

A decorative graphic consisting of three blue circles of varying sizes and two thin blue lines. One line starts from the top left and goes towards the top circle. Another line starts from the top left and goes towards the middle circle. A third line starts from the top right and goes towards the bottom circle. The circles are semi-transparent and overlap slightly.

## **Masterscriptie Onderwijskundig ontwerp & advisering**

De gepercipieerde kwaliteit van een competentie assessment programma voor docenten in het voortgezet onderwijs: een single-case study

Utrecht, december 2011

### **Universiteit Utrecht**

Naam:	R. F. Y. Njoo
Studentnummer:	3369560
1 <sup>e</sup> beoordelaar:	Dr. M. F. van der Schaaf
2 <sup>e</sup> beoordelaar:	Dr. F. J. Prins

### **Abstract**

*Het beoordelen van docenten is zowel vanuit summatief als formatief oogpunt van belang. Enerzijds kunnen consequenties uit competentiebeoordelingen betrekking hebben op contractuele beslissingen. Anderzijds dienen beoordelingen de competentieontwikkeling te stimuleren. Dit maakt de kwaliteit van competentie assessment programma's (CAP's) van cruciaal belang. In dit onderzoek is een single-casestudy uitgevoerd om de volgende onderzoeksvragen te beantwoorden: (1) Wat is de gepercipieerde kwaliteit van het CAP voor docenten in school X van het voortgezet onderwijs volgens docenten en afdelingsleiders? (2) In welke mate beïnvloeden performance assessments, als onderdeel van het CAP van school X, volgens docenten, de totale kwaliteit van het CAP? Ter beantwoording op deze vragen zijn verschillende type data verzameld op een middelbare school in Nederland. Een documentanalyse is uitgevoerd met betrekking tot het competentie-assessmentbeleid. Daarnaast is de perceptie van docenten en afdelingsleiders gemeten aan de hand van een vragenlijst. De data zijn zowel kwantitatief als kwalitatief geanalyseerd en gecombineerd volgens de datatriangulatiemethodiek. Uit de resultaten van de single-casestudy blijkt dat zowel de docenten als de afdelingsleiders niet volledig tevreden zijn met de totale kwaliteit van het CAP. De docenten zijn het meest oneens met de consistentie van het CAP en de wijze waarop het wordt uitgevoerd. Daarentegen zijn zij het meest eens met de betekenisvolheid van het CAP. Ook blijkt op school X dat de totale kwaliteit van het CAP meer dan de helft wordt bepaald door de kwaliteit van performance assessments.*

## 1. Inleiding

Internationaal wordt al jaren gezocht naar adequate methoden en werkwijzen om docenten te beoordelen met als doel de kwaliteit van docenten te waarborgen. Zo zijn er in Angelsaksische landen, zoals Australië en de Verenigde Staten nationale standaarden geformuleerd (Australian College of Educators, 2000; National Board for Professional Teaching Standards, 2002). In Nederland zijn de vereiste bekwaamheden voor onderwijzend personeel vastgelegd in de wet Beroepen In het Onderwijs (wet BIO) die sinds 1 augustus 2006 in werking is gesteld. Het doel van deze wet is een minimumniveau van kwaliteit te garanderen en wordt uitgedrukt in docentcompetenties.

Het begrip competentie is echter door de jaren heen verschillend geïnterpreteerd (Sultana, 2009). Ter verduidelijking wordt in dit onderzoek de volgende definitie gehanteerd: een competentie is een geïntegreerd geheel van kennis, vaardigheden en houding (Van Merriënboer, Van der Klink & Hendriks, 2002). Het meten van een geïntegreerd geheel van de genoemde componenten is nogal ongreepbaar en daardoor een complex proces. Dit geldt dan ook voor het beoordelen van docentcompetenties. Vanwege de complexiteit van het meten van competenties pleiten Tucker, Stronge, Gareis & Beers (2003); Baartman, Bastiaens, Kirschner & Van der Vleuten (2006) voor een combinatie van verschillende assessmentmethoden bijvoorbeeld performance assessment, portfolio assessment en self-assessment. Baartman et al. (2006) definiëren het geheel van gekozen assessments als een competentie assessment programma (CAP).

Een CAP voor docenten dient niet alleen gericht te zijn op het adequaat beoordelen van docenten, maar ook op het stimuleren van verdere professionele ontwikkeling. Het beoordelen van docenten is dus zowel vanuit summatief als formatief oogpunt van belang. Vanuit summatief oogpunt dient er rekenschap gegeven te worden van de kwaliteit van de docent. Zo zijn resultaten uit competentiebeoordelingen op Nederlandse scholen veelal leidend in beslissingen met betrekking tot contractverlenging, benoemingen voor onbepaalde tijd en recent in het kader van de functiemix<sup>1</sup>.

---

<sup>1</sup> De 'functiemix' is de verdeling van docenten over de verschillende salarisschalen. In de huidige situatie starten eerste- en tweedegraads beginnende docenten in het voortgezet onderwijs met een LB-functie dat uit 18 schalen bestaat. In vooraf gaande jaren werden voornamelijk docenten met managementtaken in een LC- of een LD-functie ingeschaald. Het voornaamste doel van de functiemix is juist, ook docenten die geen managementtaken hebben maar toch goed presteren, te belonen door hen in een hogere functie in te delen (Ministerie van OC&W, 2007).

Dit maakt de kwaliteit van beoordelen van cruciaal belang. Anderzijds is ook het formatieve doel van beoordelen van belang dat gericht is op het stimuleren en begeleiden van het leer- en ontwikkelproces van de leraar. Dit vertaalt zich concreet in het geven van feedback of het stimuleren van reflectie (Black & William, 1998). Het is daarom van belang dat bij het ontwerpen en het evalueren van de kwaliteit van een CAP, criteria gehanteerd worden die gericht zijn op zowel competentiebeoordeling als op competentieontwikkeling.

In Nederland is de wijze waarop docenten in het voortgezet onderwijs worden beoordeeld een interne aangelegenheid van de school. De school dient zelf te bepalen welke vormen van assessments opgenomen worden in een CAP en welke niet. In de Verenigde Staten is de beoordeling van docenten een nationale aangelegenheid. De vakinhoudelijke, pedagogische en didactische kennis en vaardigheden van docenten worden daar beoordeeld in de vorm van gestandaardiseerde traditionele schriftelijke toetsen de 'paper-and pencil tests'. Critici stellen echter dat deze toetsen, die basis docentcompetenties moeten meten, overgewaardeerd zijn en dat zij hoofdzakelijk gericht zijn op de summatieve assessment functie. Toch hoeft een traditionele toets niet uitgesloten te worden. In een CAP zou een dergelijke toets gecombineerd kunnen worden met nieuwere vormen van assessments zo stellen Baartman et al. (2006). Een belangrijke vraag blijft hoe beoordeeld wordt, of een CAP kwalitatief adequaat is ontworpen zodat tegemoet wordt gekomen aan zowel de summatieve als de formatieve functie.

### ***Paradigma verschuiving interpretatie kwaliteitscriteria***

De interpretatie van kwaliteit in de context van beoordelen is in de loop der jaren veranderd door een paradigmaverschuiving van een testcultuur naar een assessmentcultuur (Gipps, 1994). De traditionele interpretatie van kwaliteit wordt vertegenwoordigd door de psychometrische kwaliteitscriteria, betrouwbaarheid en validiteit. Echter, nieuwe vormen van assessments vragen een andere operationalisatie van deze traditionele kwaliteitscriteria. Waar in een testcultuur traditionele toetsen zoals multiple-choice toetsen of paper-and-pencil toetsen centraal staan, staan in een assessmentcultuur nieuwe vormen van beoordelen centraal. De mate waarin een toets als valide en betrouwbaar wordt beschouwd, doet in traditionele zin uitspraak over de mate van de kwaliteit van een toets. Baartman en collega's (2006) stellen de vraag of deze traditionele interpretatie van betrouwbaarheid en validiteit van toepassing is op nieuwere vormen van assessments. Door een literatuurstudie gebaseerd op het werk van verschillende wetenschappers (Dunbar, Koretz, & Hoover, 1991; Benett, 1993; Kane, 1992; Cronbach, Linn, Brennan, & Haetel, 1997; Jonsson, Baartman, Lennung, 2009) komen zij tot de conclusie dat de traditionele interpretatie van betrouwbaarheid niet van toepassing is op nieuwere vormen van assessments omdat psychometrie en onderwijs niet dezelfde basisassumpties delen.

### ***Betrouwbaarheid***

Een betrouwbaar instrument houdt in traditionele zin in, dat scores bij herhaalde metingen in tijd met elkaar overeen dienen te komen (stabiliteit) en dat de verschillende items onderling hetzelfde construct meten (construct representatie). Dit is bijvoorbeeld van belang bij psychologische testen bestaande uit veel verschillende kleine items. In het onderwijs is het vrij complex om stabiliteit van toetsresultaten na te streven omdat er sprake zou kunnen zijn van een leereffect. Wat overigens geheel in overeenstemming is met gestelde onderwijsdoelen. Daarnaast compliceren kenmerken van nieuwe vormen van assessments het kwalificatieproces (Cronbach et al., 1997). Bij performance assessments bijvoorbeeld gaat het veelal om het uitvoeren van een complexe taak in zijn geheel. Het vaststellen van de construct representatie wordt hierdoor een zeer complex proces. Dat de traditionele interpretatie van betrouwbaarheid niet toepasbaar is op nieuwere vormen van assessments, betekent niet dat deze niet betrouwbaar hoeven te zijn. Het veronderstelt wel een herinterpretatie van betrouwbaarheid (Gipps, 1994). Met het oog op summatieve consequenties is het juist van belang dat beoordelingen betrouwbaar zijn. Dit geldt overigen ook voor formatieve

functies van beoordelingen; immers wat heeft een docent aan bepaalde feedback als deze gebaseerd is op een onbetrouwbaar oordeel?

### *Validiteit*

De conventionele opvatting geeft validiteit weer als een eigenschap van een meetinstrument en beschrijft vier typen: 1) predictieve validiteit geeft de mate weer waarin een assessment toekomstige performance meet. Hierbij zou de vraag bijvoorbeeld gesteld kunnen worden of een hoge score op een toelatingsexamen ook voorspelt hoe goed een student het doet tijdens zijn studie 2) inhoudsvaliditeit heeft te maken met de vraag of de items het te meten (hypothetische) constructdomein dekken, bijvoorbeeld dekt de toets de leerstof? 3) concurrente validiteit geeft inzicht in (cor)relatie tussen metingen van twee toetsen die hetzelfde (hypothetische) construct beogen te meten en 4) construct validiteit stelt de vraag of het assessment het (hypothetische) construct meet dat beoogd wordt te meten; met andere woorden, verbindt de meting de theorie over het (hypothetisch) construct? Een voorbeeld hiervan is de vraag, hoe men onderzoeksvaardigheden meet. Dit houdt in dat er in de eerste plaats een adequate theoretische uiteenzetting noodzakelijk is. Vervolgens kunnen er constructen gemaakt worden die geoperationaliseerd worden in een taak. De taak die wordt gemeten, geeft inzicht in de mate van beheersing van de onderzoeksvaardigheden. In deze conventionele benadering is het van belang te beschikken over een valide meetinstrument.

Messick (1989) & Kane (1992), geïnspireerd door het gedachtegoed van Cronbach (1989), onderstrepen meer het belang van de betekenis-toekenning aan de resultaten van een beoordeling. Zij stellen dat validiteit een eigenschap is van de betekenis-toekenning aan resultaten die een meetinstrument genereert. Kortom, wat zeggen de resultaten uit bijvoorbeeld een performance assessment? Messick (1989) noemt dit consequentiële validiteit. In aansluiting hierop introduceert Kane (1992) the *'argument-based approach'*, ook een benadering van validiteit met betrekking tot de interpretatie van resultaten uit beoordelingen. Het aanreiken van empirische steun (argumenten) voor een bepaalde interpretatie van een beoordeling staat hierin centraal. In deze benadering wordt gesteld dat een valide interpretatie van een beoordeling onder andere af hangt van de plausibiliteit van de argumenten waarop een interpretatie is gebaseerd. De plausibiliteit hangt vervolgens af van de kwaliteit van de argumenten. Een argument kan op zich inherent plausibel zijn of ondersteund worden door verschillende typen bewijzen. Dit kunnen bijvoorbeeld ervaringen zijn of resultaten uit observaties, documentanalyses en enquêtes. Daarnaast kunnen de verschillende type bewijzen bestempeld worden als zwak, bijvoorbeeld persoonlijke eigen ervaring, en sterk, resultaten uit verschillende enquêtes (Clark & Sampson, 2005). Hierbij geldt dat de sterke en zwakke kanten van het bewijs telkens in overweging genomen moeten worden bij validiteitsargumenten. Daarnaast is ook het verzamelen van procedureel bewijs dat betrekking heeft op bijvoorbeeld een beoordelingsprocedure, intern bewijs wat inzicht geeft in de consistentie van het bewijs en extern bewijs dat betrekking heeft op de vergelijkbaarheid met andere beoordelingsmethoden en bronnen. Een methode om verschillende typen bewijzen te verzamelen met het ondersteunen van validiteitsargumenten als doel is het *'chain-model'* (Crooks, Kane & Cohen, 1996). Het *'chain-model'* bestaat uit acht stadia. Het eerste stadium concentreert zich op de procedurele afname van scores waarbij de focus met name gelegd wordt op de omstandigheden waarin een assessment plaatsvindt. Het tweede stadium heeft betrekking op de wijze waarop prestatie van een bepaalde taak wordt gescoord. Het derde op het genereren van een totaal score uit prestaties op individuele taken. Generalisatie van een specifieke taak staat centraal in stadium vier. Het vijfde stadium heeft betrekking op het schatten of voorspellen van prestaties uit het beoordeelde domein naar het doeldomein. Hierbij is normaalgesproken het beoordeelde domein een aftreksel van het doeldomein of authentieke situatie. De set taken zijn random gekozen en kunnen gezien worden als een steekproef van het beoordeelde domein. Per definitie omvat dit niet alle taken waardoor er altijd sprake is van bias. Stadium zes richt zich op het vormen van een beoordeling op basis van de

verzameling gegevens. De beslissingen die genomen moeten worden als gevolg van de beoordeling worden in stadium zeven geformuleerd. Deze kunnen formatieve gevolgen hebben maar ook summatieve. In het laatste stadium staat de impact als gevolg van het assessment proces, de interpretaties en besluiten op de beoordeelde en andere betrokkenen centraal. De acht stadia kunnen worden gezien als acht schakels van een ketting waarbij de zwakte van (alleen) één schakel, de ketting als geheel zwak maakt. Het 'chain-model' suggereert dat validiteit gelimiteerd is tot de zwakste schakel en dat daardoor, pogingen om andere links sterker te maken ter compensatie, zinloos is (Crooks, Kane & Cohen, 1996).

### ***Psychometrische en Edumetrische kwaliteitscriteria***

Als gevolg van de genoemde paradigmaverschuiving (van een testcultuur naar een assessmentcultuur) zijn er kwaliteitscriteria toegevoegd aan de bestaande psychometrische kwaliteitscriteria: *herhaalbaarheid, vergelijkbaarheid, rechtvaardigheid en cognitieve complexiteit*. De toegevoegde edumetrische kwaliteitscriteria zijn respectievelijk: *acceptatie, transparantie, authenticiteit, rechtvaardigheid, betekenisvolheid, onderwijsgevolgen en zelfsturend leren*. In tabel 1 worden deze twaalf kwaliteitscriteria verder toegelicht.

De achterliggende gedachte van dit onderzoek is geheel in overeenstemming met het principe van het chain-model van Crooks, Kane & Cohen (1996); de kwaliteit van een CAP is zo sterk als de zwakste schakel. Een kwaliteitscriterium dat niet voldoende wordt vertegenwoordigd in een CAP, kan niet worden gecompenseerd door een hoge score op een ander kwaliteitscriterium. Als blijkt dat een CAP niet rechtvaardig is, kan dat niet worden gecompenseerd met een hoge mate van betekenisvolheid. Volgens Baartman et al. (2006) geldt dit niet voor een afzonderlijk assessment. Zij stellen dat een kwaliteitscriterium in meer of mindere mate vertegenwoordigd kan zijn in een afzonderlijk assessment. Een traditionele "paper and pencil toets" en een performance assessment kunnen beide opgenomen worden in een CAP. De combinatie van deze twee beoordelingsmethoden zou als doel kunnen hebben om vaardigheden met bijbehorende onderliggende kennis te toetsen. Een traditionele toets is minder authentiek in tegenstelling tot een performance assessment. In dit geval compenseren de twee beoordelingsmethoden elkaar voor het kwaliteitscriterium authenticiteit. In dit onderzoek wordt deze opvatting gedeeld. Tegelijkertijd wordt er in dit onderzoek benadrukt dat de totale kwaliteit van een afzonderlijk assessment wel voldoende moet zijn. De veronderstelling is namelijk dat de totale kwaliteit van afzonderlijke assessments van invloed is op de totale kwaliteit van een CAP. Bovendien zou men zich af kunnen vragen wat de meerwaarde is van een laag gekwalificeerd afzonderlijk assessment.

Dit onderzoek heeft zich gericht op het meten van de kwaliteit van CAP's voor docenten in het voortgezet onderwijs, zodat indien nodig verbeteringen aangebracht kunnen worden. In een single-casestudy is de kwaliteit van een CAP gemeten aan de hand van percepties van deelnemers. Perceptie is een indicatie voor kwaliteit. Daarnaast gaat er ook een sturende werking uit naar het denken van de verschillende betrokkenen op het moment dat er naar perceptie wordt gevraagd. Investeren in het denken over assessments en assessmentbeleid is cruciaal voor het welslagen van een dergelijk beleid (Brown, 2004). Anderzijds heeft dit onderzoek zich gericht op het verkennen van een methodiek om de gepercipieerde kwaliteit en invloed van afzonderlijke assessments ten opzichte van een CAP vast te stellen. Hypothetisch kunnen alle afzonderlijke assessments evenredig veel aandeel hebben in de totale kwaliteit van een CAP. Het is echter ook mogelijk dat de totale kwaliteit bepaald wordt door één of twee assessmentmethoden en doen anderen er minder toe. In aansluiting hierop zou overwogen kunnen worden om assessments achterwege te laten of de kwaliteit daarvan te verbeteren zodat dit ten goede komt aan de totale kwaliteit van een CAP. Echter, in dit onderzoek zijn in plaats van alle afzonderlijke assessmentmethoden, alleen performance assessments als onderdeel van het CAP onderzocht. Deze assessmentmethode wordt in de praktijk veelvuldig

gebruikt en is hiermee exemplarisch voor andere afzonderlijke assessments. De centrale vragen in dit onderzoek zijn:

*(1) Wat is de gepercipieerde kwaliteit van het CAP voor docenten in school X van het voortgezet onderwijs volgens docenten en afdelingsleiders?*

*(2) In welke mate beïnvloeden performance assessments als onderdeel van het CAP voor docenten op school X, de totale kwaliteit van het CAP?*

De gepercipieerde kwaliteit van het CAP en performance assessments is gemeten middels het raamwerk van kwaliteitscriteria van Baartman et al. (2006). Dit raamwerk is gekozen omdat het criteria bevat die voor zowel summatieve als formatieve doeleinden gebruikt kunnen worden. Daarnaast zijn de kwaliteitscriteria gericht op competentiegericht beoordelen waarbij het gedachtegoed van Cizek (1997, p.10) zoals geciteerd in Baartman, Bastiaens, Kirschner, & Van der Vleuten, (2007) als uitgangspunt is genomen: *(1) competentiegericht beoordelen is een planmatig proces waarin relevante informatie wordt verzameld en gesynthetiseerd met betrekking tot de volgende doelen (a) ontdekken en documenteren van sterke en zwakke punten van de docent (b) het plannen en verbeteren van instructie (training en bijscholing) (c) voortgang evalueren en beslissingen maken over de docent (2) competentiegericht beoordelen is een planmatig proces waarin relevante informatie wordt verzameld en gesynthetiseerd met betrekking tot het proces, instrument of methode om informatie te verzamelen.* Ten slotte is dit raamwerk ontleend aan een grondige theoretische analyse gebaseerd op ideeën over validiteit en betrouwbaarheid met betrekking tot nieuwe vormen van assessment (Linn, Baker & Dunbar, 1991).

Tabel 1. Korte omschrijving van de twaalf kwaliteitscriteria voor CAP's (Baartman et al., 2006)

Criteria	Korte omschrijving
Doelen	Dit kwaliteitscriterium heeft betrekking op de koppeling tussen standaarden, curriculum, instructie en assessment. De beoordelingsdoelen en beoordelingsmethoden dienen in dezelfde lijn te liggen als de professionele doelen.
Herhaalbaarheid	Dit kwaliteitscriterium geeft aan dat beslissingen op basis van de resultaten van het CAP accuraat en constant dienen te zijn in verschillende situaties en tussen verschillende beoordelaars. Een beslissing zou niet mogen afhangen van de beoordelaar of een specifieke assessmentsituatie.
Transparantie	Een CAP dient duidelijk en begrijpbaar voor alle betrokkenen te zijn (docenten, leidinggevende, beoordelaars). Externe onderwijsorganen moeten een duidelijk beeld kunnen vormen van de wijze waarop een CAP is ontwikkeld en wordt uitgevoerd.
Acceptatie	Alle betrokkenen (docenten, leidinggevend, beoordelaars) dienen zich te kunnen vinden in de beoordelingscriteria en de wijze waarop een CAP wordt uitgevoerd. Zij dienen vertrouwen te hebben in de kwaliteit van een CAP.
Vergelijkbaarheid	Een CAP zou consistent en op een verantwoordelijke wijze uitgevoerd moeten worden. De omstandigheden waaronder een assessment wordt uitgevoerd, zou zo veel mogelijk, voor alle docenten hetzelfde moeten zijn. Daarnaast dient het beoordelen op een consistente wijze plaats te vinden waarbij dezelfde criteria gelden voor alle docenten.
Rechtvaardigheid	Docenten dienen een eerlijke kans te krijgen om hun competenties te demonstreren bijvoorbeeld door zichzelf op verschillende manieren uit te drukken. Dit betekent ook het uitsluiten van biasen onder beoordelaars.

---

Zelfsturend leren	Een CAP zou zelfregulerend leren van docenten moeten stimuleren. Een CAP zou specifieke methoden moeten bevatten die zelfregulerend leren stimuleren zoals zelf-assessment en feedback ontvangen.
Betekenisvolheid	Dit kwaliteitscriterium houdt in dat een CAP een significante waarde heeft voor alle betrokkenen (docenten, beoordelaars). Een manier om de betekenisvolheid te verhogen is door docenten te betrekken bij (de ontwikkeling van) het assessmentproces.
Cognitieve complexiteit	Dit kwaliteitscriterium heeft raakvlak met het kwaliteitscriterium authenticiteit en houdt in dat een CAP de benodigde cognitieve vaardigheden zou moeten vertegenwoordigen die nodig zijn voor het uitoefenen van het beroep. Uit de beoordelingsmethoden zou ook het denkproces achter de cognitieve vaardigheden zichtbaar gemaakt moeten worden.
Authenticiteit	Dit kwaliteitscriterium verwijst naar de mate waarin een CAP overeenkomt met de beroepspraktijk. Met dit kwaliteitscriterium wordt onderstreept dat alleen relevante competenties van het beroep beoordeeld dienen te worden. Authenticiteit kan vertegenwoordigd zijn in vijf dimensies: de beoordelingsopdracht, de fysieke context, de sociale context, een beoordelingsformulier en de assessmentcriteria.
Professionele gevolgen	Dit kwaliteitscriterium wordt veel ondersteund door verschillende auteurs en heeft betrekking op het effect van een CAP op de uitoefening van het beroep. In welke mate heeft een CAP positieve gevolgen en worden negatieve effecten als gevolg van het CAP geminimaliseerd.
Tijd en Kosten	Dit kwaliteitscriterium onderstreept de kosten en de baten van een CAP. Vergeleken bij het afnemen van traditionele toetsen kosten nieuwere vormen van assessments meer tijd en investering. De kosten dienen tegen de baten worden afgenomen. In dit onderzoek zijn docenten niet betrokken bij dit proces.

---

## 2. METHODE

### *Design*

Dit onderzoek is te typeren als een representatieve single-case study. In de context van dit onderzoek betekent dit, dat het competentie-assessmentbeleid en -procedure representatief is voor andere middelbare scholen in Nederland. Het onderzoek is uitgevoerd in twee fases van dataverzameling. In de eerste fase zijn documenten verzameld en geanalyseerd met betrekking tot het competentie-assessmentbeleid en -procedure van school X. Een gedeelte van de verzamelde documentatie is tevens als input gebruikt voor de tweede fase van dataverzameling. Er zijn bijvoorbeeld observatieformulieren en uitwerkingen van docentcompetenties gebruikt als bijlagen bij de vragenlijsten. In de tweede fase is een gecombineerde kwantitatieve en kwalitatieve survey uitgevoerd. 29 docenten van school X hebben twee vragenlijsten ingevuld. Eén waarin de gepercipieerde kwaliteit van het CAP werd gemeten en twee de performance assessments. Bij twee van de vijf afdelingsleiders is ook de gepercipieerde kwaliteit van het CAP gemeten door middel van dezelfde vragenlijst. Het kwantitatieve gedeelte bestond uit schaalscores en het kwalitatieve gedeelte uit schriftelijke motivaties bij de schaalscores en feedback over de duidelijkheid en begrijpelijkheid van de vragen in de vragenlijst.

### **Selectie casus**

De representativiteit van de geselecteerde casus is beoordeeld door middel van criteria die gebaseerd zijn op een evaluatierapport van het Landelijk Platform Beroepen in het Onderwijs (2010). Hieruit blijkt dat scholen in het voortgezet onderwijs in Nederland algemeen genomen wel beschikken over een beoordelingsbeleid en –procedure gebaseerd op docentcompetenties maar dat het werken met docentcompetenties nog niet ligt verankerd in de school. Met name de ontwikkeling en beoordeling, de wijze van invoering en de daadwerkelijke uitvoering van onderdelen van het personeelsbeleid. Op 20 scholen in het voortgezet onderwijs zijn schoolleiders eerst schriftelijk en vervolgens telefonisch benaderd. De selectiecriteria gebaseerd op het evaluatierapport van het Landelijk Platform Beroepen in het Onderwijs (2010) zijn geoperationaliseerd in een aantal vragen en zijn vervolgens gesteld in een telefonisch interview met de schoolleider. Op deze wijze is vastgesteld of een school in aanmerking kwam voor deelname. Op zeven van de 20 aangeschreven scholen is het tot een telefonisch interview gekomen. Slechts drie van de zeven scholen kwamen in aanmerking voor deelname. Eén school (school X) was bereid daadwerkelijk deel te nemen aan dit onderzoek. School X is een scholengemeenschap voor MAVO, HAVO en VWO en bevindt zich in de beginfase van het beoordelen van docenten door middel van docentcompetenties. Dit houdt in dat men is begonnen met het documenteren van ontwikkel- en beoordelingsprocedures, dat de schoolleiding begonnen is met het beoordelen van een aantal docenten volgens deze procedures en dat er een aantal ontwikkelgesprekken op basis van docentcompetenties heeft plaatsgevonden.

### **Deelnemers**

Van de 57 docenten hebben 29 docenten van verschillende vakken deelgenomen aan dit onderzoek. De docenten geven les op verschillende jaarlagen en aan alle niveaus. Alle afdelingsleiders, in totaal vijf, zijn benaderd voor deelname. Slechts twee hebben daadwerkelijk deelgenomen aan dit onderzoek.

### **Instrumenten**

#### *Vragenlijst gepercipieerde kwaliteit CAP*

Het raamwerk van kwaliteitscriteria van Baartman et al. (2006) dat oorspronkelijk is bedoeld om CAP's voor studenten in het MBO te evalueren, is gebruikt als meetinstrument. Dit meetinstrument is als betrouwbaar te kwalificeren. Bevindingen uit onderzoek van Baartman et al. (2006) laten na hercalculatie, voor elk kwaliteitscriterium een Cronbach's Alpha waarde zien boven de .60. Behalve voor het kwaliteitscriterium herhaalbaarheid (.59) en transparantie (.58). Het raamwerk is verwerkt in een vragenlijst en aangepast aan de context van dit onderzoek. De psychometrische en edumetrische kwaliteitscriteria zijn geoperationaliseerd in vijf of zes items (zie tabel 3). De items zijn verwerkt in stellingen die steeds betrekking hebben op drie dimensies, te weten: inhoud, actoren en instrumenten. De inhoud heeft betrekking op de invulling die gegeven wordt aan een kwaliteitscriterium en de kwaliteit van beoordelingscriteria & normering die hierbij gelden. Een voorbeeld van de operationalisatie van twee items binnen het kwaliteitscriterium zelfsturend leren met betrekking tot de inhoud: "Het CAP stimuleert tot reflectie op de eigen ontwikkeling" en "de beoordelingscriteria zijn (ook) gericht op de ontwikkeling van zelfsturend leren". De actoren binnen een kwaliteitscriterium zijn de docenten (de beoordeelde) en de leidinggevenden (de beoordelaars). Een voorbeeld hiervan zijn de volgende items: "de docenten beoordelen zichzelf en elkaar" en "de beoordelaar geeft geleidelijk het initiatief bij de beoordeling over aan de student". De laatste dimensie instrumenten heeft betrekking op de kwaliteit van beoordelingsopdrachten, de context en de uitvoering en procedure van een kwaliteitscriterium. Op een vijfpuntsschaal geven docenten of leidinggevenden aan in hoeverre zij het eens of oneens zijn met de stelling. De schaal



varieert van helemaal oneens tot helemaal eens, de optie ik weet het niet en niet van toepassing. Na elk kwaliteitscriterium volgen twee open ruimtes. Ruimte één is bedoeld om antwoorden te motiveren aan de hand van eigen ervaringen met het CAP. Ruimte twee is bedoeld om feedback te geven op de duidelijkheid en begrijpelijkheid van de vragen.

#### *Vragenlijst gepercipieerde kwaliteit performance assessment*

Om de gepercipieerde kwaliteit van performance assessments te meten is gebruik gemaakt van dezelfde vragenlijst van Baartman et al. (2006). De context van de stellingen is echter aangepast zodat de gepercipieerde kwaliteit van performance assessments wordt gemeten. Bijvoorbeeld: docenten weten en begrijpen welke competenties worden beoordeeld tijdens een lesobservatie. In deze vragenlijst is alleen ruimte om antwoorden te motiveren aan de hand van eigen ervaring met performance assessments. Er is geen ruimte om feedback te geven over de duidelijkheid en begrijpelijkheid van de vragen.

#### *Pilotonderzoek*

Een pilot-onderzoek heeft plaatsgevonden om de toepasbaarheid en kwaliteit van de vragenlijsten te exploreren en waar mogelijk te verbeteren. Zes docenten hebben drie vragenlijsten ingevuld, één om de kwaliteit van het CAP, één om performance assessments en één om leerlingenquêtes te meten. De docenten zijn geselecteerd op basis van vrijwilligheid. De docenten gaven aan dat het invullen van de drie vragenlijsten over de kwaliteit van respectievelijk, het CAP, performance assessments en leerlingenquêtes teveel tijd vergt van docenten (een vragenlijst bestaat ongeveer uit 65 stellingen). Ook het invullen van twee vragenlijsten wordt als veel werk beschouwd. De docenten bleken wel bereidwillig om alle vragen te beantwoorden. Zij hadden echter moeite om de vragen correct te interpreteren of te begrijpen. Veel vragen zijn niet beantwoord waardoor het uitvoeren van een betrouwbaarheid- en een factoranalyse niet mogelijk was. De feedback van de docenten is meegenomen en verwerkt in een herziene versie vragenlijsten. Daar waar stellingen onduidelijk waren, is gekozen om voorbeelden te geven om het originele instrument zoveel mogelijk intact te laten. Met betrekking tot het criterium authenticiteit is gebleken dat drie van de vijf indicatoren telkens als niet van toepassing zijn gemarkeerd: “De beoordelingsopdrachten lijken op taken uit het toekomstig beroep”, “de fysieke en sociale werkomstandigheden lijken op die in het toekomstig beroep”. Als argument werd gegeven dat er geen sprake was van een toekomstige beroep omdat zij al werkzaam zijn als docent: de situatie is dus al authentiek. Om deze reden en vanwege theoretische onderbouwing is gekozen om dit criterium niet mee te nemen in dit onderzoek. In de vragenlijst die de gepercipieerde kwaliteit van performance assessments meet, is het criterium Tijd & Kosten achterwege gelaten. Docenten gaven als feedback dat dit criterium in deze context niet van toepassing is. Dit kwaliteitscriterium heeft meer betrekking op het gehele CAP in plaats van alleen performance assessments.

#### *Procedure*

##### *Documentanalyse*

Het doel van de documentanalyse was gericht op het verkrijgen van inzicht in het competentie assessmentbeleid dat is opgesteld en wordt uitgevoerd door het management team van de school. Aan de schoolleider is gevraagd welke documentatie beschikbaar is met betrekking tot het beoordelingsbeleid en –procedure. Vervolgens is aan elke afdelingsleiders gevraagd welke documenten zij hanteren bij competentiebeoordeling en –ontwikkeling van docenten en hoe de procedures hiervan zijn. Bijvoorbeeld protocollen waarin de competentie ontwikkelingsprocedure staat omschreven of documenten met beoordelingscriteria die worden gebruikt bij performance assessments.

Als volgende stap is de verzamelde documentatie tegen het licht gehouden van de volgende kenmerken van goed assessmentbeleid (Ketelaar, 2008); *verbinding competentie assessmentbeleid met strategie school, draagvlak creëren voor competentie assessment, duidelijke omschrijving competenties, communicatie in de organisatie met betrekking tot competentie assessmentbeleid, integratie van competentie management in de praktijk van het personeelsbeleid, mix van beoordelingsactiviteiten met bijbehorende beoordelingscriteria*. De kenmerken zijn deels gebaseerd op nationale regelgeving op dit gebied en zijn hierdoor relevant voor de context van dit onderzoek.

#### Kwantitatieve en Kwalitatieve survey

In de tweede fase is de gepercipieerde kwaliteit van het CAP en performance assessments onderzocht. De docenten hebben een enveloppe ontvangen met daarin: een begeleidende brief waarin kort het onderzoek is uitgelegd en waarin de contextvariabelen van docenten is geoperationaliseerd. Daarnaast is door middel van drie items ook gecontroleerd op perceptie van docenten ten aanzien van docentcompetenties. Hoe docenten tegen docentcompetenties aankijken zou van invloed kunnen zijn op de wijze waarop de vragenlijsten worden ingevuld (Brown, 2004). Indien men over het algemeen negatief tegenover docentcompetenties staat, zou dat zijn weerslag kunnen hebben op het eindresultaat van de vragenlijst (kwaliteitscore van een CAP). Daarnaast is in de enveloppe een overzicht van de vormgeving van het CAP, twee vragenlijsten (kwaliteit CAP en performance assessment) en een retourenveloppe toegevoegd. De afdelingsleiders hebben hetzelfde pakket ontvangen minus de vragenlijst over de kwaliteit van performance assessments. De contextvariabelen zijn voor hen gemarkeerd als niet van toepassing. Zowel de docenten en de afdelingsleiders hebben een maand de tijd gehad om de vragenlijsten in te vullen en te retourneren aan het secretariaat van de school.

#### Analyse

Om de eerste onderzoeksvraag te beantwoorden met betrekking tot de gepercipieerde kwaliteit van het CAP van school X, zijn de verschillende typen data gecombineerd volgens de datatriangulatiemethodiek. Met als doel om vanuit verschillende invalshoeken antwoord te krijgen op de onderzoeksvraag. De schriftelijke motivaties van docenten en afdelingsleiders zijn per kwaliteitscriterium gecategoriseerd om de kwantitatieve schaalscores adequaat te interpreteren. Vervolgens is bekeken in hoeverre motivaties een gemiddelde schaalscore van een kwaliteitscriterium ondersteunen of juist niet. De resultaten uit de documentanalyse zijn gecombineerd met zowel de kwantitatieve schaalscores als met de kwalitatieve schriftelijke motivaties. Ten eerste is nagegaan welk kwaliteitscriterium betrekking heeft op een van de kenmerken van goed assessmentbeleid. Vervolgens is nagegaan in hoeverre de resultaten uit de documentanalyse, de gemiddelde schaalscores per kwaliteitscriterium wel of niet ondersteunen of juist een ander invalshoek belichten. Door de verschillende type data op deze wijze te combineren worden enerzijds beleidsmatige aspecten op het gebied van competentiebeoordeling en ontwikkeling belicht. Met het oog op het validiteitsaspect bevordert dit anderzijds de betekenisgeving aan de resultaten van de verschillende typen data. (Cronbach, 1989; Messick, 1989; Kane, 1992).

#### **FASE I**

##### *Documentanalyse*

De volgende documentatie met betrekking tot het competentie assessmentbeleid zijn verzameld: (a) benoeming, begeleiding en beoordeling van nieuwe docenten. Hierin zijn de taken en verantwoordelijkheden van de rector, schoolopleider en sectiebegeleider omschreven (b) functioneringsreglement voor zittende docenten (c) overzicht van de SLB docentcompetenties met een korte algemene omschrijving (d) overzicht van docentcompetenties die voortvloeien uit de

missie van de school (e) een observatieformulier zoals gehanteerd wordt door de onderwijsinspectie (f) een observatieformulier: kijkwijzer voor de “ideale les” (g) een observatieformulier met beoordelingscriteria en (h) een lesverslag met beoordelingscriteria.

De documentanalyse is uitgevoerd op basis van onderstaande kenmerken van goed assessmentbeleid dat gebaseerd is op een cyclus van vijf stappen van goed competentie management beleid voor scholen in het voortgezet onderwijs (Ketelaar, 2008). De stappen zijn daarnaast onderbouwd en aangevuld door het werk van verschillende wetenschappers op dit terrein.

#### *Verbinding competentie assessmentbeleid met strategie school:*

Uit documentatie zou het doel van het competentie assessmentbeleid moeten blijken en hoe dit vervolgens is verbonden met de strategische doelen van de school.

#### *Draagvlak creëren voor competentie assessment*

Een belangrijk kenmerk voor goed competentie assessmentbeleid is het creëren van draagvlak onder betrokkenen voor competentie assessment. Dit geldt voor zowel de beoordelaars als de beoordeelden.

#### *Duidelijke omschrijving competenties*

Ketelaar (2008) noemt het opstellen van competentieprofielen als onderdeel van goed assessmentbeleid: “Een competentieprofiel is een verzameling van competenties die relevant zijn voor het succesvol optreden in een functie” (Ketelaar, 2008, p. 89). Darling-Hammond (2001) stelt dat het ontwikkelen van een duidelijke en geloofwaardige raamwerk van standaarden essentieel is om de professionalisering van docenten te stimuleren.

#### *Communicatie in de organisatie*

Van belang is dat er binnen de organisatie over het doel, de inhoud en procedure van competentie assessment gecommuniceerd wordt.

#### *Integratie van competentie management in de praktijk van het personeelsbeleid*

In het personeelsbeleid dient de rol van competenties vastgesteld te zijn. Dit betreft bijvoorbeeld de afstemming in het beleid ten aanzien van werving en selectie, functioneren en beoordelen, opleiden in ontwikkelen of dossiervorming.

#### *Mix van beoordelingsactiviteiten met bijbehorende beoordelingscriteria*

Goed assessmentbeleid is gebaseerd op het aanbieden van een mix van verschillende beoordelingsactiviteiten met bijbehorende beoordelingscriteria.

## **FASE II**

### *Kwantitatieve analyse vragenlijst gepercipieerde kwaliteit CAP*

Een confirmerende factoranalyse is uitgevoerd om de dimensionaliteit van de subschalen uit te wijzen. De kwaliteit van een CAP is immers geoperationaliseerd in psychometrische en edumetrische kwaliteitscriteria waardoor er sprake is van verschillende subschalen. Vervolgens is er een betrouwbaarheidsanalyse uitgevoerd om de betrouwbaarheid van het meetinstrument te bepalen. Om de totale score van de gepercipieerde kwaliteit van het CAP adequaat te kunnen interpreteren zijn gemiddelde schaalscores en standaarddeviaties per kwaliteitscriterium berekend. De vierpuntsschaal is zodanig gecodeerd dat een hoge gemiddelde score (vier) verwijst naar een hoge gepercipieerde kwaliteit van het desbetreffende kwaliteitscriterium en een lage gemiddelde score (één) verwijst naar een lage gepercipieerde kwaliteit. De antwoordcode: helemaal mee oneens heeft waarde één, mee oneens heeft waarde twee, mee eens heeft waarde drie, helemaal mee eens heeft waarde vier. De antwoordmogelijkheden “ik weet het niet” en missing values zijn verschillend gecodeerd en maken beide geen deel uit van de schaalscore. Met betrekking tot de interpretatie van de totaalscore van het CAP geeft een minimum score van één dus aan dat docenten het helemaal niet eens zijn met de kwaliteit van het CAP, twee het oneens zijn, drie het eens zijn en vier het helemaal eens zijn met de kwaliteit van het CAP.

### *Kwalitatieve analyse vragenlijst gepercipieerde kwaliteit CAP*

De kwalitatieve data uit de vragenlijst bestaan uit twee delen en zijn letterlijk getranscribeerd en kwalitatief geanalyseerd. Het eerste gedeelte, de schriftelijke motivaties, heeft het onderbouwen van de kwantitatieve schaalscores als doel. Dit is van belang om de resultaten valide te kunnen interpreteren (consequentiële validiteit). Zoals eerder beschreven hangt een valide interpretatie af van de kwaliteit van het bewijs (Kane, 1992). In dit onderzoek bestaat het bewijs uit schriftelijke motivaties van docenten en afdelingsleiders, al dan niet onderbouwd. De typologie van Clark & Sampson (2005) is gebruikt om de motivaties te kwalificeren. Zij kwalificeren motivaties aan de hand van een typologie dat onderscheid maakt tussen motivaties zonder onderbouwing (level 0), motivaties met uitleg als onderbouwing (level 1), motivaties met bewijs als onderbouwing (level 2), en motivaties met verschillende bewijsstukken als onderbouwing (level 3). Deze indeling maakt het mogelijk om de kwaliteit van de motivaties te verifiëren. Vanzelfsprekend werd een hoge mate van motivaties met verschillende bewijsstukken als onderbouwing beoogd. Uitleg als onderbouwing in level 1 bestaat uit persoonlijke ervaringen ten aanzien van een kwaliteitscriterium. Motivaties die buiten de typologie vielen, zijn apart gecodeerd. Het tweede gedeelte, schriftelijke toelichting of opmerkingen van docenten en afdelingsleiders over de duidelijkheid en begrijpelijkheid van de vragen in de vragenlijst heeft niet alleen betrekking op de interne validiteit van het meetinstrument. Ook wordt deels tegemoet gekomen aan de argument-based approach van Kane (1992), waarin wordt gesteld dat de kwaliteit van de dataverzameling (procedure) voorwaardelijk is voor de validiteit.

Om de tweede onderzoeksvraag te beantwoorden: *“In welke mate beïnvloeden performance assessments als onderdeel van het CAP voor docenten in school X, volgens docenten, de totale kwaliteit van het CAP?”*, zijn er drie analysestappen ondernomen.

Ten eerste is de gepercipieerde kwaliteit van de performance assessments op school X volgens docenten vastgesteld. Hiervoor zijn de kwantitatieve en de kwalitatieve data uit de vragenlijst over de kwaliteit van performance assessments op dezelfde manier geanalyseerd als de vragenlijst met betrekking tot de kwaliteit van het CAP. De kwalitatieve data bestaan echter alleen uit motivaties bij schaalscores. Daarnaast is ook voor deze vragenlijst een factor- en een betrouwbaarheidsanalyse uitgevoerd.

Ten tweede, is de relatie tussen de gepercipieerde kwaliteit van het CAP en de gepercipieerde kwaliteit van performance assessments onderzocht. De hypothese is dat er een relatie bestaat tussen een CAP en performance assessments omdat deze beoordelingsactiviteit onderdeel is van een CAP. De vraag is alleen hoe sterk het verband is. Hiervoor is gebruik gemaakt van een enkelvoudige regressie analyse na het controleren van de assumpties.

De uitvoering van de derde stap geeft inzicht in welke mate de kwaliteit van performance assessments mee weegt in de totale kwaliteit van een CAP of in welke mate een predictor bijdraagt in Y. Hiervoor is de proportie verklaarde variantie ( $R^2$ ) van de predictor op Y bepaald. Een significantie level van .05 en een relevantie level van  $> .15$  zijn als acceptabel beschouwd.

## **3. RESULTATEN**

### **Kwaliteit vragenlijsten**

Een confirmerende factoranalyse (principal component analyse) heeft de eendimensionaliteit van de subschalen bevestigd. In tabel 3 en 5 zijn de item-restcorrelaties en de Cronbach's Alpha waarden af te lezen voor elk kwaliteitscriterium. Zowel statistische als inhoudelijke criteria zijn gehanteerd om de keuze voor één of meerdere factoroplossingen te bepalen en om eventueel indicatoren te verwijderen uit de schalen. Hierbij is een Cronbach's Alpha waarde van minimaal (.60) en een item-restcorrelatie van minimaal (.35) gehanteerd. De Cronbach's Alpha waarden van alle kwaliteitscriteria voor zowel het CAP als de performance assessments kunnen als hoog worden gekwalificeerd behalve voor het kwaliteitscriterium Tijd & Kosten (-.31), tabel 3, en het

kwaliteitscriterium cognitieve complexiteit (.57), tabel 5. Een reden voor een negatieve Cronbach's Alpha voor het kwaliteitscriterium Tijd & Kosten zou de hoge mate van voorkomen van de antwoordmogelijkheid "ik weet het niet" voor dit kwaliteitscriterium kunnen zijn. 41% van de docenten heeft bijvoorbeeld aangegeven niet te weten of de beoordelaars het CAP redelijkerwijs uitvoeren. 48% weet niet of er voorafgaand aan de implementatie van het CAP een inschatting gemaakt wordt van tijd en kosten en 66% weet niet of er op regelmatige basis wordt geëvalueerd of het CAP efficiënter kan worden uitgevoerd. Hoewel "ik weet het niet" als missing value is verwerkt, is tijdens het coderen wel onderscheid gemaakt tussen "ik weet het niet" en missing values om inzicht te krijgen of docenten bepaalde zaken echt niet weten of dat er vragen niet zijn beantwoord. Indien dit onderscheid niet gemaakt zou zijn, zou er belangrijke data verloren zijn gegaan. Namelijk dat docenten soms echt geen antwoord kunnen geven op bepaalde vragen omdat zij het simpelweg niet weten. Dit gegeven kan gekwalificeerd worden als waardevolle feedback voor de school.

Een verklaring voor een lage Alpha waarde voor het kwaliteitscriterium cognitieve complexiteit zou een kleine variantie tussen de antwoordmogelijkheden van de verschillende indicatoren kunnen zijn, iets wat tevens resulteert in een lage item-restcorrelatie. Zo is 62% van de docenten het er over eens dat een performance assessment het denkproces oproept dat docenten hanteren. 0% is het hier helemaal niet mee eens, wat heeft geresulteerd in een item-restcorrelatie van (.11). Een acceptabele Cronbach's Alpha waarde (.60) voor dit kwaliteitscriterium zou bereikt kunnen worden door deze indicator te verwijderen. Vanwege inhoudelijke argumenten is hier echter niet voor gekozen. Deze indicator maakt namelijk inzichtelijk dat performance assessments hoog gekwalificeerd worden als beoordelingsactiviteit om competenties te meten. Verwijdering van de indicatoren vijf "de beoordeling van een performance assessment vindt plaats in een nieuwe complexe situatie" en zes "bij het uitvoeren van een performance assessment moet ik uitleggen waarom ik bepaalde keuzes heb gemaakt" heeft geen stijging van een acceptabele Cronbach's Alpha als gevolg en zijn om deze reden niet verwijderd uit de schaal.

Uit tabel 3 blijkt tevens dat indicator drie van kwaliteitscriterium transparantie "de beoordelaars kennen en begrijpen de inhoud, criteria en uitvoering van een performance assessment" een lage item-restcorrelatie heeft (.19). Verwijdering van dit item zou een stijging van de Cronbach's Alpha waarde hebben naar .83. De Alpha waarde voor dit kwaliteitscriterium was echter al acceptabel (.67). Daarnaast is dit item inhoudelijk relevant genoeg om opgenomen te blijven in de schaal transparantie. Dit criterium geeft namelijk aan dat een performance assessment transparant moet zijn voor alle betrokkenen: de beoordelaars en de beoordeelden. De kwalitatieve gegevens met betrekking tot dit kwaliteitscriterium hebben geen relevante en bruikbare resultaten opgeleverd. Indicator vijf van het kwaliteitscriterium zelfsturend leren bij performance assessments geeft ook een lage item-restcorrelatie aan (.30). Verwijdering van dit item levert geen stijging van de Cronbach's Alpha waarde op. Bovendien is dit item inhoudelijk representatief voor het kwaliteitscriterium zelfsturend leren en is vanwege deze argumentatie behouden.

### **Resultaat documentanalyse**

De documentanalyse is uitgevoerd op basis van onderstaande kenmerken van goed assessmentbeleid dat gebaseerd is op een cyclus van vijf stappen van goed competentie management beleid voor scholen in het voortgezet onderwijs (Ketelaar, 2008).

#### Verbinding competentie assessmentbeleid met strategie school

De wijze waarop het competentie assessmentbeleid verbonden is met de strategische doelen van de school is niet terug te vinden in welke vorm van documentatie dan ook.

#### Draagvlak creëren voor competentie assessment

Documentatie over het competentie assessment beleid is opgesteld door het management team van de school. Docenten in de rol als beoordeelden en beoordelaars hebben geen rol gespeeld tijdens het maken van dit beleid, noch in de ontwikkeling van competenties en de opzet van de procedures.

#### Duidelijke omschrijving competenties

Uit de uitgevoerde documentanalyse blijkt dat vijf van de vier afdelingen de zeven SBL competenties hanteren. De omschrijving van de SBL competenties voor docenten bestaan uit een algemene omschrijving bijvoorbeeld: Pedagogische competentie: “Zorgen voor een veilige leeromgeving en bevorderen van persoonlijke, sociale en morele ontwikkeling” met vijf competentieniveaus beginnend bij: niet competent, matig competent, basis competent, ervaren competent en excellent competent. De verschillen tussen de competentieniveau zijn echter niet geëxpliciteerd. Eén afdeling hanteert een eigen set van acht competenties waarbij een koppeling gemaakt wordt met de missie van de school. Naast een korte algemene omschrijving zijn er vier competentieniveaus omschreven. Van elke niveau worden voorbeelden gegeven.

#### Communicatie in de organisatie

Communicatie over de procedure van competentiebeoordeling en ontwikkeling bestaat voornamelijk uit schriftelijke documentatie. Uit de verzamelde documentatie blijkt dat er onderscheid gemaakt wordt tussen enerzijds benoeming en beoordeling van nieuwe docenten en anderzijds het functioneren van zittende docenten. De procedure van deze onderdelen zijn globaal beschreven.

#### Integratie van competentie management in de praktijk van het personeelsbeleid

In het personeelsbeleid staat de rol van competenties niet expliciet beschreven. Er is wel documentatie beschikbaar waarin het personeelsbeleid ten aanzien van werving en selectie, functioneren en beoordelen, opleiden in ontwikkelen of dossiervorming wordt geëxpliciteerd. Opmerkelijk is dat de functie van docentcompetenties hier niet in staan omschreven terwijl er in de praktijk docenten wel worden beoordeeld op basis van docentcompetenties.

#### Mix van beoordelingsactiviteiten met bijbehorende beoordelingscriteria

Er is een overzicht beschikbaar van beoordelingsactiviteiten die uitgevoerd worden tijdens de procedure van benoeming en beoordeling van nieuwe docenten en het functioneren van zittende docenten. Bij twee afdelingsleiders komen de procedures redelijk overeen in tegenstelling tot de drie overige afdelingsleiders. Daarnaast wordt per afdeling niet exact dezelfde beoordelingscriteria gehanteerd bij performance assessments.

### **Controle bruikbaarheid motivaties bij schaalscores en feedback op duidelijkheid en begrijpelijkheid**

De kwalitatieve data in dit onderzoek bestaan uit schriftelijke motivaties van docenten en afdelingsleiders op schaalscores en feedback op de duidelijkheid en begrijpelijkheid van de vragen in de vragenlijst. Dit heeft enerzijds als doel om de kwantitatieve schaalscores beter te kunnen interpreteren en anderzijds om inzicht te krijgen in de kwaliteit van de dataverzameling. Om te verifiëren of de kwalitatieve data daadwerkelijk voor deze doelen bruikbaar zijn, is de typologie van Clark & Sampson (2005) gehanteerd. Volgens deze typologie zou het merendeel van de motivaties gekwalificeerd worden als level 0, enkele motivaties als level 1. Level 2, 3 en 4 van de typologie komen niet voor. Daarnaast viel een groot aantal motivaties buiten de typologie. Zij impliceerden bijvoorbeeld een wens: “*ik zou graag willen dat de beoordelaar op lesbezoek zou komen*” of een aanvulling op een stelling. Bij de stelling: “*in het CAP wordt voldoende aandacht aan kennis besteed*” plaatste een docent de opmerking: “*dit is ook afhankelijk van de deskundigheid van de collega/leidinggevende*”. Dit type opmerking is geen motivatie op een schaalscore, noch zegt het iets over de kwaliteit van een CAP. De hoeveelheid van dit type motivatie resulteerde in een hoge mate van onbruikbare en irrelevante kwalitatieve data. Met een lage kwalitatieve respons als gevolg. Dit geldt ook voor de feedback over de duidelijkheid en begrijpelijkheid van de vragen. De mogelijkheid om feedback te geven op de duidelijkheid en begrijpelijkheid van de vragen is weinig tot niet benut. Er zijn slechts 16 opmerkingen geplaatst. De meest gemaakte opmerking is: “*ik snap de vraag niet*” bij het kwaliteitscriterium cognitieve complexiteit. De rest van de gegeven feedback sluit niet aan op een stelling of kwaliteitscriterium. Zo schreef een docent bij het kwaliteitscriterium doelen: “*CAP is alleen geschikt voor iemand met wie je samenwerkt/ die je meemaakt op het werk. Zo niet, dan is het te globaal*” een andere docent schreef als opmerking bij de stelling: er wordt in meerdere verschillende werksituaties beoordeeld: “*hiermee bedoel ik dat ik verschillende dingen doe*”

bijvoorbeeld klassikaal, activerende werkvormen”. Ten opzichte van de feedback over de duidelijkheid en begrijpelijkheid van de vragen geven de bruikbare motivaties daarentegen wel een indicatie waardoor schaalscores beter geïnterpreteerd kunnen worden. Om deze reden is gekozen om de motivaties toch inzichtelijk te maken. Bij de interpretatie van de kwantitatieve schaalscores worden de motivaties benut en aangehaald.

### Schaalscores, motivaties en feedback op Competentie Assessment Programma

De kwantitatieve schaalscores van de gepercipieerde kwaliteit van het CAP volgens docenten zijn weergegeven in tabel 2a. Hierin zijn de standaarddeviaties en de gemiddelde schaalscores weergegeven per kwaliteitscriterium van laag naar hoog. Eveneens is in tabel 2a de totaalscore van het gehele CAP volgens docenten opgenomen, uitgedrukt in een gemiddelde schaalscore. De kwantitatieve schaalscores van de gepercipieerde kwaliteit van het CAP volgens afdelingsleiders zijn weergegeven in tabel 2b. De resultaten zijn echter niet representatief voor alle afdelingsleiders omdat van de vijf afdelingsleiders er maar twee, de vragenlijst hebben ingevuld. Toch is de keuze gemaakt om de gemiddelde schaalscores per kwaliteitscriterium te presenteren om een indicatie te krijgen van de perceptie van de afdelingsleiders. Om een compleet beeld te krijgen van de kwaliteit van een CAP zijn namelijk inzichten van verschillende betrokkenen van belang.

Ter verduidelijking van de gemiddelde schaalscores per kwaliteitscriterium, gebaseerd op de perceptie van docenten, zijn in tabel 3 de schaalscores per stelling in percentages weergegeven. De percentages geven een weergave van de docenten die de antwoordmogelijkheden “helemaal mee oneens”, “mee oneens”, “mee oneens”, “helemaal mee eens” en “ik weet het niet” hebben gegeven. Een dergelijk overzicht van de afdelingsleiders is overigens niet gepresenteerd omdat er maar twee van de vijf hebben deelgenomen. De kwalitatieve motivaties van de docenten zijn gepresenteerd in tabel 4 en 5. Een onderverdeling is gemaakt tussen motivatie zonder onderbouwing (level 0) en met persoonlijke ervaring als onderbouwing (level 1).

Tabel 2a. Gemiddelde schaalscores per kwaliteitscriterium en totaalscore van CAP bij beoordeling door docenten

Kwaliteitscriterium CAP	k	n	Minimum	Maximum	Gemiddelde	Standaardafwijking
Vergelijkbaarheid	5	29	1	4	2.17	0.81
Herhaalbaarheid	7	29	1	4	2.53	0.64
Rechtvaardigheid	6	28	1	4	2.66	0.52
Cognitieve Complexiteit	6	28	1	4	2.70	0.47
Professionele gevolgen	5	27	1	4	2.72	0.66
Zelfsturend leren	6	28	1	4	2.78	0.49
Doelen	6	29	1	4	2.80	0.50
Tijd & Kosten	6	26	1	4	2.85	0.52
Acceptatie	4	29	1	4	2.86	0.46
Transparantie	4	28	1	4	3.01	0.56
Betekenisvolheid	4	28	1	4	3.06	0.54
<b>Totaal kwaliteit CAP</b>	<b>11</b>	<b>29</b>	<b>1</b>	<b>4</b>	<b>2.73</b>	<b>0.39</b>

Tabel 2b. Gemiddelde schaalscores per kwaliteitscriterium en totaalscore van CAP bij beoordeling door afdelingsleiders

Kwaliteitscriterium CAP	k	n	Minimum	Maximum	Gemiddelde	Standaardafwijking
Zelfsturend leren	6	2	1	4	2.47	0.19
Cognitieve Complexiteit	6	2	1	4	2.50	0.71
Tijd & Kosten	6	2	1	4	2.58	0.59
Doelen	6	2	1	4	2.67	0.24
Vergelijkbaarheid	5	2	1	4	2.70	0.14
Transparantie	4	2	1	4	2.71	0.06
Acceptatie	4	2	1	4	2.75	0.35
Herhaalbaarheid	7	2	1	4	2.80	0.24

Professionele gevolgen	5	2	1	4	2.80	0.28
Betekenisvolheid	4	2	1	4	2.88	0.17
Rechtvaardigheid	6	2	1	4	<u>3.00</u>	0.00
<b>Totaal kwaliteit CAP</b>	<b>11</b>	<b>2</b>	<b>1</b>	<b>4</b>	<b>2.72</b>	<b>0.26</b>

Tabel 3. Beschrijvende statistiek CAP

Kwaliteitscriterium en indicatoren CAP	Helemaal mee oneens	Mee oneens	Mee eens	Helemaal mee eens	Ik weet het niet	Cronbachs alpha & Item- restcorrelaties
<b>Doelen</b>						<b>.84</b>
1 Alle competenties beoordeeld in CAP	13.8	13.8	62.1	6.9	3.4	.76
2 In CAP voldoende aandacht voor kennis	3.4	37.9	44.8	6.9	6.9	.75
3 Vaardigheden voldoende in CAP	0.0	10.3	75.9	13.8	0.0	.54
4 Aandacht voor professionele houding	3.4	37.9	41.4	10.3	3.4	.52
5 Beoordelingscriteria gericht op K/V/H	3.4	27.6	55.2	10.3	0.0	.75
6 Beoordelingsactiviteiten vereisen integratie van K/V/H	0.0	10.3	65.5	17.2	6.9	.47
<b>Herhaalbaarheid</b>						<b>.87</b>
1 Competentie meerdere keren beoordeeld	20.7	37.9	27.6	3.4	6.9	.69
2 Combinatie eindoorddeel meerdere beoordelaars	13.8	27.6	27.6	24.1	6.9	.80
3 Beoordelaars met verschillende achtergrond	3.4	17.2	41.4	34.5	0.0	.66
4 Eindoorddeel gebaseerd op verschillende bewijzen	10.3	20.7	41.4	20.7	6.9	.68
5 Beoordeling in verschillende werksituaties	6.9	44.8	37.9	6.9	3.4	.44
6 Onderbouwing oordeel door beoordelaars	13.8	44.8	17.2	13.8	10.3	.55
7 Beoordelaars komen tot hetzelfde oordeel	17.2	31.0	31.0	0.00	20.7	.72
<b>Transparantie</b>						<b>.67</b>
1 Docent begrijpt competenties	0	10.3	62.1	24.1	3.4	.61
2 Docent begrijpt beoordelingscriteria	3.4	20.7	55.2	13.8	3.4	.65
3 Docent begrijpt inhoud, criteria, uitvoering CAP	0	24.1	34.5	13.8	27.6	<u>.19</u>
4 Docent kent uitvoering beoordelingsactiviteiten	0	10.3	55.2	24.1	6.9	.51
<b>Acceptatie</b>						<b>.71</b>
1 Goedkeuring docent mbt beoordelingscriteria	0	13.8	75.9	6.9	3.4	.46
2 Goedkeuring beoordelaar mbt uitvoering CAP	0	3.4	51.7	0	<u>44.8</u>	.46
3 Goedkeuring docent mbt beoordelingactiviteiten	0	10.4	69.0	24.1	3.4	.71
4 Goedkeuring docent mbt uitvoering CAP	10.3	37.9	31.0	13.8	6.9	.61
<b>Vergelijkbaarheid</b>						<b>.92</b>
1 Beoordeling op basis van vergelijkbare competenties	13.8	34.5	27.6	10.3	10.3	.84
2 Beoordelingscriteria voor docenten vergelijkbaar	24.1	13.8	34.5	13.8	10.3	.80
3 Beoordelaar beoordeeld op vergelijkbare wijze	17.2	6.9	17.2	6.9	48.3	.85
4 Beoordelingsactiviteiten voor docenten vergelijkbaar	17.2	13.8	13.0	3.4	31.0	.90
5 Werkomstandigheden voor docenten vergelijkbaar	44.8	34.5	13.8	3.4	3.4	.62
<b>Rechtvaardigheid</b>						<b>.84</b>
1 Beoordelaar oordeelkundig tav competenties	6.9	24.1	48.3	0	17.2	.40
2 Docenten ervaren het CAP als eerlijk	18.8	17.2	48.3	6.9	10.3	.85
3 Beoordelaar competent mbt beoordelingsactiviteiten	0.0	17.2	41.4	3.4	34.5	.80
4 Beoordelaar ervaart het CAP als eerlijk	0.0	0.0	24.1	10.3	62.1	.58
5 Beoordelingsactiviteiten biedt docent dezelfde kansen	6.9	13.8	34.5	0.0	41.4	.61
6 De docent kan bezwaar maken tegen een beoordeling	6.9	10.3	41.4	6.9	31.0	.86
<b>Zelfsturend leren</b>						<b>.79</b>
1 CAP stimuleert tot reflectie op professionele ontwikkeling docent	0.0	20.7	51.7	20.7	3.4	.73
2 Beoordelingscriteria gericht op ontwikkeling zelfsturend leren	6.9	13.8	48.3	13.8	13.8	.48
3 Docenten beoordelen zichzelf en elkaar	13.8	37.9	27.6	13.8	3.4	.50
4 Beoordelaar geeft geleidelijk initiatief bij beoordeling aan docent	6.9	17.2	37.9	13.8	20.7	.52
5 Docent heeft eigen inbreng bij keuze beoordelingsactiviteiten	13.8	6.9	37.9	3.4	34.5	.60
6 Docent formuleert eigen leerdoelen nav beoordelingen	0.0	13.8	72.4	6.9	3.4	.56
<b>Betekenisvolheid</b>						<b>.85</b>
1 Beoordelingscriteria goede basis om feedback op te baseren	0.0	17.2	58.6	20.7	0.0	.54
2 Beoordelaar geeft zinvolle feedback op basis van het CAP	0.0	20.7	55.2	10.3	10.3	.64
3 Docent ervaart uitvoeren beoordelingsactiviteiten als leermoment	3.4	10.3	44.8	27.6	10.3	.75
4 Docent vindt feedback op de beoordelingen zinvol	0.0	10.3	51.7	27.6	6.9	.83



Cognitieve Complexiteit							<b>.90</b>
1 CAP vereist denkniveau dat docenten nodig hebben	0.0	10.3	65.5	13.8	6.9		.69
2 CAP beoordeelt ook de toepassing van kennis	0.0	27.6	55.2	6.9	6.9		.63
3 Beoordelingscriteria zijn gericht op product en het denkproces	0.0	27.6	62.1	3.4	3.4		.80
4 Beoordelingsactiviteiten roepen denkproces op	0.0	24.1	37.9	10.3	24.1		.86
5 Beoordeling vindt plaats in nieuwe, complexe situatie	20.7	37.9	0.0	6.9	31.0		.69
6 Bij beoordelingsactiviteit moet de docent keuzes verklaren	0.0	31.0	41.4	13.8	6.9		.88
Professionele gevolgen							<b>.89</b>
1 Docent past lespraktijk aan indien CAP aanleiding geeft	0.0	13.8	58.6	17.2	6.9		.53
2 Docent wordt gemotiveerd door CAP	10.3	24.1	34.5	17.2	10.3		.86
3 CAP roept het gewenste gedrag op	3.4	31.0	41.4	13.8	6.9		.89
4 CAP roept gewenste gedrag op in voorbereiding op beoordelingen	6.9	44.8	24.1	10.3	10.3		.75
5 Docent krijgt inzicht in vereiste bekwaamheden beroep door CAP	10.3	27.6	34.5	13.5	10.3		.69
Tijd & Kosten							<b>-.31</b>
1 Docenten kunnen CAP redelijkerwijs uitvoeren	0.0	10.3	48.3	13.8	24.1		<u>-.27</u>
2 Beoordelaar kan het CAP redelijkerwijs uitvoeren	0.0	13.8	31.0	10.3	41.4		<u>-.27</u>
3 Tijd en kosten van invloed op keuze beoordelingsactiviteiten	3.4	13.8	24.1	31.0	24.1		<u>-.14</u>
4 CAP is redelijkerwijs uitvoerbaar in gekozen context	3.4	13.8	48.3	6.9	24.1		<u>-.34</u>
5 Voorafgaand wordt inschatting gemaakt van tijd en kosten CAP	13.8	13.8	13.8	6.9	48.3		.51
6 CAP wordt regelmatig geëvalueerd op efficiency	17.2	3.4	3.4	3.4	65.5		.49

Noot 1. Cronbachs  $\alpha$  if item Deleted:

Transparantie indicator 3 (.83)	Tijd & Kosten indicator 3 (.17)
Zelfsturend leren indicator 4 (.68)	Tijd & Kosten indicator 4 (.15)
Tijd & Kosten indicator 2 (.17)	Tijd & Kosten indicator 5 (.23)

### Gepercipieerde kwaliteit Competentie Assessment Programma volgens docenten

De resultaten laten een gemiddelde overall score van de kwaliteit van het CAP van 2.73 ( $SD= 0.39$ ) zien (tabel 2a). Deze score is gebaseerd op de gemiddelden van de elf kwaliteitscriteria en geeft aan dat de docenten het niet volledig eens zijn met de kwaliteit van het CAP op school X, daarentegen zijn zij het ook niet geheel oneens. De gemiddelde scores per kwaliteitscriterium geven inzicht in hoeverre de docenten het eens of oneens zijn met een kwaliteitscriterium. De kwaliteitscriteria vergelijkbaarheid ( $M= 2.17$ ;  $SD= 0.81$ ), herhaalbaarheid ( $M= 2.53$ ;  $SD= 0.64$ ) en rechtvaardigheid ( $M= 2.66$ ;  $SD= 0.52$ ) scoren het laagst. De kwaliteitscriteria acceptatie ( $M= 2.86$ ;  $SD= 0.46$ ), transparantie ( $M= 3.01$ ;  $SD= 0.56$ ) en betekenisvolheid ( $M= 3.06$ ;  $SD= 0.54$ ) scoren gemiddeld het hoogst. Uit tabel 2a blijkt dat het kwaliteitscriterium vergelijkbaarheid de laagste gemiddelde schaalscore heeft ( $M= 2.17$ ) met een hogere standaarddeviatie ( $SD= 0.81$ ) ten opzichte van de andere standaarddeviaties. De percentages in tabel 3 van het kwaliteitscriterium vergelijkbaarheid bevestigen de spreiding van de verschillende antwoordmogelijkheden. Eensgezind over dit kwaliteitscriterium zijn de docenten dus allerminst. Het lage gemiddelde impliceert dat de docenten over het algemeen het meest oneens zijn met dit kwaliteitscriterium ten opzichte van de andere kwaliteitscriteria. Met andere woorden, de docenten zijn het minst eens met de wijze waarop het CAP is opgezet en wordt uitgevoerd. Hiermee wordt met name de consistentie van het CAP bedoeld. Opmerkelijk is het hoge percentage (48%) van indicator drie aangeeft dat docenten niet weten of de beoordelaars verschillende docenten op vergelijkbare wijze beoordelen. De lage score op dit kwaliteitscriterium zou deels verklaard kunnen worden door de score op indicator vijf "de werkomstandigheden zijn voor alle docenten gelijk". Slechts 45% van de docenten is het helemaal oneens (schaalscore 1) en 35% is het oneens (schaalscore 2) met deze stelling. De motivaties op dit kwaliteitscriterium hebben voornamelijk betrekking op indicator vijf waarin wordt benadrukt dat de werkomstandigheden voor alle docenten niet vergelijkbaar is (tabel 4 en 5). Uit de motivaties blijkt dat de term 'werkomstandigheden' geïnterpreteerd is als het lesgeven aan verschillende jaarlagen en niveaus. Deze interpretatie komt echter niet overeen met de feitelijke betekenis. Onder werkomstandigheden worden de omstandigheden waaronder een assessment wordt uitgevoerd bedoeld. Het kwaliteitscriterium betekenisvolheid heeft het hoogste gemiddelde

( $M= 3.06$ ) ten opzichte van de andere kwaliteitscriteria wat inhoudt dat het CAP een significante waarde heeft voor de docenten op school X. De standaarddeviatie is relatief laag ( $SD= 0.54$ ). Dit impliceert een redelijke overeenstemming over het kwaliteitscriterium betekenisvolheid onder de docenten. De percentages uit tabel 3 bevestigen dit beeld. Ongeveer de helft van de docenten is het eens met de vier stellingen van het kwaliteitscriterium betekenisvolheid. Dit houdt in dat docenten de beoordelingscriteria een goede basis vinden om feedback op te baseren, dat de beoordelaars zinvolle feedback geven op basis van het CAP, het uitvoeren van beoordelingsactiviteiten als leermoment ervaren en dat de feedback op de beoordelingen zinvol zijn. De motivaties van docenten op dit criterium laten echter een minder eensgezind beeld zien. Drie docenten reageerden negatief en twee positief. Zo geeft een docent aan dat niet alle leerlinginput even zinvol is. Een andere docent geeft juist aan door feedback, verbeterpunten te ontdekken.

### **Gepercipieerde kwaliteit Competentie Assessment Programma volgens afdelingsleiders**

#### *Kwantitatieve schaalscores*

De resultaten laten een gemiddelde overall score van de kwaliteit van het CAP van 2.72 ( $SD= 0.26$ ) zien (tabel 2b). Deze score is gebaseerd op de gemiddelden van tien kwaliteitscriteria en geeft aan dat de afdelingsleiders het niet volledig eens zijn met de kwaliteit van het CAP op school X, daarentegen zijn zij het ook niet geheel oneens. De gemiddelde scores per kwaliteitscriterium geven inzicht in hoeverre de afdelingsleiders het eens of oneens met een kwaliteitscriterium zijn. De kwaliteitscriteria zelfsturend leren ( $M= 2.47$ ;  $SD= 0.19$ ), cognitieve complexiteit ( $M= 2.50$ ;  $SD= 0.19$ ) en tijd & kosten ( $M= 2.58$ ;  $SD= 0.59$ ) scoren het laagst. De kwaliteitscriteria professionele gevolgen ( $M= 2.80$ ;  $SD= 0.28$ ), betekenisvolheid ( $M= 2.88$ ;  $SD= 0.17$ ) en rechtvaardigheid ( $M= 3.00$ ;  $SD= 0.00$ ) scoren gemiddeld het hoogst. Met betrekking tot de totale kwaliteit van het CAP lijken de docenten ( $M= 2.73$ ;  $SD= 0.39$ ) en afdelingsleiders ( $M= 2.72$ ;  $SD= 0.26$ ) eensgezind. Ook hier liggen de gemiddelde schaalscores per kwaliteitscriterium tussen de twee en drie. Echter, verschilt de volgorde van de gemiddelde schaalscores per kwaliteitscriterium van laag naar hoog (tabel 2a en 2b). Opvallend is het verschil tussen docenten en afdelingsleiders met betrekking tot de gemiddelde schaalscore op het kwaliteitscriterium rechtvaardigheid. Dit criterium is één van de drie criteria waar de docenten het meest mee oneens zijn, de twee afdelingsleiders zijn het hier daarentegen, volledig mee eens ( $M= 3.00$ ;  $SD= 0.00$ ). Het kwaliteitscriterium betekenisvolheid is één van de drie kwaliteitscriteria waar zowel de docenten ( $M= 3.06$ ;  $SD= 0.54$ ) als de afdelingsleiders ( $M= 2.88$ ;  $SD= 2.88$ ) het meest mee eens zijn.

#### *Schriftelijke motivaties op kwaliteitscriteria van afdelingsleiders*

Met betrekking tot het kwaliteitscriterium professionele doelen geeft één van de twee afdelingsleiders aan dat een aantal aspecten met betrekking tot het meten van een professionele houding en vakkennis onderbelicht is. Er is echter wel aandacht voor pedagogische en didactische kennis. De andere afdelingsleider geeft als algemene motivatie op dit criterium aan ontevreden te zijn over de formulering van de docentcompetenties: “de omschrijvingen van de competenties zijn zo vaag/ algemeen dat iedereen al gauw in “ervaren” komt. Bij het criterium herhaalbaarheid wordt aangegeven dat er bij beoordelingen wel overleg is tussen sectieleider en afdelingsleider maar dat het management team uiteindelijk beslist. Een afdelingsleider geeft daarnaast aan, het competentie formulier een erg onhandig instrument te vinden, zo blijkt uit de toelichting bij het criterium transparantie. Met betrekking tot het kwaliteitscriterium acceptatie geeft een afdelingsleider aan dat de procedure van het CAP niet duidelijk is. Daarnaast interpreteren de afdelingsleiders de term “werkomstandigheden” op dezelfde wijze zoals de docenten namelijk als het lesgeven aan verschillende jaarlagen en niveaus blijkt ook hier uit de schriftelijke motivaties. Zo schrijven de afdelingsleiders: “werkomstandigheden verschillen per domein” en “verschillende klassen en domeinen”. Ten slotte geeft een afdelingsleider aan niet voldoende tijd te hebben om het CAP redelijkerwijs uit te voeren.

Tabel 4. Motivaties zonder onderbouwing bij kwaliteitscriteria CAP

Kwaliteitscriterium en indicatoren CAP	Motivatie zonder onderbouwing (level 0)
<i>Doelen:</i> aansluiting tussen docentcompetenties en beoordelingsactiviteiten	<p>Professionele houding komt in gesprekken niet aan de orde                      Alles zit er wel in maar naar mijn idee worden er te weinig onderdelen verbonden tot één conclusie. Didactisch en vakinhoudelijke kennis komt te weinig aan bod.                      Integratie laat te wensen over. De beoordeling is te veel rap-snap                      Er zit verschil tussen CAP in theorie en in uitvoering</p>
<i>Herhaalbaarheid:</i> oordelen gebaseerd op combinatie meerdere bewijzen	<p>Eindoordeel is alleen gebaseerd op enquêtes                      Er wordt ons niet verteld of er tussen beoordelaars een gelijkwaardig overleg plaatsvindt waarin iedereen zijn oordeel onderbouwt: dus geen inzicht                      Er wordt ons niet verteld of meerdere beoordelaars onafhankelijk van elkaar tot hetzelfde oordeel komen: dus geen inzicht                      Veel van wat er in het CAP staat wordt niet door verschillende personen verbonden                      Leerlingen onderbouwen hun oordeel niet omdat het anoniem via de ELO is                      Er is maar 1 beoordelaar                      Beoordeling inclusief functioneringsgesprek vindt niet of nauwelijks plaats. Als het plaats vindt, alleen met de afdelingsleider.                      Nooit met anderen over gesproken.                      Voor beoordeelde en beoordelaars vaak onduidelijk tot welk oordeel de schoolleiding (meestal) komt                      Bewijzen en verslagen worden selectief gebruikt. Eerst oordeel dan bewijs in plaats van andersom</p>
<i>Transparantie:</i> duidelijkheid en begrijpelijk voor alle betrokkenen	<p>Ik vraag me af of docenten de inhoud, criteria en uitvoering van het CAP kennen en begrijpen. Wordt het geheel aan competenties wel overzien?                      Beoordelingscriteria en normen niet duidelijk en transparant</p>
<i>Acceptatie:</i> betrokkenen kunnen zich vinden in CAP	<p>In mijn beleving is er geen structurele uitvoering van het CAP op onze school                      Het uitvoeren van de procedure gaat allemaal wel wat "losjes".                      Er is geen verbinding tussen de verschillende beoordelaars.                      Ja, er moet een duidelijke procedure zijn. Maar die is de afgelopen jaren slecht gevolgd.                      Er wordt slecht teruggekoppeld over bevindingen, door beoordelaars en directie</p>
<i>Vergelijkbaarheid:</i> Consistentie opzet en uitvoering CAP	<p>Gym, muziek, O&amp;O, geschiedenis, taal, bv, c.k.v. allemaal verschillend.                      Alles eens maar het is de vraag of bij elke beoordeling iedereen dezelfde competenties en criteria meenemen.                      Er is op dit gebied nog heel wat te ontwikkelen op onze school</p>
<i>Rechtvaardigheid:</i> eerlijk en zorgvuldig, geen bias	<p>Niet alle beoordelingsactiviteiten worden uitgevoerd.                      In de praktijk verandert er echter niets.                      Bezwaar maken tegen een beoordeling heeft weinig zin.                      Beoordelaars zijn lang niet altijd competent voor bepaalde activiteiten.                      Op onze school is het afhankelijk van persoon en tijd/ruimte wat er wel/niet gebeurt.</p>
<i>Zelfsturend leren:</i> CAP stimuleert ontwikkeling zelfsturend leren	<p>Er wordt vanuit de school te weinig met CAP gedaan op het gebied van mijn professionele ontwikkeling                      Er is wel reflectie maar dan stopt het                      Dat docenten zichzelf en elkaar beoordelen gebeurt te weinig                      We beoordelen alleen onszelf en niet elkaar.                      We beoordelen met name onszelf.</p>

	<p><i>Er kan zelf</i> aangegeven worden wanneer je als docent het initiatief wil nemen bij beoordelingen                  Docenten hebben geen eigen inbreng bij de keuze voor beoordelingsactiviteiten: nooit gebeurd.                  Ik denk dat er te weinig met de beoordelingen gedaan worden.                  Ik heb nooit ervaren op onze school dat het initiatief bij de beoordeling geleidelijk over wordt gegeven aan de docent</p>
<i>Betekenisvolheid</i> : formatief doel meer dan alleen meten	<p>Geen zinvolle feedback op de beoordelingen meegemaakt.                  Ik vind dat ik na elk gesprek iets heb om aan te werken en om mezelf verbeteren.                  Niet alle leerlingeninput was even zinvol</p>
<i>Cognitieve complexiteit</i> : inhoud, meten hogere cognitieve vaardigheden, ook proces	<p>Nooit ervaren dat het zinvol was m.b.t. mijn functioneren.                  Ik betwijfel of vakkennis echt gemeten wordt                  Meer op product gericht, dan op deel-proces</p>
<i>Professionele gevolgen</i> : effect op leren en instructie	<p>Ik zou mijn dagelijkse lespraktijk wel aanpassen als de uitkomsten van een lesobservatie hier aanleiding toegeven maar de aanleiding ontbreekt.                  Als het CAP volledig wordt uitgevoerd, zou ik nog meer gemotiveerd worden.                  Ik word niet gemotiveerd maar leer wel veel van de beoordeling bij het CAP.</p>
<i>Tijd &amp; Kosten</i> : oordelen gebaseerd op combinatie meerdere bewijzen	<p>Ik heb geen goed beeld of het CAP redelijk uit te voeren is                  Ik heb geen goed beeld of het CAP redelijkerwijs uitvoerbaar is in de gekozen context                  Het CAP is mijn inziens veelal ondoorzichtig en te weinig toegepast.</p>

Tabel 5. Motivaties met onderbouwing bij kwaliteitscriteria CAP

<b>Kwaliteitscriterium en indicatoren CAP</b>	<b>Motivatie met persoonlijke ervaring als onderbouwing (level 1)</b>
<i>Herhaalbaarheid</i> : oordelen gebaseerd op combinatie meerdere bewijzen	<p>Schaalscores zijn gebaseerd op de beoordeling van mijzelf in mijn eerste jaar en ook de beoordelingen en lesobservaties die ik zelf heb uitgevoerd bij de beoordeling van mijn collega's.                  Er waren geen meerdere beoordelaars die lessen bezocht hebben, maar slechts bezoek van de afdelingsleider</p>
<i>Transparantie</i> : duidelijkheid en begrijpelijk voor alle betrokkenen	<p>Ik werd doodgegooid met docentcompetenties op de HU dus ben hier erg bekend mee.                  Beoordelingscriteria waren voor mijn beoordeling niet duidelijk. Normering nog steeds niet.                  Ik wist niet waarop ik beoordeeld werd. Ik had toen nog geen idee van docentencompetenties.</p>
<i>Vergelijkbaarheid</i> : Consistentie opzet en uitvoering CAP	<p>Met de verschillende domeinen is de werkomgeving en zijn de -omstandigheden niet voor een ieder gelijk.                  De werkomstandigheden in 3 Mavo zijn niet geheel gelijk aan die in 5 VWO dus een lesobservatie in 3M is niet gelijk aan die van 5V.                  Elk domein is anders                  4 Mavo is wat anders dan 1Gym en vice versa. Bij lesbezoeken wordt hier te weinig rekening mee gehouden. Niet alle activiteiten worden altijd ingezet.</p>
<i>Rechtvaardigheid</i> : eerlijk en zorgvuldig, geen bias	<p>De 3 gesprekken die ik de afgelopen jaren heb gehad zijn met de afdelingsleider geweest. Ik heb veel respect voor het oordeel en deskundigheid van de afdelingsleider maar ik weet niet hoe objectief hij naar mij toe is. Hij vindt me duidelijk goed maar hij is nooit in een van mijn lessen geweest.                  Indien het rechtvaardig zou zijn, zou ik toch al langere tijd in LD moeten zitten. Ik zit nog steeds in LC.</p>

Zelfsturend leren: CAP stimuleert ontwikkeling zelfsturend leren	Voor de beoordeling wordt gebruik gemaakt van het formulier "Competenties GRC". Alleen leerlingen enquêtes worden hiervoor gebruikt. De beoordelaar geeft aan waar hij/zij de volgende keer naar wil kijken. Leerdoelen met behulp van beoordelen, vaak mag je zelf nog iets aangeven waar expliciet opgelet wordt, bijvoorbeeld eerdere doelen.
---	--

### Schaalscores, motivaties en feedback op performance assessments

De kwantitatieve resultaten van de gepercipieerde kwaliteit van performance assessments zijn weergegeven in tabel 6. Hierin zijn ook de standaarddeviaties en de gemiddelde schaalscores weergegeven per kwaliteitscriterium van laag naar hoog. In tabel 6 is eveneens de totale kwaliteit van performance assessments, uitgedrukt in een gemiddelde schaalscore opgenomen. In Tabel 7 zijn de schaalscores per indicator in percentages weergegeven. In de tabellen 8 en 9 zijn de motivaties per kwaliteitscriterium gepresenteerd.

Tabel 6. Gemiddelde schaalscores per kwaliteitscriterium en totaalscore van performance assessments

Kwaliteitscriterium	k	n	Minimum	Maximum	Gemiddelde	Standaardafwijking
<b>Performance assessments</b>						
Vergelijkbaarheid	5	27	1	4	<u>2.33</u>	0.63
Herhaalbaarheid	7	26	1	4	2.51	0.62
Doelen	6	27	1	4	2.66	0.37
Rechtvaardigheid	6	27	1	4	2.71	0.47
Professionele gevolgen	5	26	1	4	2.77	0.50
Acceptatie	4	26	1	4	2.80	0.44
Zelfsturend leren	6	26	1	4	2.80	0.38
Cognitieve Complexiteit	6	27	1	4	2.84	0.39
Transparantie	4	27	1	4	2.99	0.58
Betekenisvolheid	4	27	1	4	<u>3.00</u>	0.44
<b>Totaal kwaliteit</b>	<b>10</b>	<b>28</b>	<b>1</b>	<b>4</b>	<b>2.76</b>	<b>0.33</b>
<b>Performance assessments</b>						

### Gepercipieerde kwaliteit performance assessments

In tabel 4 blijkt de gemiddelde totaalscore van de kwaliteit van performance assessments ( $M= 2.76$ ;  $SD= 0.44$ ) te zijn. Geen groot verschil met de gemiddelde totaal score van de kwaliteit van het CAP. Deze score is gebaseerd op de gemiddelden van de tien kwaliteitscriteria en geeft aan dat docenten het niet volledig eens zijn met de kwaliteit van performance assessments. Geheel ontevreden zijn ze echter ook niet. De kwaliteitscriteria vergelijkbaarheid ( $M=2.33$ ;  $SD= 0.63$ ), herhaalbaarheid ( $M= 2.51$ ;  $SD= 0.62$ ) en doelen ( $M= 2.66$ ;  $SD= 0.37$ ) scoren het laagst. De kwaliteitscriteria cognitieve complexiteit ( $M= 2.84$ ;  $SD= 0.39$ ), transparantie ( $M= 2.99$ ;  $SD= 0.58$ ) en betekenisvolheid ( $M= 3.00$ ;  $SD= 0.44$ ) scoren gemiddeld het hoogst.

Tabel 7. Beschrijvende statistiek performance assessments

Kwaliteitscriterium en indicatoren performance assessments	Helemaal mee oneens	Mee oneens	Mee eens	Helemaal mee eens	Ik weet het niet	Cronbachs alpha & Item-restcorrelaties
<b>Doelen</b>						
1 Alle competenties beoordeeld in lesobservaties	10.3	65.5	17.2	0.0	3.4	.67
2 Tijdens lesobservaties voldoende aandacht voor kennis	6.9	31.0	48.3	0.0	6.9	.57
3 Vaardigden voldoende in lesobservaties	3.4	27.6	55.2	3.4	6.9	.37
4 Aandacht voor professionele houding	3.4	41.4	48.3	0.0	3.4	.37
5 Beoordelingscriteria gericht op K/V/H	3.4	17.2	65.5	6.9	3.4	.56
6 Beoordelingsactiviteiten vereisen integratie van K/V/H	3.4	44.8	44.8	0.0	3.4	.15
<b>Herhaalbaarheid</b>						
1 Competentie meerdere keren beoordeeld	13.8	41.4	24.1	3.4	13.8	.91
2 Combinatie eendoordeel meerdere beoordelaars	20.7	13.8	44.8	10.3	6.9	.88
3 Beoordelaars met verschillende achtergrond	0.0	6.9	51.7	27.6	6.9	.85
4 Eindoordeel gebaseerd op verschillende bewijzen	13.8	24.1	44.8	3.4	6.9	.52
5 Beoordeling in verschillende werksituaties	3.4	31.0	44.8	3.4	10.3	.75
6 Onderbouwing oordeel door beoordelaars	17.2	34.5	27.6	3.4	10.3	.39
7 Beoordelaars komen tot hetzelfde oordeel	20.7	17.2	13.8	6.9	34.5	.93
<b>Transparantie</b>						
1 Docent begrijpt competenties	6.9	6.9	65.5	13.8	0.0	.89
2 Docent begrijpt beoordelingscriteria	3.4	17.2	51.7	17.2	3.4	.65
3 Docent begrijpt inhoud, criteria, uitvoering lesobservaties	0.0	10.3	48.3	10.3	24.1	.87
4 Docent kent en begrijpt uitvoering lesobservatie	0.0	3.4	72.4	17.2	0.0	.81
<b>Acceptatie</b>						
1 Goedkeuring docent mbt beoordelingscriteria	3.4	17.2	62.1	3.4	6.9	.78
2 Goedkeuring beoordelaar mbt uitvoering lesobservatie	0.0	20.7	37.9	3.4	31.0	.91
3 Goedkeuring docent mbt uitvoering lesobservatie	0.0	10.3	75.9	3.4	3.4	.82
4 Goedkeuring docent mbt procedure lesobservatie	0.0	24.1	58.6	3.4	6.9	.84
<b>Vergelijkbaarheid</b>						
1 Beoordeling op basis van vergelijkbare competenties	10.3	24.1	44.8	0.0	10.3	.88
2 Beoordelingscriteria voor docenten vergelijkbaar	13.8	20.7	41.4	0.0	13.8	.70
3 Beoordelaar beoordeeld op vergelijkbare wijze	10.3	13.8	27.6	0.0	41.4	.90
4 Lesobservatie voor docenten vergelijkbaar	17.2	24.1	41.4	0.0	10.3	.77
5 Werkomstandigheden voor docenten vergelijkbaar	27.6	27.6	37.9	0.0	0.0	.82
<b>Rechtvaardigheid</b>						
1 Beoordelaar oordeelkundig tav competenties	0.0	17.2	44.8	3.4	27.6	.46
2 Docenten ervaren lesobservaties als eerlijk	10.3	10.3	65.5	3.4	3.4	.89
3 Beoordelaar competent mbt uitvoering lesobservaties	3.4	17.2	55.2	3.4	13.8	.84
4 Beoordelaar ervaart lesobservaties als eerlijk	0.0	6.9	44.8	6.9	34.5	.62
5 Lesobservatie biedt docent dezelfde kansen	10.3	27.6	20.7	0.0	34.5	.54
6 De docent kan bezwaar maken tegen een beoordeling	6.9	13.8	37.9	6.9	27.6	.90
<b>Zelfsturend leren</b>						
1 Stimuleert reflectie op professionele ontwikkeling	0.0	3.4	55.2	31.0	3.4	.65
2 Beoordelingscriteria gericht op ontwikkeling zelfsturend leren	3.4	6.9	55.2	6.9	20.7	.75
3 Docenten beoordelen zichzelf en elkaar	6.9	44.8	31.0	6.9	3.4	.40
4 Beoordelaar geeft geleidelijk initiatief aan docent	0.0	20.7	48.3	6.9	17.2	.55
5 Keuze in wel/niet beoordeeld worden dmv lesobservaties	6.9	37.9	31.0	3.4	13.8	.63
6 Docent formuleert eigen leerdoelen nav lesobservaties	0.0	17.2	62.1	6.9	6.9	.30
<b>Betekenisvolheid</b>						
1 Beoordelingscriteria goede basis om feedback op te baseren	0.0	10.3	65.5	13.8	3.4	.30
2 Beoordelaar geeft zinvolle feedback op basis lesobservaties	3.4	17.2	55.2	6.9	10.3	.55
3 Docent ervaart uitvoeren lesobservaties als leermoment	0.0	10.3	65.5	10.3	6.9	.72
4 Docent vindt feedback op lesobservaties zinvol	0.0	10.3	55.2	20.7	6.9	.52
<b>Cognitieve Complexiteit</b>						
1 Lesobservatie vereist denkniveau dat docenten nodig hebben	0.0	0.0	69.0	17.2	6.9	.80
2 Lesobservatie beoordeelt ook de toepassing van kennis	3.4	24.1	51.7	13.8	0.0	.57
3 Beoordelingscriteria gericht op product en het denkproces	3.4	20.7	55.2	3.4	10.3	.39
4 Lesobservaties roepen denkproces op van het beroep	0.0	6.9	62.1	3.4	17.2	.56
5 Beoordeling vindt plaats in nieuwe, complexe situatie	6.9	41.4	10.3	0.0	34.5	.45
6 Bij lesobservaties moet de docent keuzes verklaren	0.0	17.2	48.3	20.7	6.9	.11
						.23
						.27

Professionele gevolgen						.81
1 Docent past lespraktijk aan indien aanleiding is	0.0	0.0	69.0	20.7	3.4	.56
2 Docent wordt gemotiveerd door lesobservaties	3.4	31.0	41.4	10.3	6.9	.66
3 Lesobservatie roept het gewenste gedrag op	0.0	41.4	27.6	6.9	17.2	.76
4 Lesobservatie roept gewenste gedrag op in voorbereiding op beoordeling	3.4	37.9	17.2	10.3	24.1	.81
5 Docent krijgt inzicht in vereiste bekwaamheden beroep door lesobservatie	0.0	27.6	55.2	6.9	3.4	.25

**Noot.** Cronbachs  $\alpha$  if item Deleted:

Cognitieve Complexiteit indicator 1 (.50)

Cognitieve Complexiteit indicator 4 (.60)

Cognitieve Complexiteit indicator 2 (.40)

Cognitieve Complexiteit indicator 5 (.57)

Cognitieve Complexiteit indicator 3 (.48)

Cognitieve Complexiteit indicator 6 (.55)

Opmerkelijk is dat het kwaliteitscriterium vergelijkbaarheid ook hier het laagst ( $M= 2.33$ ) en eveneens de standaarddeviatie het hoogst scoort ( $SD= 0.63$ ) ten opzichte van de andere kwaliteitscriteria. Hier geeft het lage gemiddelde aan dat de docenten het minst eens zijn met de wijze waarop lesobservaties zijn opgezet en uitgevoerd worden. Uit tabel 5 blijkt dat niemand van de docenten het helemaal eens is met geen van de indicatoren van dit kwaliteitscriterium. Zo is niemand het helemaal eens met de volgende stellingen: “iedere docent wordt op vergelijkbare competenties beoordeeld, de beoordelingscriteria en normering zijn voor alle docenten vergelijkbaar, de beoordelaars beoordelen verschillende docenten op vergelijkbare wijze, de beoordelingsactiviteiten zijn voor alle docenten vergelijkbaar en de werkomstandigheden zijn voor alle docenten vergelijkbaar”. Wat tevens leidt tot een lagere gemiddelde is het hoge percentage 41% “ik weet het niet” bij indicator drie “de beoordelaars beoordelen verschillende docenten op vergelijkbare wijze”. Daarnaast vinden meer dan de helft van de docenten dat de werkomstandigheden voor alle docenten niet vergelijkbaar zijn. Drie van de vijf bruikbare kwantitatieve gegevens onderbouwen dit.

Het kwaliteitscriterium betekenisvolheid scoort ook op de kwaliteit van lesobservaties het hoogst ( $M= 3.00$ ;  $SD= 0.44$ ) ten opzichte van de andere kwaliteitscriteria. Ook lesobservaties hebben een significante waarde voor docenten. Een ruime meerderheid van de docenten is het steeds eens met de vier indicatoren van dit kwaliteitscriterium blijkt uit tabel 5.

### Relatie tussen performance assessments en competentie assessment programma's

De assumpties voor een regressieanalyse blijken te voldoen aan de voorwaarden. Er blijkt een relatief sterke positieve samenhang ( $r= 0.75$ ;  $p < 0,001$ ;  $n= 28$ ) te zijn tussen de totale kwaliteit van een CAP en de totale kwaliteit van performance assessments. Dit sterke verband is verklaarbaar doordat performance assessments daadwerkelijk onderdeel zijn van het CAP. Een andere verklaring heeft te maken met de wijze waarop docenten de vragenlijst voor de kwaliteit van het CAP hebben ingevuld. Uit de schriftelijke motivaties blijkt dat de docenten tijdens het invullen van deze vragenlijst niet het gehele CAP in gedachte te hadden maar voornamelijk één beoordelingsactiviteit te weten, lesobservaties (performance assessment). Zo stond bovenaan de vragenlijst voor de kwaliteit van het CAP: “Ingevuld als lesobservatie. Naar mijn weten zijn er geen andere beoordelingsactiviteiten”. Een andere docent heeft bij het criterium zelfsturend leren, wat inhoudt dat het gehele CAP de ontwikkeling van zelfsturend leren zou moeten stimuleren, als opmerking geplaatst: “De beoordelaar geeft aan waar hij/zij de volgende keer naar wil kijken”. Dit voorbeeld impliceert dat de docent niet het gehele CAP in gedachte heeft maar alleen performance assessments.

Van de variantie van de totale kwaliteit van een CAP kan 56% verklaard worden door de totale kwaliteit van lesobservaties ( $R^2= 0.56$ ;  $p < 0,000$ ;  $F=33,14$ ;  $n= 28$ ). Er is dus sprake van een sterk lineair verband waarbij  $R^2 > .15$  als criteria is gehanteerd (wat acceptabel is bij één predictor). Dit betekent dat 44% van de variantie van de totale kwaliteit van CAP wordt verklaard door andere factoren die niet meegenomen zijn in dit onderzoek.

Tabel 8. Motivaties zonder onderbouwing bij kwaliteitscriteria performance assessments

Kwaliteitscriterium en indicatoren lesobservatie	Motivatie zonder onderbouwing (level 0)
<i>Doelen:</i> aansluiting tussen docentcompetenties en beoordelingsactiviteiten	<p>Competentie 5,6 en 7 worden niet door lesobservatie beoordeeld. Competentie 4 is een momentopname.</p> <p>In slechts 1 les krijg je als docent niet altijd te maken met probleemoplossend handelen, terwijl dat volgens mij juist iets is waar je als (beginnend) docent feedback op wil/moet krijgen.</p> <p><i>Alle te bewijzen docentcompetenties worden beoordeeld tijdens een lesobservatie:</i> niet elke les hoeft dat te bieden; aanvulling met gesprek nodig; gaat niet via vaststaande formulieren.</p> <p>Vakkennis is door beoordelaars is moeilijk in te schatten</p> <p>Vaardigheden komen voldoende voor tijdens een lesobservatie: dat is afhankelijk van de kwaliteit van de observant en het gebruikte instrument</p> <p>Ik twijfel sterk of alle competenties tijdens de lesobservatie beoordeeld worden. Niet alle aspecten komen bij een lesobservatie ook aan de orde.</p>
<i>Herhaalbaarheid:</i> oordelen gebaseerd op combinatie meerdere bewijzen	<p>Overleg is onzichtbaar</p> <p>Je wordt niet vergeleken wanneer er sprake is van haast. Er is geen toelichting gevraagd van de beoordelaar</p>
Transparantie: duidelijkheid en begrijpelijkheid voor alle betrokkenen	<p>Ben op de HU alleen maar met competenties bezig geweest. Docenten die al langer in het VO werken hebben hier minder ervaringen mee.</p>
<i>Acceptatie:</i> betrokkenen kunnen zich vinden in CAP	<p>Het belangrijkste hiervoor is het nabespreken, niet de observatiecriteria ansich</p> <p>Je hoort te laten zien dat je de procedure op de juiste wijze uitvoert</p> <p>Domeinen zorgen ervoor dat werksituatie voor docenten niet vergelijkbaar is.</p>
<i>Vergelijkbaarheid:</i> consistentie opzet en uitvoer lesobservatie	<p>Op dit gebied moet er nog hard gewerkt worden.</p> <p>Juist omdat de omstandigheden zo verschillen is er een framework (CAP) nodig zodat toch voor alle competenties aandacht kan zijn.</p> <p>Lesobservaties zijn vaak sporadisch om goed te kunnen beoordelen, met betrekking tot vergelijkbaarheid.</p> <p>Mavo 3/4 is wat anders dan 1Gym. Beoordelaars houden daar weinig/geen rekening mee.</p>
<i>Rechtvaardigheid:</i> eerlijk en zorgvuldig, geen bias	<p>Beoordelaars zijn niet altijd vakdidactisch kundig</p> <p>De beoordelaars zijn niet altijd competent voor het afnemen van lesobservaties</p> <p>Ik kan bezwaar maken tegen een beoordeling van een lesobservatie: Er is altijd ruimte voor een gesprek</p> <p>Ik krijg soms de indruk dat sommige docenten niet weten wat een bepaalde competentie nou daadwerkelijk inhoudt.</p> <p>Bij een observatie gaat het over hoe het overkomt bij de beoordelaar. Dit is voor 50 % afhankelijk van de persoon van de beoordelaar en voor 50% van de activiteiten van de beoordeelde</p>
<i>Zelfsturend leren:</i> lesobservatie stimuleert ontwikkeling zelfsturend leren	<p>Ik betwijfel of de beoordelingscriteria op onze school gericht zijn op zelfsturend leren</p> <p><i>De beoordelaar geeft geleidelijk het initiatief bij de beoordeling over aan de docent:</i> Er kan aangegeven worden waar expliciet op gelet moet worden</p>
<i>Betekenisvolheid:</i> formatief doel, meer dan alleen meten	<p>Het is prettig om bevestiging te krijgen dat je het goed doet en om verbeterpunten te ontdekken en daar dan ook daadwerkelijk iets mee te doen!</p> <p>De lesobservatie ervaar ik als een "showing-off" moment. Leren doe je van feedback; wat neem je daar van aan, hoe komt wat je doet over en wil je daar iets aan veranderen.</p>



Tabel 9. Motivaties met onderbouwing bij kwaliteitscriteria performance assessments

Kwaliteitscriterium en indicatoren lesobservatie	Motivatie met persoonlijke ervaring als onderbouwing (level 1)
<i>Doelen:</i> aansluiting tussen docentcompetenties en beoordelingsactiviteiten	Er vindt vrijwel geen lesobservatie plaats bij mij. In al die jaren 1 à 2 keer. Er is geen lesobservatie geweest in de laatste 7 jaar. Wel 2 functioneringsgesprekken. Diverse competenties zijn niet aanbod gekomen tijdens de lesobservatie zoals toetsen maken, nakijken, organiseren, mailen. De formulieren zoals voorbeeld hierbij zitten, zijn bij mij nooit gebruikt op onze school
<i>Herhaalbaarheid:</i> oordelen gebaseerd op combinatie meerdere bewijzen	Mijn ervaring is dat een paar mensen 1x zo'n observatie invullen en dat is echt te weinig, vooral als het niet om een geweldige docent gaat. Overleg wordt niet zo georganiseerd dat sectiegenoot + afdelingsleider beide met de docent op hetzelfde tijdstip een gesprek hebben. Eerst oordeel, daarna bewijs wat daarbij aansluit. In mijn beleving ben ik nooit echt beoordeeld, en heb ik nooit (gelijkwaardig) overleg gevoerd met medebeoordelaars. Leerlingen komen soms toch wel tot een afwijkend oordeel.
<i>Acceptatie:</i> betrokkenen kunnen zich vinden in een lesobservatie	Het verslag is makkelijker te hanteren, maar splitst niet uit naar pedagogisch, didactisch en relationeel handelen. De wijze waarop een lesobservatie wordt uitgevoerd (procedure) gaat een beetje "houtje-touwtje". Er zou een formulier moeten zijn waarmee iedereen de les observeert met vaststaande criteria. Er zijn twee formulieren bijgevoegd. Er vind ook observatie plaats zonder formulieren. Voor onze school zou een nieuw instrument ontwikkeld moeten worden, waarin het "domeinwerken" is opgenomen.
<i>Zelfsturend leren:</i> CAP stimuleert ontwikkeling zelfsturend leren	<i>Wij, docenten beoordelen onszelf en elkaar door middel van lesobservaties:</i> dat doen we in de pauzes, tussen de lessen in gesprekken en vergaderingen. De beoordelaar geeft geleidelijk het initiatief bij de beoordeling: gebeurde bij mij wel in gesprek met afdelingsleider Toen ik nog op school zat formuleerde ik eigen leerdoelen op papier, nu in mijn hoofd. Vooral door observaties door iemand van buitenaf heb ik veel feedback gehad.
Betekenisvolheid: formatief doel, meer dan alleen meten.	Tot nu toe heb ik altijd wel goede feedback gekregen. Het hangt echter wel af van de observant wat voor soort feedback ik krijg. De een geeft meer vakdidactische feedback en de ander meer interpersoonlijk/pedagogisch.
<i>Professionele gevolgen:</i> effect op leren en instructie	Ik vind het jammer dat er niet altijd iets met beoordelingen gedaan wordt. Er lijken geen consequenties aan te zitten en het gebeurt te weinig. Dit geldt niet alleen voor docenten die misschien een wat negatieve beoordeling gekregen hebben. Ook docenten die goed functioneren moeten vaker aangezet worden tot het reflecteren op hun eigen handelen + lespraktijk. Dat kan mijn inziens tot positievere resultaten leiden.

#### 4. CONCLUSIE en DISCUSSIE

Dit onderzoek had het meten van de kwaliteit van een CAP voor docenten in het voortgezet onderwijs als doel waarin de perceptie van docenten en afdelingsleiders centraal stonden. Een single-casestudy is uitgevoerd ter illustratie voor scholen in het voortgezet onderwijs. De geselecteerde casus betrof een middelbare school in Nederland waarin het competentie-beoordelingsbeleid nog niet helemaal ligt verankerd in de school. Het raamwerk van kwaliteitscriteria voor CAP's van Baartman et al. (2006) is gebruikt als meetinstrument. Verschillende type data zijn gecombineerd om een representatief beeld te krijgen en om kwantitatieve data adequaat te kunnen interpreteren. Zo is er een documentanalyse uitgevoerd met betrekking tot het competentie-assessmentbeleid. Vervolgens heeft een survey plaatsgevonden waarin de gepercipieerde kwaliteit van het CAP van docenten en afdelingsleiders is gemeten. Dit resulteerde in kwantitatieve schaalscores met bijbehorende kwalitatieve motivaties. Daarnaast hebben docenten feedback kunnen geven over de begrijpelijkheid en duidelijkheid van de vragen. De data zijn door middel van datatriangulatie zowel kwantitatief als kwalitatief geanalyseerd.

*“Wat is de gepercipieerde kwaliteit van het CAP voor docenten in het voortgezet onderwijs in school X volgens docenten en afdelingsleiders”.*

In dit onderzoek blijkt dat de docenten van school X het niet volledig eens zijn met de kwaliteit van het CAP. Volledig ontevreden hierover zijn zij daarentegen ook niet. Dit niet eenduidige resultaat lijkt onbevredigend en weinig duidelijkheid te geven. Met name wanneer alleen een verklaring wordt gezocht in de gemiddelde schaalscores per kwaliteitscriterium. Uit de resultaten blijkt namelijk dat deze gemiddelden tussen twee (mee oneens) en drie (mee eens) liggen wat geen eenduidig resultaat oplevert. Echter, uit de resultaten van de beschrijvende statistiek blijkt dat er veelal sprake is van twee meerderheden; enerzijds docenten die het voornamelijk eens tot helemaal eens zijn met een stelling en anderzijds docenten die het oneens tot helemaal oneens zijn. Dit zou kunnen verklaren waarom de totaalscore “in het midden ligt”. Wanneer de schriftelijke motivaties nader bekeken worden, lijken deze enigszins ontevredenheid over de kwaliteit van het CAP te ondersteunen; de meerderheid van de motivaties hebben een negatieve lading. Er zijn echter kanttekeningen te plaatsen bij de generaliseerbaarheid van de verzamelde motivaties. Het merendeel van de docenten heeft de schaalscores namelijk niet gemotiveerd wat heeft geresulteerd in een lage respons. De motivaties zijn hierdoor niet representatief voor de docenten van school X. Alhoewel er op basis van de motivaties geen harde conclusies getrokken kunnen worden, zijn ze wel indicatief voor de gepercipieerde kwaliteit van het CAP. Daarnaast zijn ook kanttekeningen te plaatsen bij de kwaliteit van de motivaties. Volgens de typologie van Clark & Sampson (2005) is de meerderheid van de motivaties te classificeren als level 0 (motivatie zonder onderbouwing). De docenten geven vaker hun mening op een schaalscore dan dat zij een schaalscore onderbouwen door middel van bewijzen. In onderzoek van Baartman, Prins, Kirschner & van der Vleuten (2007) waarin een CAP van een MBO school is geëvalueerd, blijkt ook dat docenten moeite hebben met het ondersteunen van een stelling door middel van bewijzen. In vervolgonderzoek is het dan ook wenselijk om docenten expliciet te vragen om schaalscores te motiveren en om aandacht te besteden aan voorbeelden van typen bewijzen. Daarnaast is het wenselijk om tijdens het categoriseren van de motivaties volgens de typologie van Clark & Sampson (2005) een tweede beoordelaar te betrekken met het oog op de interbeoordelaarsbetrouwbaarheid.

Met betrekking tot de totale kwaliteit van het CAP zijn de docenten en de afdelingsleiders het met elkaar eens. Ook de afdelingsleiders zijn het niet volledig eens met de kwaliteit van het CAP. De afdelingsleiders verschillen daarentegen wel van mening met de docenten ten aanzien van de kwaliteit per kwaliteitscriterium, met uitzondering van het kwaliteitscriterium betekenisvolheid. Zo zijn de afdelingsleiders het minst eens met het zelfsturend karakter van het CAP, de cognitieve complexiteit en het kwaliteitscriterium Tijd & Kosten. Het meest eens zijn zij met de professionele gevolgen, de betekenisvolheid en de rechtvaardigheid van het CAP. Dit resultaat is echter niet representatief omdat slechts twee van de vijf afdelingsleiders deel genomen hebben aan het

onderzoek. De resultaten geven daarentegen wel een indicatie van de perceptie van de afdelingsleiders ten aanzien van het CAP. De docenten zijn het minst eens met de vergelijkbaarheid, de herhaalbaarheid en de rechtvaardigheid van het CAP. Daarentegen zijn zij het meest eens met de kwaliteitscriteria acceptatie, transparantie en betekenisvolheid ten opzichte van de andere kwaliteitscriteria.

Algemeen genomen zijn de docenten het minst eens met de consistentie van het CAP en de wijze waarop het CAP wordt uitgevoerd. Het kwaliteitscriterium *vergelijkbaarheid* is het minst gekwalificeerd ten opzichte van de andere kwaliteitscriteria. Echter zijn de docenten ten opzichte van andere kwaliteitscriteria hierover het minst eensgezind. Zo vindt een kleine minderheid dat zij op basis van vergelijkbare competenties beoordeeld worden. Ongeveer de helft van de docenten is het hier niet mee eens. Resultaten uit de documentanalyse ondersteunen dit. Eén afdeling hanteert namelijk eigen docentcompetenties in vergelijking tot de andere afdelingen. Ook een docent betwijfelt of bij elke beoordeling dezelfde docentcompetenties gehanteerd worden zo blijkt uit de motivaties. Een kleine meerderheid van de docenten vindt daarnaast dat de beoordelingscriteria voor alle docenten vergelijkbaar zijn, iets minder dan de helft weet het niet of is het hier niet mee eens. Resultaten uit de documentanalyse bevestigen dat niet alle afdelingen dezelfde beoordelingscriteria hanteren. Het meest eensgezind zijn de docenten over de vergelijkbaarheid van de beoordelingsactiviteiten. Meer dan de helft van de docenten weet niet of de beoordelingsactiviteiten voor alle docenten vergelijkbaar zijn of zijn het hier niet mee eens. Alhoewel het lage gemiddelde van het kwaliteitscriterium onderbouwd wordt door bovengenoemde resultaten, zou een deel ook verklaard kunnen worden door het niet adequaat interpreteren van een stelling, zo blijkt uit de schriftelijke motivaties. Het merendeel van de docenten (80%) was het oneens met de stelling. De vraag is of dit kwaliteitscriterium het minst gekwalificeerd blijft, indien de stelling wel adequaat wordt geïnterpreteerd.

Na de consistentie en de uitvoering van het CAP zijn de docenten het minst eens met de wijze waarop oordelen en beslissingen ten aanzien van docentcompetenties tot stand komen. Het kwaliteitscriterium *herhaalbaarheid* heeft de meeste motivaties gegenereerd en tegelijkertijd de meest negatieve, al dan wel of niet onderbouwd. Dit zou het belang dat docenten hechten aan dit kwaliteitscriterium kunnen aantonen of de mate van ontevredenheid. In de omschrijving van de procedures van competentiebeoordeling en ontwikkeling staat dan ook niets vermeld over de wijze waarop beoordelingen tot stand komen, zo blijkt uit de documentanalyse. Het enige positieve aspect met betrekking tot het kwaliteitscriterium herhaalbaarheid is dat het merendeel van de docenten van mening is dat het eindoordeel over een docentcompetentie gebaseerd is op meerdere verschillende bewijzen, zoals resultaten uit lesobservaties en leerlingenenquêtes. Echter zijn zij niet eensgezind of de verschillende bewijzen worden gecombineerd. Aansluitend is ongeveer de helft van de docenten van mening dat er geen gelijkwaardig overleg plaats vindt waarin de beoordelaars hun oordeel onderbouwen. Ook zijn zij van mening dat de beoordelaars niet onafhankelijk van elkaar tot hetzelfde oordeel komen. Uit een schriftelijke motivatie blijkt zelfs dat voor zowel de beoordeelde als de beoordelaars vaak onduidelijk is tot welk oordeel de schoolleiding komt.

Het kwaliteitscriterium *acceptatie* is één van de drie kwaliteitscriteria waar de docenten het meest mee eens zijn. Geconcludeerd kan worden dat een meerderheid van de docenten zich kan vinden in de beoordelingscriteria en in de uit te voeren beoordelingsactiviteiten. De helft van de docenten denkt dat de beoordelaars zich kunnen vinden in de inhoud, criteria en uitvoering van het CAP, de andere helft geeft aan dit niet te weten. Het 'twee kampen fenomeen' wordt ook zichtbaar in de stelling of docenten zich kunnen vinden in de wijze waarop het CAP wordt uitgevoerd. De helft van de docenten is het hier mee eens tot helemaal eens en de andere helft is het hier mee oneens tot helemaal oneens. De twee schriftelijke motivaties hebben voornamelijk betrekking op de uitvoering van het CAP waarin een adequate uitvoering van het CAP wordt betwijfeld.

De resultaten wijzen daarnaast uit dat het CAP een significante waarde heeft voor zowel de docenten als de afdelingsleiders. Drie kwart van de docenten is het namelijk volledig eens met de

*betekenisvolheid* van een CAP. Het merendeel van de docenten vindt dat de beoordelingscriteria een goede basis vormen om feedback op te baseren, dat de beoordelaars zinvolle feedback geven op basis van het CAP en dat feedback op de beoordelingen zinvol zijn. Daarnaast ervaart de meerderheid van de docenten het uitvoeren van beoordelingsactiviteiten als leermoment. De resultaten van de documentanalyse laten zien dat op basis van kenmerken van goed assessmentbeleid (Ketelaar, 2008) er enigszins nagedacht is over de opzet en uitvoering van een competentie assessment programma. Met uitzondering van één kenmerk: verbinding competentie assessmentbeleid met strategie school, zijn er richtlijnen in de vorm van procedures en protocollen beschreven. Hierbij dient wel een kanttekening geplaatst te worden. De selectie van de documentatie is namelijk overgelaten aan de afdelingsleiders en aan de schoolleiding. Het risico hiervan is dat niet alle relevante documenten geselecteerd hoeven te zijn. Of dat documenten juist selectief gekozen zijn (Yin, 2009). Een documentanalyse limiteert zich daarnaast tot het verzamelen en analyseren van documentatie over ontwikkeld beleid, beschrijvingen van procedures en protocollen. In een vervolgonderzoek zou de uitvoering en implementatie van beschreven procedures en beleid ook een focus kunnen zijn. Hierdoor wordt inzichtelijk gemaakt in hoeverre in de praktijk tegemoet wordt gekomen aan de kenmerken van goed assessmentbeleid. Hoewel dit niet is onderzocht in dit onderzoek geeft de uitgevoerde documentanalyse wel een indicatie in hoeverre het competentie-assessmentbeleid is doordacht.

Dit onderzoek heeft tevens een methodiek geëxploreerd om de gepercipieerde kwaliteit en invloed van afzonderlijke assessments ten opzichte van een CAP vast te stellen. Hierbij waren performance assessments exemplarisch voor de afzonderlijke assessments. Ten eerste is onderzocht wat de gepercipieerde kwaliteit is van performance assessments volgens docenten. Vervolgens is de relatie van de gepercipieerde kwaliteit van performance assessment ten opzichte van de gepercipieerde kwaliteit van het CAP berekend door de proportie verklaarde variantie vast te stellen.

*“In welke mate beïnvloeden performance assessments als onderdeel van het CAP voor zittende docenten op school X, volgens docenten, de totale kwaliteit van het CAP?”*

De resultaten wijzen uit dat meer dan de helft van de totale kwaliteit van een CAP (56%) wordt bepaald door de totale kwaliteit van performance assessments. Dit impliceert dat de docenten op school X veel waarde hechten aan deze vorm van beoordelen om docentcompetenties te meten. Uit de resultaten blijkt dat performance assessments een significante waarde hebben voor de docenten op school X. Het kwaliteitscriterium betekenisvolheid is namelijk het enige kwaliteitscriterium waar de docenten het volledig mee eens zijn. Uit onderzoek van Jonsson, Baartman & Lennung (2009) blijkt dat performance assessments zowel geschikt zijn voor zowel summatieve als formatieve doeleinden en hiermee een cruciale bijdragen leveren aan de beoordeling van docentcompetenties. Dit zou de invloed van performance assessments op de totale kwaliteit van het CAP kunnen verklaren. Er dienen echter wel kritische kanttekeningen geplaatst te worden bij de totstandkoming van het bovengenoemde resultaat. Een aantal docenten bleek bij het kwalificeren van de totale kwaliteit van het CAP niet alle onderdelen van het CAP in gedachte te hebben maar alleen performance assessments. In onderzoek van Baartman et al. (2007) waarin een CAP voor MBO-studenten werd gekwalificeerd door docenten blijkt ook dat zij moeite hadden om het gehele CAP in gedachte te nemen. Dit impliceert dat deze groep docenten de twee vragenlijsten vrijwel identiek ingevuld hebben waardoor een sterk verband tussen de totale kwaliteit van een CAP en de kwaliteit van performance assessments in de hand wordt gewerkt. Bovendien zou de vraag gesteld kunnen worden of performance assessments herhaaldelijk meer dan de helft van de kwaliteit van een CAP bepalen indien er meerdere beoordelingsactiviteiten meegenomen worden in een vervolgonderzoek bijvoorbeeld: self-assessments, portfolio assessments en leerlingenenquêtes. Het zou interessant zijn om te onderzoeken in welke mate andere beoordelingsactiviteiten/ methoden bepalend zijn voor de totale kwaliteit van een CAP. Met betrekking tot de methodiek zou in plaats van een enkelvoudige

regressieanalyse een multiple regressieanalyse toegepast moeten worden. In het kader van generaliseerbaarheid is het daarnaast van belang om in plaats van een single-casestudy, een multiple casestudy uit te voeren. Gestreefd zou moeten worden naar een grotere steekproef. Ook zou een correlatie tussen de kwaliteit van een CAP en performance assessments nader onderzocht kunnen worden. Indien performance assessments goed worden gekwalificeerd, zou dit vervolgens een kwalitatief goede CAP inhouden. Indien een daadwerkelijke correlatie blijkt, zou een investering in de kwaliteit van performance assessments een wezenlijke bijdrage leveren aan de kwaliteit van een CAP en daarmee ook aan de beoordeling van docentcompetenties.

In dit onderzoek is het meetinstrument van Baartman et.al (2006) gebruikt om de gepercipieerde kwaliteit van zowel een CAP te meten als van performance assessments. Alhoewel dit onderzoek niet gericht was om dit instrument specifiek voor deze context te valideren, zou dit wel een meerwaarde opleveren. De resultaten uit dit onderzoek geven hier namelijk aanleiding toe. Zo blijkt uit de feedback over de duidelijkheid en de begrijpelijkheid van het meetinstrument. Met betrekking tot de procedurele validiteit zou het wenselijk zijn om in vervolgonderzoek de kwaliteitscriteria toe te lichten aan docenten en afdelingsleiders. Een bijeenkomst zou georganiseerd kunnen worden om de stellingen door te nemen voorafgaand aan het afnemen van de kwantitatieve survey. Onduidelijkheden kunnen hierdoor worden weggenomen. Daarnaast zou in een dergelijke bijeenkomst benadrukt kunnen worden dat alle onderdelen van een CAP in ogenschouw genomen moeten worden bij de kwalificering van de totale kwaliteit van het CAP.

## Referenties

- Australian College of Educators (2000). *Standards of professional practice for accomplished teaching in Australian classrooms*. Gevonden op 12 maart 2010, op [www.austcolled.com.au/Disper.pdf](http://www.austcolled.com.au/Disper.pdf).
- Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & van der Vleuten, C. P. M. (2006). The wheel of competency assessment: presenting quality criteria for competency assessment programs. *Studies in Educational Evaluation*, 32, 153-170.
- Baartman, L. K. J., Prins, F. J., Kirschner, P. A., & van der Vleuten, C. P. M. (2007). Determining the quality of competence assessment programs: A self-evaluation procedure. *Studies in Educational Evaluation*, 33, 258-281
- Black, P., & William, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5, 7-75.
- Bennet, Y. (1993). The validity and reliability of assessments and self-assessments of work-based learning. *Assessment & Evaluation in Higher Education*, 18, 83-95.
- Boeije, H. (2005). *Analyseren in kwalitatief onderzoek. Denken en doen*. Hoofddorp: Boomonderwijs.
- Brown, G. T. L. (2004). Teachers' conceptions of assessment: implications for policy and professional development. *Assessment in Education*, 11, 301-318.
- Clark, D. B., & Sampson, V. D. (2005). Analyzing the quality of argumentation supported by personally-seeded discussions. In Baartman, L. K. J., Prins, F. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2007). Determining the quality of competence assessment programs: a self-evaluation procedure.
- Cronbach, L. J. (1989). Construct validation after thirty years. In R. E. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 147-171). Urbana: University of Illinois Press.
- Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement*, 57, 373-399.
- Crooks, T. J., Kane, M. T., & Cohen, A. S. (1996). *Assessment in Education: Principles, Policy & Practice*, 3, 265-285.

- Darling-Hammond, L. (2001). Standard setting in teaching: Changes in licensing, certification, and assessment. In V. Richardson (Ed.), *Handbook of research on teaching* (4<sup>th</sup> ed.). Washington: American Educational Research Association.
- Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessment. *Applied Measurement in Education*, 4, 289-303.
- Hickey, D. T., Zuiker, S. J., Taasobshirazi, G., Schafer, N., & Michael, M. A. (2006). Balancing varied assessment functions to attain systemic validity: Three is the magic number. *Studies in Educational Evaluation*, 32, 180-201.
- Gipps, C. V. (1994). *Beyond Testing. Towards a theory of educational assessment*. New York: Routledge Falmer.
- Jonsson, A., Baartman, L. K. J., & Lennung, S. A. (2009). Estimating the quality of performance assessments: The case of an 'interactive examination' for teacher competencies. *Learning Environments Research*, 12, 225-241.
- Kane, M.T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527-535.
- Landelijk Platform Beroepen in het Onderwijs (2010). *Bekwaamheidseisen in de school. Op weg naar versterking beroepskwaliteit docenten*. Utrecht.
- Linn, R. L., Baker, J., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, 20, 15-21.
- Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18, 5-11.
- Ministerie van Onderwijs, Cultuur en Wetenschap (2007). Actieplan Leerkracht van Nederland. Beleidsreactie op het advies van de commissie docenten. Scheveningen.
- National Board of Professional Teaching Standards (2002). *What teachers should know and be able to do*. Arlington: NBPTS. <http://www.nbpts.org/pdf/coreprops.pdf>
- Sultana, R. G. (2009). Competence and competence frameworks in career guidance: complex and contested concepts. *International Journal for Educational and Vocational Guidance*, 9, 15-30.
- Tucker, P. D., Stronge, J. H., Gareis, C. R., & Beers, C. S. (2003). The efficacy of portfolios for teacher evaluation and professional Development: Do they make a difference? *Educational Administration Quarterly*, 39, 572-602.
- Van Merriënboer, J. J. G., Van der Klink, M. R., & Hendriks, M. (2002). Competenties: van complicaties tot compromis. Den Haag: Onderwijsraad.
- Yin, R. K. (2009). *Case Study Research. Design and Methods*. California, USA: SAGE Inc.

