 Conclusie en discussie

Dit onderzoek beoogde te achterhalen of en op welke manier ontwerpprincipes een bijdrage leveren aan het waarborgen van de betrouwbaarheid en validiteit van competentiebeoordelingen. Nagegaan is 1) in hoeverre de coachcompetentie van docenten in het mbo op basis van een videodossier betrouwbaar wordt gescoord door beoordelaars en 2) in hoeverre de beoordelingen van de *coach-performance* van docenten in afzonderlijke videofragmenten generaliseerbaar zijn naar het beoogde universum van videofragmenten.

5.1 Het scoren door beoordelaars

De eerste conclusie die getrokken kan worden op basis van de resultaten is dat sommige beoordelaars die een collega zijn van de beoordeelde docenten extreme scores toekennen. Dit geldt voor zowel de scores die de beoordelaars toekennen aan de afzonderlijke videofragmenten van kritische situaties als voor de overallscores die deze beoordelaars toekennen. Beoordelaars die extreem scoren, doen dit zowel extreem streng als extreem mild. Vanuit de literatuur is bekend dat beoordelaars geneigd zijn een positiever oordeel te geven aan de personen die dicht bij de beoordelaar staan (Aronson, Wilson, & Akert, 2007). Dit wordt ook wel het nabijheideffect genoemd. Dit effect zou kunnen verklaren waarom de beoordelaars hun collega's mild beoordelen. Uit de resultaten van dit onderzoek blijkt dat er ook beoordelaars zijn die hun collega's juist strenger beoordelen. Hier is geen duidelijke verklaring voor; wellicht heeft dit te maken met de persoonlijke eigenschappen van de individuele beoordelaars. Er valt overigens geen uitspraak te doen over de validiteit of de gepastheid van de toegekende scores door de beoordelaars die hun collega's hebben beoordeeld. Het zou kunnen zijn dat deze beoordelaars meer valide beoordelingen geven, omdat zij beschikken over meer informatie die relevant is voor het beoordelingskader (Schutz & Moss, 2004). Het kan ook zijn dat deze beoordelaars teveel worden beïnvloed door vooroordelen en verwachtingen die zij hebben ten aanzien van hun collega.

Een tweede conclusie die getrokken kan worden, is dat beoordelaars op een acceptabel niveau de coachcompetentie van docenten scoren en beoordelen, wanneer zij de

schaal met de vier onderscheiden competentieniveaus gebruiken. Er wordt een acceptabele tot hoge overeenstemming gevonden voor het scoren en beoordelen van afzonderlijke videofragmenten uit het dossier (0,71 tot 0,80). Er is over het algemeen zelfs sprake van een hoge overeenstemming voor het toekennen van overallscores aan een videodossier (0,76 tot 0,93). Er wordt een iets lagere, maar nog steeds acceptabele overeenstemming gevonden tussen beoordelaars, wanneer de beoordelaars met extreme toegekende scores worden meegenomen in deze analyses (0,68-0,75 voor de videofragmenten en 0,68-0,85 voor de overallscores). Het verschil in overeenstemming tussen beoordelaars ten aanzien van videofragmenten en overallscores komt overeen met eerder onderzoek naar het beoordelen op basis van een videodossier (Bakker, Beijaard, Roelofs, Tigelaar, Sanders, & Verloop, 2007). Uit dat onderzoek bleek dat beoordelaars het lastig vinden om afzonderlijke videofragmenten van een kritische situatie te beoordelen, omdat deze maar een klein stukje laten zien van wat er allemaal gebeurt tussen leerlingen en de docent. Tevens bleek dat ze na het bekijken van vijf tot zes videofragmenten wel een goed beeld kunnen krijgen van het coachen van de docent.

Een derde conclusie is dat toegekende scores op de schaal met vier competentieniveaus goed te generaliseren zijn over beoordelaars. Uit de resultaten blijkt dat de gemiddelde toegekende score op basis van twee beoordelaars een overeenstemming heeft met de gemiddelde toegekende score over tien beoordelaars van 0,88 tot 0,91. Wanneer de extreme beoordelaars worden meegenomen in deze analyses ligt deze overeenstemming iets lager (0,72 tot 0,90). Deze resultaten houden in dat het in de praktijk haalbaar is om op basis van twee beoordelaars een acceptabele overeenstemming tussen beoordelaars te realiseren voor het beoordelen van de coachcompetentie op basis van een videodossier. Dit is een belangrijke conclusie, omdat het in de praktijk vanwege tijd en kosten niet realistisch is 10 of 12 beoordelaars in te zetten bij een beoordeling.

Op basis van deze drie conclusies mag verondersteld worden dat de genomen ontwerpmaatregelen de eerste gevolgtrekking in de argumentatieketen van validiteitsbepaling (Kane, 2004) ondersteunen. Het ontwikkelde beoordelingskader, de competentieniveaus, de scoringsvoorschriften, de training en de samenstelling van het dossier gaan over het algemeen samen met een betrouwbare scoring door beoordelaars.

5.2 Het generaliseren over videofragmenten

De mate waarin toegekende scores aan videofragmenten generaliseerbaar zijn naar het universum van videofragmenten, de tweede gevolgtrekking van de argumentatieketen, laat een verdeeld beeld zien. De resultaten van de analyses op de toegekende scores aan videofragmenten waarin een docent coacht op vakinhoud, duiden erop dat de scores van deze videofragmenten redelijk te generaliseren zijn naar het universum van videofragmenten. Voor de praktijk betekent dit dat er relatief weinig videofragmenten hoeven te worden opgenomen in een videodossier. De generaliseerbaarheid van toegekende scores aan videofragmenten waarin docenten coachen op reguleren, is voor de docenten één en twee over het algemeen goed en voor de docenten drie en vier over het algemeen acceptabel. Op basis van de resultaten is er geen verklaring te geven waarom de generaliseerbaarheid van de videofragmenten waarin de docent coacht op reguleren voor de docenten één en twee hoger ligt dan voor de docenten drie en vier. Het zou kunnen zijn dat de docenten één en twee op hetzelfde reageren in de situaties waarin ze moeten coachen en dat de docenten drie en vier verschillend van elkaar reageren in die situaties. Het zou ook zo kunnen zijn dat beoordelaars de docenten één en twee wel consistent beoordelen en dit om de een of andere reden niet doen bij docenten drie en vier. De generaliseerbaarheid van videofragmenten waarin docent drie coacht op samenwerken is acceptabel en voor videofragmenten van docent vier acceptabel tot goed. Ook voor deze resultaten is niet te zeggen waardoor het verschil in generaliseerbaarheid is te verklaren. De generaliseerbaarheid van de scores die toegekend zijn aan de videofragmenten waarin de docent coacht op leerhouding, blijkt

problematisch. Een aannemelijke verklaring hiervoor is dat het coachen van de leerhouding heel subtiel gebeurt en vaak verweven is met het coachen op andere leeractiviteiten, waardoor het moeilijk is voor beoordelaars om het coachen op deze leerhouding consistent te beoordelen. In het beoordelingskader dient mogelijk scherper omschreven te worden hoe het coachen op de leerhouding op de vier competentieniveaus er uitziet.

Op basis van dit onderzoek kunnen ten aanzien van de generaliseerbaarheid over assessmenttaken alleen tendensen worden beschreven. Definitieve uitspraken over het minimale aantal videofragmenten dat nodig is om uitspraken te doen over het coachen van de docent op de verschillende leeractiviteiten, kunnen dan ook niet gedaan worden. Het standaardiseren van de videofragmenten op basis van de definitie van een kritische situatie lijkt samen te gaan met positieve resultaten op het gebied van het generaliseren van toegekende scores over videofragmenten. De overeenstemming tussen toegekende scores aan een videofragment en de gemiddelde toegekende scores aan de rest van de videofragmenten is over het algemeen acceptabel tot goed; alleen de generaliseerbaarheid van toegekende scores aan de videofragmenten waarin de docent coacht op leerhouding is problematisch.

5.3 Valide extrapoleren naar de coachperformance in de praktijk

De derde gevolgtrekking in de argumentatieketen voor validiteit, het extrapoleren van de *coachperformance* in de videodossiers naar de *coachperformance* van docenten in de praktijk, is niet onderzocht in deze studie. Met zeer natuurgetrouwe coachsituaties en de gerealiseerde spreiding in de kritische situaties is een voorwaarde vervuld voor de extrapoleerbaarheid van de resultaten naar de *coachperformance* van docenten in de praktijk. Hierbij moet wel de kanttekening worden geplaatst dat het veel tijd kost om voldoende authentieke situaties te verzamelen die genoeg variatie bezitten om alle aspecten van coachen op te roepen, zoals het ondersteunen van verschillende leeractiviteiten. De in dit onderzoek verzamelde videofragmenten van kritische situaties zijn ontstaan in de natuurlijke coachsituatie, waarbij de toevallige coachbehoefte van de leerlingen bepaalt in welk opzicht een situatie kritisch is. Daardoor kon het voorkomen dat er bij een docent veel kritische situaties voorkwamen waarin werd gecoacht op bijvoorbeeld vakinhoud en heel weinig op reguleren. Een gevolg hiervan was dat het lastig was om voor deze docent een videodossier samen te stellen met een representatieve steekproef van kritische situaties voor alle aspecten van coaching. Het bewust oproepen van situaties waarin alle aspecten van coachen worden ontlokt, lijkt psychometrisch gezien aantrekkelijk, maar is wellicht moeilijk haalbaar en ook niet wenselijk. Om vast te stellen welke steekproef van situaties naar aard en aantal een representatieve afspiegeling vormt van criteriumsituaties in de praktijk, zou aanvullend onderzoek in de vorm van bijvoorbeeld een *job*-analyse wenselijk zijn.

5.4 Vervolgonderzoek

In dit onderzoek is nagegaan in hoeverre beoordelaars in staat zijn om overeenstemmend te scoren. Echter, om volledig zicht te krijgen op de validiteit van de assessmentprocedure zal ook moeten worden onderzocht op welke wijze scores tot stand komen. Hierbij gaat het om de vraag in hoeverre beoordelaars scoren en beoordelen op basis van het beoordelingskader dat ze verondersteld worden te gebruiken bij het beoordelen. Om dit te achterhalen, zullen aangevoerde bewijzen en argumenten waarmee beoordelaars de beoordeling onderbouwen nader moeten worden onderzocht op basis van kwalitatieve analyses. Met deze gegevens zou tevens een antwoord kunnen worden gevonden op de vraag of beoordelaars die collega zijn van de beoordeelde docent meer of minder valide beoordelen.

Om hardere uitspraken te doen over het minimale aantal videofragmenten dat in een videodossier zou moeten worden opgenomen, is onderzoek op basis van een groter aantal kritische lessituaties essentieel. Wanneer grotere aantallen geregistreerde videofragmenten van kritische lessituaties betrokken worden in het onderzoek, is het mogelijk om een generaliseerbaarheidstudie uit te voeren op toegekende scores. Op basis van deze analyse is te bepalen hoeveel variantie in de

toegekende scores toegeschreven kan worden aan de videofragmenten van kritische lessituaties en hoeveel variantie aan andere facetten van de assessmentprocedure. Aan de hand van dergelijke gegevens kunnen hardere uitspraken gedaan worden over de generaliseerbaarheid van oordelen, hetgeen een belangrijke stap is in de argumentatieketen voor de bepaling van validiteit.

Noot

1 Dit onderzoek is gefinancierd door NWO/ PROO (projectnummer 411-02-207).

Literatuur

Aronson, E., Wilson, T. D., & Akert, R. M. (2007). *Social psychology* (5th ed.). Amsterdam: Pearson Education Benelux BV.

Bakker, M. E. J., Beijaard, D., Roelofs, E., Tigelaar D., Sanders, P., & Verloop, N. (2007). Video portfolios: *The development and practical utility of an authentic teacher assessment procedure.* Paper gepresenteerd op de VOR divisie conferentie, Utrecht, Nederland.

Boekaerts, M. (1999). Self-regulated learning: Where we are today. *International Journal of Educational Research*, *31*, 445 - 457.

Boekaerts, M., & Simons, P.R.J. (1995). *Leren en instructie. Psychologie van de leerling en het leerproces* (2e druk). Assen, Nederland: Van Gorcum.

Bolhuis, S. (2000). *Naar zelfstandig leren: Wat doen en denken docenten.* Apeldoorn, Nederland: Garant.

Borsboom, D., & Mellenbergh, G.J. (2004). The concept of validity. *Psychological Review, 111,* 1061 - 1071.

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24*, 339 - 353.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical syntheses. *Review of Educational Research*, *65,* 245 - 281.

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L.B. Resnick (Ed), *Knowing, learning and instruction: Essays in honor of Robert Glaser* (pp. 453 - 494). Hillsdale, NJ: Erlbaum.

Crooks, T. J., Kane, M. T., & Cohen, S. A. (1996). Threats to the valid use of assessments. *Assessment in Education: Principles, Policy, & Practice, 3*, 265 - 285.

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.

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

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. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education*, *4*, 289 - 303.

Dwyer, C. A. (1998). Psychometrics of praxis III: Classroom performance assessments. *Journal of Personnel Evaluation in Education, 12*, 163 - 187.

Feldman, J. M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology, 66*, 127 - 148.

Frederiksen, J. R., Sipusic, M., Sherin, M., & Wolfe, E. W. (1998). Video portfolio assessment of teaching. *Educational Assessment*, *5*, 225-298.

Gipps, C. V. (1994). *Beyond testing: Towards a theory of educational assessment.* London, Washington D.C.; Falmer Press.

Haertel, E. H. (1991). New forms of teacher assessment. *Review of Research in Education, 17*, 3 - 19.

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

Johnson, D., & Johnson, R. (1994). *Learning together and alone: cooperative, competitive, and individualistic learning* (4th ed). Boston: Allyn & Bacon.

Kane, M. (2004). Certification testing as an il-

lustration of argument based validation. *Measurement*, *2*, 135 - 170.

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

Klein, S. P., & Stecher, B. M. (1998). Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education, 11*, 121 - 137.

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

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

Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task-specificity. *Educational Measurement: Issues and Practice, 13*(1)*,* 5 - 15.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed.). New York; MacMillan.

Miller, M. D., & Linn, R. L. (2000). Validation of performance assessments. *Applied Psychological Measurement 24*, 367 - 378.

Moerkamp, T., Bruijn, E. de, Kuip, I. van der, Onstenk, J., & Voncken, E. (2000). *Krachtige leeromgevingen in het MBO. Vernieuwingen van het onderwijs in beroepsopleidingen op niveau 3 en 4.* Amsterdam: SCO-Kohnstamm Instituut.

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

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

Onstenk, J. (2000). *Op zoek naar een krachtige beroepsgerichte leeromgeving: fundamenten voor een onderwijsconcept voor de bve-sector*. 's-Hertogenbosch, Nederland: CINOP.

Roelofs, E., & Sanders, P. (2007). Towards a framework for assessing teacher competence. *European Journal for Vocational Training*, *40*(1), 123 - 139.

Ruiz-Primo, M. A., Baxter, G. P., & Shavelson, R. J. (1993). On the stability of performance assessments. *Journal of Educational Measurement, 30*, 41 - 53.

Sanders, P. F. (1998). In W.P. van der Brink en G.J. Mellenbergh (Eds.), *Testleer en testconstructie.* Amsterdam: Boom.

Schaaf, van der, M. F., Stokking, K. M., & Verloop, N. (2005). De invloed van cognitieve representaties van beoordelaars op hun beoordeling van docentportfolio's. *Pedagogische Studiën 82*, 7 - 25.

Schutz, A. M., & Moss, P. A. (2004). Reasonable dicisions in portfolio assessment: Evaluating complex evidence of teaching. *Education Policy Analysis Archives, 12(33).* Geraadpleegd op 7 september 2004, op http://epaa.asu.edu/v12n33/.

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

Shuell, T. J. (1993). Toward an integrated theory of teaching and learning. *Educational Psychologist*, *28*, 291 – 311.

Slavin, R. (1990). *Cooperative learning: theory, research, and practice*. Englewood Cliffs: NJ, Prentice-Hall.

Stamoulis D. T., & Hauenstein, N. M. A. (1993). Rater training and rater accuracy: Training for dimensional accuracy versus training for rater differentiation. *Journal of Applied Psychology, 78*, 994 - 1003.

Vermunt, J. D., & Verloop, N. (1999). Congruence and friction between learning and teaching. *Learning and Instruction, 9,* 257 - 280.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University press.

Woerh, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 64*, 189 - 205.

Zegers, F. E. (1989). Het meten van overeenstemming. *Nederlands Tijdschrift voor de Psychologie, 44*, 145 - 156.

Manuscript aanvaard: 22 mei 2008

Auteurs

Mirjam Bakker is als promovenda verbonden aan het Interfacultair Centrum voor Lerarenopleiding, Onderwijsontwikkeling en Nascholing (ICLON) en het Cito.

Piet Sanders is directeur van het Research Center voor Examinering en Certificering (RCEC), een samenwerkingsverband van Cito en de Universiteit Twente, dat onderzoek doet op het gebied van examinering en certificering.

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

Erik Roelofs is werkzaam als senior toetsdeskundige bij Cito in Arnhem.

Dineke Tigelaar is als universitair docent verbonden aan het Interfacultair Centrum voor Lerarenopleiding, Onderwijsontwikkeling en Nascholing (ICLON), Universiteit Leiden.

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

Correspondentieadres: Mirjam Bakker, ICLON, Universiteit Leiden, Postbus 9555, 2300 RB Leiden, e-mail: *mbakker@iclon.leidenuniv.nl*

Abstract

Reliability and generalizability of competence assessments in video portfolio

Authentic teacher assessments are increasingly developed and used in practice. An important issue in designing authentic performance assessment is in what way the reliability and validity of these assessments can be guaranteed. In the literature, several design principles are discussed that should contribute to more reliable and valid assessments, such as increasing the number of assessors and assessment tasks in the assessment, standardizing assessment tasks, and using high-fidelity tasks in the assessments. However, not much empirical evidence is available that proves that these principles really contribute to reliable and valid assessments. The aim of this research is to find out whether these design principles lead to reliable and valid assessments. Previous to this study, an authentic performance assessment was constructed based on the design principles. The constructed assessment can be used for assessing teachers' coaching competence in the context of senior secondary vocational education. Video episodes of teaching situations in the classroom are the main elements of the constructed assessment procedure. Additional data sources were included that inform about the context of the videotaped teaching situations. This combination of video episodes and context information is called a video portfolio. After the construction of the video portfolios, the validity was determined by answering the following research questions: (a) To which extent do assessors score teachers' coaching competence in a reliable way based on the video portfolios? (b) to which extent are assigned scores to separate video episodes of teachers' coaching performance generalizable to the intended universe of video episodes? In order to answer these research questions, twelve assessors scored four video portfolios. Scorecards were gathered and several analyses were performed. An acceptable to high level of interrater-agreement was found for assigned scores to video episodes and a high level of interrater-agreement was found for assigned overall scores. Furthermore, except for one assessment scale (coaching with regard to students' attitude towards learning) an acceptable to high level of similarity was found between assigned scores to a video episode and the average of the assigned scores to the other video episodes on the assessment scale. The conclusion of this study is that the design principles go together with positive results concerning assessors' scoring as well as the generalizability of assigned scores across video episodes.