Universiteit Utrecht, Utrecht

Masterthesis

Subjectiviteit van assessmentbeoordelingen

en mogelijke invloeden op de beoordelingen

Schölvinck, B. A.

Juni 2010

Faculteit Sociale Wetenschappen
Universiteit Utrecht

Brechje Schölvinck *(3223 019)*

In opdracht van:
Hogeschool Domstad, te Utrecht

Interne begeleider: Dr. F. J. Prins
Externe begeleider: Drs. J. H. Nuijten

Eerste beoordelaar: Dr. F. J. Prins
Tweede beoordelaar: Drs. S. Werdmuller von Elgg

Schölvinck, B. A.
FSW Universiteit Utrecht

Subjectiviteit van assessmentbeoordelingen

en mogelijke invloeden op de beoordelingen

Schölvinck, B. A.

Samenvatting

In 2007 is de onderwijsstructuur van Hogeschool Domstad te Utrecht drastisch veranderd, waarbij men is overgegaan van aanbodgestuurd naar vraaggestuurd onderwijs. Deze ontwikkeling heeft ook invloed gehad op de beoordelingsmethode. In dit onderzoek is gekeken naar de objectiviteit van de nieuwe beoordelingsmethode en aspecten die mogelijk invloed uitoefenen op de betrouwbaarheid en validiteit van de beoordeling. De achtergrond van de assessor (beoordelaar), het moment dat de assessor zijn oordeel vormt en de vorm van het beoordelingsformulier, zijn invloeden die onderzocht zijn. Door middel van de Hardop-denk-methode zijn overwegingen tijdens het beoordelingsproces van assessoren uitgesproken. De resultaten hiervan zijn een aanvulling op de verkregen kwantitatieve data. Uit het onderzoek is geen aanwijzing gekomen dat de assessmentbeoordelingen subjectief verlopen. De resultaten worden besproken en er worden suggesties gedaan voor verder onderzoek.

1. Inleiding

Hogeschool Domstad te Utrecht leidt studenten op tot leraar in het basisonderwijs. In 2007 heeft er een drastische wijziging van de opleidingsstructuur plaatsgevonden, waarbij ook de beoordelingsmethode veranderd is. De onderwijsstructuur van de Hogeschool is gewijzigd van aanbodgestuurd onderwijs naar vraaggestuurd onderwijs. Studenten werken in het huidige systeem gedurende de laatste twee jaar van de opleiding aan eigen leervragen. Deze leervragen zijn enerzijds gestoeld op een competentiematrix, anderzijds worden deze gevormd door de persoonlijke (stage)situatie van de student. De competentiematrix is een schematisch weergegeven leerlijn die de student volgt in de opleiding tot een startbekwame leraar voor het basisonderwijs.

Dit onderzoek richt zich op de beoordelingsmethode, waarbij specifiek gekeken is naar de afwegingen die assessoren, beoordelaars, maken bij het beoordelen van materiaal van de student, het moment waarop zij hun mening vormen en het materiaal dat zij bij de beoordeling gebruiken. Mogelijk hebben de resultaten

van dit onderzoek in de toekomst invloed op de beoordelingsmethode, aangezien objectieve en kwalitatief hoogstaande beoordelingen een belangrijk aspect vormen in de beoogde opleidingsstructuur van Hogeschool Domstad.

2. Theoretische achtergrond

2.1. Veranderingen in het onderwijs

De onderwijsontwikkelingen bij Hogeschool Domstad staan niet op zich zelf. De wijze van beoordelen binnen het onderwijs is de afgelopen decennia sterk veranderd. Niet alleen de frequentie van beoordelen is toegenomen (Myers & Myers, 2007), maar ook de manier van beoordelen heeft grote veranderingen ondergaan. Dit houdt over het algemeen in dat men overgegaan is van een test- en examineercultuur (alleen gericht op het eindproduct) naar een assessmentcultuur (zowel op het proces als op het product gericht) (Gipps, 1994). Ten tijde van de test- en examineercultuur lag de nadruk op kennisoverdracht en werd het eindresultaat getoetst (Cluitmans & Klarus, 2005), terwijl bij assessments niet alleen gekeken wordt naar het eindproduct van de lerende, maar daarnaast ook naar het gehele leerproces. Binnen dit leerproces kan de nadruk op verschillende doelen liggen, zoals het aanmoedigen van leren en het verschaffen van informatie over prestaties. De vorm en de wijze van beoordelen van het assessment wordt door het doel van het leerproces bepaald (Gipps, 1994). Een voordeel is dat lerenden met verschillende leerstijlen (niet uitsluitend gericht op kennisoverdracht) de mogelijkheid krijgen op hun eigen wijze te leren (Cluitmans & Klarus, 2005). Stiggins en Bridgeford (1985) signaleren echter voor het eerst in een artikel dat er problemen bij het toepassen van assessments kunnen ontstaan. Bij de beoordeling ontstaat vaak de lastige situatie dat het resultaat van het assessment vergeleken moet kunnen worden met andere resultaten en dus een betrouwbaar beeld moet kunnen geven van de uitkomst van het leerproces (Van der Schaaf, Stokking & Verloop, 2008). Daarom wordt inmiddels het assessment vaak opgesplitst in een gedeelte om de lerende te begeleiden tijdens het leerproces (formatief assessment) en om de lerende te toetsen aan het eind van het leerproces (summatief assessment) (Cluitmans & Klarus, 2005; Gipps, 1994). Door de opsplitsing in formatieve en summatieve assessments zijn ook de functies van personen die de assessments afnemen, aan het veranderen. Een assessor bij een formatief assessment zal zich eerder begeleidend dan toetsend

opstellen. Deze assessor zal de lerende begeleiden in de periode tot het summatieve assessment. De assessor die aan het eind van het leerproces van de lerende toetsend zal optreden, spreekt alleen een oordeel uit over het eindresultaat van het proces en het product (Gipps, 1994). Belangrijk is hierbij wel dat het assessment uit meervoudig bewijsmateriaal bestaat, zodat de assessor een compleet beeld van de lerende en van zijn of haar leerproces krijgt (Cluitmans & Klarus, 2005).

Toch lijkt uit onderzoek het beeld te ontstaan dat de opsplitsing van het assessment in een formatief en summatief deel, en de eis dat het assessment uit meervoudig bewijsmateriaal bestaat, niet voldoende is om de kwaliteit rondom het nieuwe toetsen te waarborgen. Naast problemen, voorwaarden en facetten bij het meten van competenties, lijken ook eigenschappen van assessoren en standaarden in beheersingsniveaus bij het meten van invloed op de kwaliteit te zijn.

2.2. Problemen, voorwaarden en facetten bij het meten van competenties

Luken (2004) richt zich op wat er in het onderwijs gemeten wordt. Uit zijn betoog vloeit het dilemma voort: “We kunnen niet meten wat we in competentiegericht onderwijs zouden willen meten.” (Luken, 2004, p. 38). Hiermee bedoelt hij dat het welhaast onmogelijk is lerenden objectief te toetsen, aangezien het resultaat moeilijk vergeleken kan worden met paralleltoetsen. Dit wordt eveneens onderschreven door Van der Schaaf en collega’s (2008). Het is een nieuwe ontwikkeling in het onderwijs dat de lerende werkt aan (en getoetst wordt op) de beheersing van competenties, waarbij het begrip competentie omschreven kan worden als “een vermogen dat kennis-, houdings- en vaardigheidsaspecten omvat, om in concrete taaksituaties doelen te bereiken.” (Luken, 2004, p. 39). Al lijken de aspecten kennis, houding en vaardigheid zeer van elkaar te verschillen, zowel Luken (2004) als Van Merriënboer, Van der Klink en Hendriks (2002) benadrukken dat deze verschillende aspecten ondeelbaar de kern van het competentiebegrip bevatten. Eraut (1994) en Gonczi (1994) stellen dat naast deze drie aspecten binnen competenties ook persoonskenmerken invloed hebben op het uitvoeren van professionele taken in een beroepsmatige context. Cluitmans en Klarus (2005) richten zich derhalve op deze beroepsmatige context; zij stellen dat competenties in wezen taaksituaties zijn naar de vraag van de beroepspraktijk. In dit artikel worden de persoonskenmerken buiten beschouwing gelaten; de beroepsmatige context wordt wel meegenomen.

Klarus (1998), Luken (2004) en Messick (1989) hebben zich gebogen over de manier waarop competenties beoordeeld moeten worden. Daarbij heeft Klarus (1998) aandacht besteed aan de voorwaarden die van toepassing zijn op de validiteit van de beoordeling. Messick (1989) richt zich eveneens op de validiteit; hij splitst deze op in zes facetten. Daarentegen voorziet Luken (2004) juist een achttal problemen bij de beoordeling. Logischerwijs grijpen de problemen, de voorwaarden en de facetten in elkaar en kunnen deze met elkaar vergeleken worden. Met behulp van de voorwaarden en de facetten worden de problemen getoetst op hun houdbaarheid. Met de vergelijking kan een afweging gemaakt worden of deze manier van toetsen werkelijk betrouwbaar, valide en zinvol is. Deze vergelijking zal in dit artikel gemaakt worden.

Het eerste probleem dat Luken (2004) voorziet, is dat competentie geen homogeen begrip is; “als een meetinstrument uit verschillende items bestaat en deze items bij een meting verschillende kanten opwijzen, is het onmogelijk om een betrouwbare meting te doen,” aldus Luken (2004, p. 39). Volgens Van der Schaaf en collega’s (2008) verwijst betrouwbaarheid naar de stabiliteit bij herhaalde metingen en de consistentie tussen de beoordelingen. Wanneer de items binnen een meting verschillende kanten opwijzen, zoals Luken (2004) beschrijft, kunnen zij dit bij meerdere metingen achter elkaar doen en is er volgens de definitie van betrouwbaarheid van Van der Schaaf en collega’s (2008) wel sprake van betrouwbaarheid. Het is de vraag of Luken (2004) door te wijzen naar de betrouwbaarheid wel de vinger op de juiste plek legt en of hij niet eigenlijk de validiteit bedoelt, aangezien validiteit zich richt op de vraag of de beoordeling de beoogde constructen meet en niet iets anders. De validiteit van competentiebeoordelingen zal nog besproken worden aan de hand van de voorwaarden van Klarus (1998) en de validiteitfacetten van Messick (1989). Naast de vraag die Luken (2004) stelt of het überhaupt mogelijk is om een betrouwbare meting te doen, benoemt hij in hetzelfde artikel toch dat de aspecten kennis, houding en vaardigheden de kern van het competentiebegrip bevatten en ondeelbaar zijn. Kuijpers (2003) adviseert daarentegen in haar promotieonderzoek om competenties die vermogen en gedrag definiëren niet binnen één competentie op te nemen. Zij ziet geen empirische basis voor het vasthouden aan de ondeelbaarheid van de definitie van competenties. Volgens Van der Schaaf en collega’s (2008) woedt er momenteel een discussie om de hoge kwaliteitseisen wat betreft betrouwbaarheid en validiteit los te laten, waarmee de definitie van competentie weer geheel ter discussie komt te staan. Van der Schaaf en collega’s (2008) houden overigens vast aan de

hoge betrouwbaarheids- en validiteiteisen waaraan beoordelingen moeten blijven voldoen. Kortom, er is tot op heden geen consensus over de ondeelbaarheid van de definitie van het begrip competentie en hieruit voortvloeiend staat ook de betrouwbaarheid van de metingen ter discussie.

De volgende drie problemen die Luken (2004) signaleert, zijn gericht op de relatie tussen onderwijs en arbeid; (1) *Competentiebeoordelingen zijn gekoppeld aan personen, terwijl competentiebegrip ook de context omvat*, (2) *Onderwijs is een andere context dan arbeid* en (3) *Competentiebeoordelingen zijn in het onderwijs niet zo zinvol*. Wat betreft het eerste probleem vraagt Luken (2004) zich af of het mogelijk is om aan de tegenstrijdige eisen te voldoen: enerzijds moeten de competenties voldoende algemeen geformuleerd worden om de context te kunnen omvatten, anderzijds moeten zij ook voldoende specifiek zijn om in situaties werkelijk geldig te kunnen zijn. Vervolgens benadrukt Luken (2004) dat onderwijs een heel andere context dan arbeid is en dat een lerende pas werkelijk competent is als deze 'echte' doelen kan bereiken, doelen waar het om gaat in de realiteit van het beroep. In het onderscheid dat Elshout-Mohr en Oostdam (2001) maken, waarbij het gaat om een holistische en analytische benadering, is de algemene en specifieke formulering van Luken (2004) terug te zien. De arbeidsmarkt gaat volgens Elshout-Mohr en Oostdam (2001) meer uit van de gehele context, terwijl het onderwijs meer analytisch verschillende onderdelen apart onder de aandacht zou brengen bij de lerenden. Luken (2004) gaat er vanuit dat beide benaderingen, zowel algemeen als specifiek (zowel holistisch als analytisch) niet tegelijk mogelijk zijn en trekt dus de conclusie dat competentiebeoordelingen niet zinvol zouden zijn in het onderwijs. Toch is de verandering in examineercultuur mede gestart door het feit dat het onderwijs niet goed genoeg op de arbeidsmarkt aansloot en startende werknemers voor verrassingen kwamen te staan waar zij niet voor opgeleid waren. Zowel Klarus (1998) als Van der Schaaf en collega's (2008) onderbouwen een ander standpunt dan de mening van Luken (2004). Klarus (1998) heeft in zijn voorwaarden eveneens aandacht besteed aan de relatie tussen de omgeving en het doel van onderwijs; *authentieke beroepssituaties*. Hij geeft daarentegen aan dat het wel degelijk mogelijk is beide aspecten te betrekken, op voorwaarde dat de omgeving belangrijk is, maar dat generalisatie mogelijk moet blijven. Klarus (1998) legt dus de nadruk op het specifieke en analytische aspect van respectievelijk Luken (2004) en Elshout-Mohr en Oostdam (2001). Net zoals Klarus (1998) wijzen Van der Schaaf en collega's (2008) er op dat gegevens vaak uniek, situationeel, persoonlijk en beschrijvend zijn, waardoor het noodzakelijk is het

materiaal eerst te interpreteren voordat het beoordeeld wordt. De consequentie hiervan is dat de betrouwbaarheid en de validiteit moeilijk te bepalen zijn. Van der Schaaf en collega's (2008) hebben daarom gekozen voor het verzamelen van meerdere gegevens. Dit ligt in de lijn van wat Gipps (1994) en Cluitmans en Klarus (2005) adviseren, namelijk de opsplitsing van een beoordeling in een formatief en summatief deel. Hierbij wordt aandacht besteed aan de theoretische en praktische kant van de leerstof en de wens van Cluitmans en Klarus (2005) dat het assessment uit meervoudig bewijsmateriaal bestaat, zodat de assessor een compleet beeld van de lerende en zijn of haar leerproces krijgt. Uit de beschreven literatuur is gebleken dat deze probleemstelling van Luken (2004) door verschillende auteurs is behandeld. Uiteindelijk is de benadering van de context (holistisch dan wel analytisch) cruciaal in het beoordelingsproces.

De noodzaak tot interpreteren (Van der Schaaf et al., 2008) stuit op een volgend probleem – mogelijk het grootste en meest overkoepelende probleem– van Luken (2004); *Competentiebeoordelingen zijn subjectief*. Luken (2004) somt verschillende bedreigingen op, zoals persoonlijke waardeoordelen, voorkeuren, vooroordelen en afhankelijkheidsrol van de beoordelaar binnen de beoordeling. Deze bedreigingen hebben invloed op de mate van overeenkomst tussen de beoordelingen van verschillende mensen. Toch benadrukt Luken (2004) zelf dat hiermee niet gezegd is dat competenties niet beoordeeld kunnen worden. Vier van de zeven voorwaarden die Klarus (1998) heeft opgesteld om de validiteit van competentiebeoordelingen te waarborgen, raken dit probleem. Deze voorwaarden; (1) *het vermogen beoordeeld te worden*, (2) *congruentie*, (3) *doorzichtigheid* en (4) *communicatie*, zorgen ervoor dat er zicht is op het proces dat tot de uiteindelijke beoordeling leidt. Ook twee van de zes facetten van Messick (1989) zijn verwant aan het probleem van Luken (2004); (1) *construct* en (2) *proces*. Het vermogen beoordeeld te worden gaat enerzijds uit van de vraag of de lerende zelf bewijzen bij de competenties zoekt en aandraagt of dat de beoordelaar de bewijzen selecteert. Anderzijds kan men zich afvragen of het überhaupt mogelijk is bewijs te vinden, waarmee aangetoond wordt dat de competentie beheerst wordt. Hierbij kan de subjectiviteit van een beoordelaar (Luken, 2004) nog steeds een rol spelen. Wat betreft congruentie en doorzichtigheid: de beoordelingsmethode moet aansluiten op het competentiedomein en de vereisten voor de competentie en de procedure bij de beoordeling moet voor alle betrokkenen helder en duidelijk zijn. Beide vereisten zijn zowel terug te vinden in de voorwaarden van Klarus (1998) als bij de validiteitfacetten van Messick (1989). De subjectiviteit van de beoordelaar (Luken, 2004) wordt door deze twee voorwaarden

ingeperkt. De laatste voorwaarde van Klarus (1998), *communicatie*, komt overeen met de noodzaak tot interpreteren van Van der Schaaf en collega's (2008); door middel van interactie tussen assessor en beoordeelde moet een beeld geschapen of verhelderd worden van de beheersing van de competentie door de lerende, waarna de beoordeling kan plaatsvinden. Dit wordt eveneens beschreven door Stokking, Van der Schaaf, Jaspers en Erkens (2004) in een van de suggesties om de kwaliteit van de beoordeling te verbeteren; *de criteria van de beoordeling moeten expliciet gemaakt worden voor de studenten*. Om terug te komen op het probleem van subjectieve competentiebeoordelingen dat Luken (2004) voorziet, wordt dit met de voorwaarden die Klarus (1998) stelt, over het algemeen teniet gedaan. Wanneer meerdere malen door verschillende mensen die hetzelfde materiaal beoordelen dezelfde beoordelingen worden gegeven, dan kan men wellicht toch spreken van objectiviteit in de beoordeling. Interessant is om dan te achterhalen met welke overwegingen de beoordelaars tot eenzelfde beoordeling komen. Dit wordt nader besproken als het gaat om verschillende standaarden in beheersingsniveaus van de beoordelingen.

Los van eerdere problemen van Luken (2004), de voorwaarden die Klarus (1998) stelt en de facetten van Messick (1989) om de validiteit te waarborgen, waarschuwt Luken (2004) voor nog een probleem: *Competentiebegrip gaat om 'vermogen', terwijl het belangrijker is of de prestatie werkelijk geleverd wordt*. Hij beschrijft de discrepantie tussen de belangen van de beoordeelde en van de beoordelaar, waarbij de beoordeelde een korte tijd kan 'doen alsof hij het gevraagde beheerst' om op die manier de gewenste beoordeling te halen. Dit probleem wordt echter weggenomen wanneer er sprake is van een formatieve en summatieve beoordeling (Cluitmans & Klarus, 2005; Gipps, 1994) en meervoudig bewijsmateriaal (Cluitmans & Klarus, 2005). Stokking en collega's (2004) suggereren eveneens de splitsing tussen formatieve en summatieve beoordeling; "Er moet een duidelijk onderscheid gemaakt worden door de leraar tussen formatieve en summatieve assessments." (Stokking et al., 2004, p. 97). De verschillende belangen van de lerende en de assessor, kunnen deels ondervangen worden door het gebruik van meervoudig bewijsmateriaal en het instellen van formatieve en summatieve assessments. Dit zal de omvang en urgentie van het probleem van Luken (2004) tot het minimum reduceren.

Competenties zijn niet stabiel en *Als men toch competenties probeert te meten dreigt terugval* zijn de laatste twee problemen die Luken (2004) signaleert bij de beoordeling van competenties. Van Merriënboer en collega's (2002) bevestigen de instabiliteit van competenties. Zij stellen dat een

competentie uit zes kenmerken bestaat. Een van deze kenmerken is *veranderlijkheid*. Hierdoor is het sowieso onmogelijk dat competenties stabiel kunnen zijn. Mede daarom is het noodzakelijk dat ook het *proces*, een van de validiteitfacetten van Messick (1989), meegenomen wordt in de beoordeling. Het tweede probleem omvat de angst dat tijdens een beoordeling teruggegrepen wordt naar het afvinken van vaardigheden aan de hand van vastgestelde criteria in contextarme situaties. Met behulp van het competentiesysteem werd juist geprobeerd om uit deze vorm van beoordelen te ontsnappen, mede om de aansluiting van het onderwijs tot de arbeidsmarkt te versoepelen. Een van de criteriumgerelateerde voorwaarden van Klarus (1998) besteedt hier aandacht aan: de criteria moeten vergeleken zijn met de vooraf gestelde kwaliteitseisen. Toch benadrukt Eraut (1994) dat deze criteria niet te strak en star moeten zijn, omdat dit amateuristisch beoordelen uitlokt. Daarnaast zal dan de unieke leersituatie en ontwikkeling van de lerende niet meer centraal staan. Van der Schaaf en collega's (2008) vinden het cruciaal dat duidelijk is welke eisen gesteld worden en bij de beoordeling eerst geïnterpreteerd wordt, voordat er beoordeeld wordt. De aandacht die uitgaat naar het scoren bij de beoordeling behoort eveneens tot een van de zes validiteitfacetten van Messick (1989). De stabiliteit van competenties kan dus alléén gewaarborgd worden indien ook het proces wordt meegenomen in de beoordeling. De valkuil dat competentiebeoordelingen dreigen af te glijden tot simpele afvinklijstjes, wordt hiermee voorkomen.

Ook Crooks en Kane (1996) hebben gericht gekeken naar de bedreigingen van de validiteit die kunnen optreden bij de beoordeling van assessments. Zij stellen dat een assessment opgedeeld kan worden in acht stadia en dat elk stadium eigen validiteitproblemen heeft. Crooks en Kane (1996) doen de suggestie per stadium in een assessment te kijken welke problemen zich voor kunnen doen en deze dan te bestrijden. Zij gaan niet in op de mogelijke bestrijdingstactieken.

Uit de toetsing van de problemen die Luken (2004) signaleert, blijkt dat het merendeel weggenomen wordt door de voorwaarden van Klarus (1998) om de validiteit te waarborgen en door de validiteitfacetten, opgesteld door Messick (1989). Echter, er blijkt ook dat één probleem alle andere problemen kan omvatten: *Competentiebeoordelingen zijn subjectief*. Wanneer uit het onderzoek blijkt dat er meerdere malen significante verschillen bestaan tussen de beoordelingen, dan kan terecht gesproken worden over instabiele competenties en beoordelingen. Competentie is in dit geval geen homogeen begrip en

competentiebeoordeling zal dan niet zinvol zijn in het onderwijs. Mocht blijken dat er zelden significante verschillen bestaan tussen de beoordelingen, dan mag wellicht juist gesproken worden over objectieve beoordelingen en dat de problemen waarvoor Luken (2004) waarschuwt, niet van toepassing zijn op de gebruikte competenties.

Frappant is dat geen van de problemen die Luken (2004) signaleert, de voorwaarden die Klarus (1998) stelt en de validiteitfacetten van Messick (1989) zich richten op de eigenschappen van de beoordelaars. Stokking en collega's (2004) doen in deze richting wel een suggestie.

2.3. Eigenschappen van assessoren

In een assessmentcultuur zijn de kwaliteiten van de assessor vele malen belangrijker dan in een test- en examineercultuur, waar het uiteindelijk behaalde resultaat louter door regels is bepaald. Eraut (1994) stelt dat een assessor in het huidige systeem aan een viertal eisen moet voldoen: (1) *hij moet het relevante systeem van assessment kennen en begrijpen*; dit ligt in het verlengde van één van de voorwaarden van Klarus (1998) die voorschrijft dat de procedures doorzichtig moeten zijn, zodat ze voor alle betrokkenen helder en duidelijk zijn. (2) *De assessor moet de standaarden van het assessment systeem op de juiste manier interpreteren.* Moss en Schutz (2001) hebben in twee cases, naar voorbeeld van Habermas (1990), gekeken hoe assessoren tot consensus komen in de beoordeling. Hierbij benoemen zij ook dat er consensus moet ontstaan over de standaarden zelf. Dit gebeurt nog voordat er sprake is van een beoordeling van materiaal van een lerende. Moss en Schutz (2001) concluderen dat het niet erg is als er geen consensus in de standaarden bereikt wordt. Zij menen dat het misschien zelfs wel beter is dan wanneer er wèl consensus bereikt wordt. Er ontstaat dan een compleet beeld van de situatie, waarin aandacht is voor de verschillende perspectieven op een bepaald aspect. Toch realiseren zij zich dat dit lastig is voor de mensen die de standaarden moeten gebruiken, aangezien er bepaald moet worden of iets voldoende dan wel onvoldoende is. Moss en Schutz (2001) komen dus tot de conclusie dat consensus over de inhoud van de standaarden een vereiste is. (3) *De assessor moet valide bewijsmateriaal verzamelen voor het assessment door middel van bijvoorbeeld observatie, mondelinge ondervraging en toetsing*. Hier lijkt een discrepantie te bestaan met een van de voorwaarden voor validiteit van Klarus (1998), namelijk dat het materiaal het vermogen moet hebben om beoordeeld te worden en dat er overeenstemming moet zijn of de lerende materiaal ter

beoordeling aandraagt of dat de beoordelaar kijkt naar welk materiaal geschikt is. Volgens Klarus (1998) moet er in elk geval op dit punt duidelijkheid zijn, hij spreekt geen voorkeur uit. (4) *De assessor moet de vastgestelde procedures en criteria toepassen bij het beoordelen van de lerende*. Deze eis komt eveneens overeen met een validiteitvoorwaarde van Klarus (1998), criterium gerelateerd, waarbij Eraut (1994) benadrukt dat te veel regels amateuristisch beoordelen uitlokken.

De tweede en vierde eis van Eraut (1994) richt zich op het interpreteren van de standaarden in beheersingsniveaus en de procedure van de beoordelingen. Beide hebben grote invloed op de beoordeling door de assessoren. Verschil in de standaarden kan bepalen of de procedure voor de assessoren helder en eenduidig is. Met de uitkomst tussen het vergelijken van verschillende standaarden, kan bepaald worden of de beoordelingsmethode adequaat is.

2.4. Standaarden in beheersingsniveaus

Een van de vereiste eigenschappen van een assessor houdt in dat hij het beoordelingssysteem moet begrijpen (Eraut, 1994). Om dit te kunnen is het in eerste instantie nodig dat hij een bepaalde basiskennis over de onderwijsinhoud bezit, dit kan beoordelen op niveau en vervolgens een algeheel beeld van de bewijzen kan geven. Naast kennis over de onderwijsinhoud is het rekening houden met eigen vooroordelen door de assessor misschien wel belangrijker. Door bewust te zijn van eigen waardeoordelen, voorkeuren en mogelijke vooroordelen voorkomt de assessor dat deze vervalt in het probleem dat Luken (2004) voorziet; het geven van een subjectieve beoordeling. De bewustwording van de eigen rol als assessor is een van de manieren om niet in subjectieve beoordelingen te vervallen, standaarden in beheersingsniveaus spelen eveneens een rol.

Robson (2002) benadrukt dat gebruik van verschillende schalen, standaarden in beheersingsniveaus, inzicht geeft in wat mensen ergens van vinden. Hierbij is het belangrijk dat er eveneens extreme positieve en negatieve items in de criteria opgenomen worden. Het lastige hiervan is dat het aantal geselecteerde items samen een consistent beeld moeten vormen. Een mogelijkheid om tegemoet te komen aan dit dilemma, is het consistent vasthouden aan eenduidigheid in de standaarden. Een mogelijk dilemma bij een assessor kan ontstaan wanneer er onduidelijkheid is of de intervallen tussen de antwoordmogelijkheden (mogelijk) niet gelijk zijn. Om te voorkomen dat niet eenduidige

beoordelingsmethoden worden gebruikt, adviseert Robson (2002) het materiaal door meerdere assessoren onafhankelijk van elkaar te gebruiken. Wanneer er geen significante verschillen tussen de antwoorden van de assessoren bestaan, kan er gesproken worden over eenduidige standaarden en een heldere beoordelingsmethode. Daarnaast kunnen assessoren gevraagd worden naar hun overwegingen bij het toekennen van een standaard. Een andere manier om de overwegingen bij het toekennen van standaarden te onderzoeken is het gebruik maken van een verschillend aantal standaarden bij dezelfde criteria. Mogelijk zijn de overwegingen genuanceerder als de assessor moet kiezen tussen 1, 2, 3, 4 of 5 (waarbij 1 voor onvoldoende staat en vervolgens gelijkmatig oplopend naar 5, als voldoende) dan wanneer hij kiest tussen uitsluitend onvoldoende en voldoende. Een laatste aandachtspunt binnen de beoordelingsmethode richt zich op de vastgestelde procedures, deze moeten bekend en toepasbaar zijn voor de assessor (Eraut, 1994; Klarus, 1998; Messick, 1989). In de procedures moet aandacht besteed worden aan de mogelijkheid dat assessoren tijdens het beoordelingsproces door nieuw bewijsmateriaal van oordeel veranderen. Er moet duidelijkheid zijn over het laatste moment dat de assessor zijn oordeel mag en kan wijzigen.

De helderheid en duidelijkheid die Messick (1989) en Klarus (1998) als voorwaarden stellen, kunnen gesteund worden door inzicht te krijgen in de afwegingen die een assessor tijdens de beoordeling maakt. Daarvoor is het nodig te weten wat een assessor denkt tijdens het beoordelen van het materiaal.

2.5. Hardop-denk-methode

Eer Ericsson en Simon (1984) met de methode voor verslaglegging van denkprocessen aan de slag kunnen, zijn er een vijftal kwesties die zij behandelen. Ten eerste achten zij het van groot belang een passend antwoord te vinden op de twijfels die bij veel psychologen bestaan over de geschiktheid van verbalisaties als fundamentele data in wetenschappelijk onderzoek. Hieruit voortvloeiend volgt een discussie over de methode ter omzetting van (verbaal) gedrag in data. Daarbij realiseren Ericsson en Simon (1984) zich dat deze data ‘hard’, oftewel objectief en eenduidig moeten zijn. Het coderingsproces moet daartoe theoretisch sterk onderbouwd zijn. Een laatste kwestie die aangesneden wordt, is het feit dat de processen gespecificeerd moeten worden waarmee terug gegaan kan worden van de data naar respectievelijk gedrag en gedachten.

Het zwaartepunt in de hierboven genoemde aandachtspunten ligt bij het omzetten van gedachten en gedrag naar harde data. Ericsson en Simon (1984) hebben verschillende niveaus van verbalisaties verwoord om grip te krijgen op het objectief maken van de gegevens.

1. Verbaliseren van bedekte informatie; hierbij verwoordt de deelnemer wat deze doet, denkt of leest.
2. Verbaliseren en expliciteren of labelen van de inhoud; hierbij verwoordt de deelnemer niet alleen wat hij doet, denkt of leest, maar verbindt hij het met elkaar en geeft mogelijk zijn eigen mening.
3. Verbalisatie koppelen aan eerdere gedachten; de deelnemer interpreteert wat hij doet, denkt of leest en verbindt het met reeds verworven kennis.

Ericsson en Simon (1984) kwamen tot de formulering van de Hardop-denk-methode, waarbij de deelnemer aangemoedigd wordt te praten zonder de intentie te hebben te communiceren. Hij moet zijn gedachten proberen te verwoorden. Deze werkwijze wordt vaker gebruikt dan voorzien; docenten vragen studenten regelmatig hun gedachten en stappenplannen te verwoorden tijdens het oplossen van een vraagstuk. Door deze methode komt de oplossingsstrategie bloot te liggen en kan de docent zien waar de student een fout maakt. De docent kan hierdoor instructie geven gericht op de kern van het probleem.

Om deelnemers aan een onderzoek te stimuleren hardop te denken, adviseren Ericsson en Simon (1984) om voordat het onderzoek begint te oefenen met het hardop denken. Daarnaast is het noodzakelijk om tijdens het onderzoek de deelnemer te stimuleren alles te zeggen wat hij denkt. Deze aanbevelingen zijn (voor zover mogelijk was) uiteraard ter harte genomen in het onderstaande onderzoek, waarbij zowel de subjectiviteit van beoordelingen als de effectiviteit van de Hardop-denk-methode aan bod is gekomen.

2.6. Onderzoeksvragen

Subjectiviteit van competentiebeoordeling en de dreiging terug te vallen in beoordelen naar vaste maatstaven, zijn twee problemen die Luken (2004) voorziet bij de huidige ontwikkelingen in het onderwijs. Beide problemen bleken bij de toetsing met de voorwaarden voor validiteit (Klarus, 1998) en de validiteitfacetten (Messick, 1989) in het theoretisch kader niet geheel stand te houden. Het vervallen in vaste maatstaven om subjectiviteit in de competentiebeoordelingen te voorkomen, betekent al snel een nekslag voor de huidige ontwikkelingen binnen het onderwijs. Daarnaast waarschuwt Eraut (1994) voor amateuristisch optreden door assessoren bij een veelvoud aan beoordelingscriteria. Om zicht te krijgen op

de overwegingen die assessoren tijdens het beoordelen maken, hebben Ericsson en Simon (1984) de Hardop-denk-methode geformuleerd. Hierbij worden de achterliggende gedachten van een assessor bij een beoordeling kenbaar gemaakt, waardoor zichtbaar wordt hoe een beoordelaar tot zijn of haar oordeel komt.

Om toch een indicatie te krijgen van de subjectiviteit van beoordelingen door assessoren aan Hogeschool Domstad en om de invloeden op de mogelijke subjectiviteit te beoordelen, wordt een onderscheid gemaakt in de achtergrond van de assessoren, de momenten dat oordelen worden gevormd en de standaarden in beheersingsniveaus die in de beoordelingsmethoden gebruikt worden.

Dit kader leidt tot de volgende onderzoeksvragen:

1. Zijn de assessmentbeoordelingen subjectief?
2. Welke aspecten hebben invloed op de subjectiviteit van de assessmentbeoordelingen?

Om deze onderzoeksvraag zo volledig mogelijk te beantwoorden, moeten eerst antwoorden gevonden worden op de volgende deelvragen:

A. Speelt de achtergrond van de interne versus de externe assessoren een rol bij het beoordelen van het assessment?
B. Welke invloed hebben het dossier en het gesprek afzonderlijk van elkaar op de beoordeling? Oftewel ‘Op welk moment vormt de assessor zijn mening?’
C. Welke invloed hebben de twee- en vijf puntsschaal in beheersingsniveaus op de beoordeling die de assessor geeft?
D. Wat beweegt een assessor een bepaalde beoordeling voor het gehele assessment te geven?
E. Als er tussen twee assessoren verschillende oordelen bestaan, hoe brengen zij dit dichter bij elkaar?

2.7. Hypothesen

Binnen dit onderzoek omvat de nulhypothese dat alle assessoren na het bekijken van het assessmentdossier en na het voeren van het assessmentgesprek geheel objectief in hun beoordeling zijn, waarbij er geen verschil bestaat tussen de beoordelingen van de interne en externe assessoren, de assessoren gelijke waarden aan de standaarden hechten en uit de Hardop-denk-methode blijkt dat de assessoren volgens dezelfde overwegingen tot een oordeel komen. Daar tegenover staat de hypothese; de assessoren zijn tot op zekere hoogte subjectief in alle stadia van het beoordelingsproces.

De nulhypothese en de hypothese kunnen voor een drietal deelvragen gespecificeerd worden:

A. H_0 = De achtergrond van de interne versus de externe assessoren speelt geen rol bij de beoordeling.

   H_1 = De achtergrond van de interne versus de externe assessoren speelt wel een rol bij de beoordeling.

BI. H_0 = Het dossier en het gesprek hebben evenveel invloed op de beoordeling.

   H_1 = Het dossier en het gesprek hebben niet evenveel invloed op de beoordeling.

BII. H_0 = Tussen de voorlopige beoordeling na het voeren van het gesprek en de definitieve beoordeling veranderen interne en externe assessoren niet meer van mening.

   H_1 = Tussen de voorlopige beoordeling na het voeren van het gesprek en de definitieve beoordeling veranderen interne en externe assessoren wel van mening.

C. H_0 = Het verschil in standaarden in de beheersingsniveaus heeft geen invloed op de beoordeling.

   H_1 = Het verschil in standaarden in de beheersingsniveaus heeft wel invloed op de beoordeling.

Hierbij geldt dat wanneer uit de data blijkt dat een deelhypothese aangenomen moet worden, de algemene nulhypothese verworpen moet worden, echter de andere deelnulhypothesen kunnen dan alsnog aangenomen worden.

Luken (2004), Eraut (1994) en Cluitmans en Klarus (2005) menen dat het mogelijk is dat door de verschillende waarden van de assessoren de validiteit van de beoordeling in gevaar komt. De bestaande literatuur gaat dus uit van de bovenstaande hypothesen. Hierbij wordt verwacht dat de interne assessoren lagere beoordelingen geven dan de externe assessoren, aangezien er vanuit gegaan mag worden dat zij het curriculum beter kennen en daar dus kritischer naar kunnen kijken. Verwacht wordt dat het assessmentgesprek meer invloed op de beoordeling zal hebben dan het assessmentdossier, aangezien de student in het gesprek twijfels van assessoren kan wegnemen en vragen kan beantwoorden. Daarnaast wordt verwacht dat de vijf puntsschaal in de standaarden van beheersingsniveaus meer invloed zal hebben op de beoordeling dan de twee puntsschaal, omdat de vijf puntsschaal meer mogelijkheden geeft in differentiatie van de beoordeling. Overigens zijn deze verwachtingen zonder voorspelling in welke mate en in twee gevallen de specifieke richting er sprake zal zijn van de subjectiviteit, aangezien er geen vooronderzoek in dit specifieke gebied bestaat.

3. Methode

3.1. Onderzoeksgroep

Aan het onderzoek hebben in het totaal 28 assessoren deel genomen, waarvan 19 interne assessoren en 9 externe assessoren; $N_{tot.} = 28$, $N_{int.} = 19$ en $N_{ext.} = 9$. De interne assessoren zijn docenten van Hogeschool Domstad en de externe assessoren zijn medewerkers met bevoegdheid tot lesgeven op basisscholen, het merendeel basisscholen waar studenten van de Hogeschool stage lopen. Bijna alle assessoren zijn 44 jaar of ouder. Het aantal jaren werkervaring is heel verschillend; van 3 jaar tot 37 jaar, met een gemiddelde van 22.8 jaar. Alle assessoren hebben een opleiding gevolgd die gericht is op het onderwijs; onder andere de Pedagogische academie basisonderwijs, de universitaire opleiding Onderwijskunde, verschillende universitaire vakspecifieke opleidingen om vervolgens les te geven op een middelbare school of aan de Pedagogische academie. Momenteel zijn de meeste interne assessoren opleidingsdocent aan Hogeschool Domstad, een enkeling vervult een managersfunctie of is daarnaast ook studiebegeleider. De meeste externe assessoren zijn groepsleerkracht op een basisschool, enkelen vervullen daarnaast een ondersteunende onderwijsinhoudelijke functie op de basisschool.

Aangezien het onderzoek uit drie delen bestaat en niet iedereen in staat was aan alle delen deel te nemen, is het aantal assessoren dat per onderdeel mee heeft gewerkt verschillend. Bij het eerste deel van het vooronderzoek hebben alle assessoren meegewerkt; $N1_{tot.} = 28$, $N1_{int.} = 19$ en $N1_{ext.} = 9$. Aan het tweede deel van het vooronderzoek hebben 13 interne assessoren en 5 externe assessoren meegewerkt; $N2_{tot.} = 18$, $N2_{int.} = 13$ en $N2_{ext.} = 5$. Bij de interventie is zowel onderscheid gemaakt tussen interne en externe assessoren, als tussen de controle groep en experimentele groep; er hebben in het totaal 19 assessoren meegewerkt; ($N3_{tot.} = 19$). Daarvan zaten 12 assessoren in de controle groep, waarvan 8 interne assessoren en 4 externe assessoren ($N3_{contr.} = 12$, $N3_{contr.; int.} = 8$ en $N3_{contr.; ext.} = 4$). De experimentele groep bestond uit 7 assessoren, waarvan 3 interne assessoren en 4 externe assessoren ($N3_{exp.} = 7$, $N3_{exp.; int.} = 3$ en $N3_{exp.; ext.} = 4$). Van de controle groep hebben 10 assessoren in 5 assessorenparen meegewerkt aan de Hardop-denk-methode ($N3_{HDMcontr.} = 10$) en van de experimentele groep hebben 6 assessoren in 3 assessorenparen meegewerkt aan deze methode ($N3_{HDMexp.} = 6$). Beide experimentele groepen zijn kleiner dan de controle groepen, aangezien er enkele assessments uitvielen door terugtrekking door de student.

3.2. Situatie

In het vernieuwde onderwijssysteem van de lerarenopleiding basisonderwijs Hogeschool Domstad te Utrecht, sluiten de derde- en vierdejaars studenten elk half jaar af door deel te nemen aan een assessment. Met behulp van dit assessment –in de vorm van een assessmentdossier en een assessmentgesprek– behalen zij de benodigde studiepunten voor dat halve jaar. In dit onderwijssysteem is er geen andere mogelijkheid om studiepunten te behalen. Voorafgaand aan het assessmentgesprek stelt de student aan de hand van richtlijnen (algemene richtlijnen en specifieke competenties uit de competentiematrix) een assessmentdossier samen met materiaal waar de student ongeveer een half jaar aan heeft gewerkt. De studiebegeleider controleert het assessmentdossier op volledigheid en schat in of er een reële kans bestaat dat de student tenminste 20 studiepunten, van de maximale 30 studiepunten, zal behalen. Hierna levert de student het assessmentdossier in bij de assessoren. Elk assessment wordt afgenomen door een eerste assessor (in de meeste gevallen een docent van de Hogeschool) en een tweede assessor, vaak een externe assessor. De ideaalverdeling (het assessment wordt afgenomen door een interne en externe assessor) is gemaakt om tijdens het assessment de aandacht voor de theoretische en praktische kant van het materiaal en het vak te waarborgen. Beide assessoren bekijken afzonderlijk van elkaar, voorafgaand aan het assessmentgesprek, het assessmentdossier van de student; de eerste assessor doet dit uitgebreider dan de tweede assessor en zal ook tijdens het gesprek de leiding nemen. Na het gesprek bepalen de assessoren hoeveel studiepunten de student behaald heeft. De inspraak in de beoordeling per assessor is niet nader bepaald, dit gaat in overleg. De assessoren kunnen de student respectievelijk 0, 20 of 30 studiepunten geven. Bij 0 punten heeft de student het assessment en daarmee het halve studiejaar niet gehaald. Bij 20 punten is een aanvulling nodig, deze wordt door de assessoren geformuleerd. De student zal zelf –mogelijk met hulp van de studiebegeleider– de aanvulling maken en deze weer bij de assessoren inleveren, waarna zij bepalen of de student alsnog de resterende 10 studiepunten toegekend krijgt. 30 studiepunten is het maximale aantal studiepunten per assessment, de student hoeft dan niets meer te doen.

3.3. Design en procedure

Het onderzoek betreft een designonderzoek naar subjectiviteit van, en mogelijke invloeden op de beoordelingen van assessments van studenten bij Hogeschool Domstad te Utrecht door assessoren. De uitkomsten van het onderzoek kunnen leiden tot een aanbeveling tot een structurele aanpassing van het beoordelingssysteem van de assessments bij Hogeschool Domstad.

Het onderzoek bestaat uit drie delen; het vooronderzoek bestaat uit twee delen die gezamenlijk een zo compleet mogelijk beeld geven van de bestaande situatie. Het derde deel, de interventie, is ontwikkeld met behulp van de gegevens uit het vooronderzoek.

Hogeschool Domstad gaat uit van de mogelijkheid om 0, 20 of 30 studiepunten toe te kennen. Aangezien deze verdeling geen gelijkmatig beeld zou geven van de resultaten, is uitgegaan van een verdeling van 0, 10 of 20 punten, waarbij 20 studiepunten met 10 punten correleert en 30 studiepunten met 20 punten.

3.3.1. Vooronderzoek deel 1

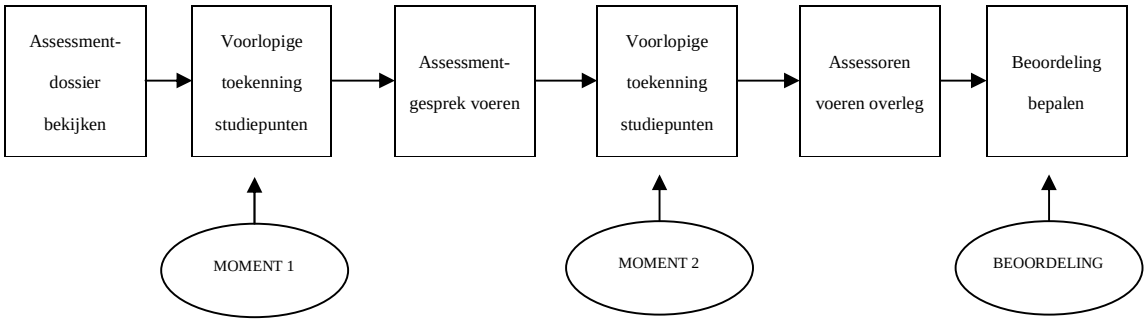
Voor dit deel worden gegevens van het afgelopen studiejaar gebruikt. De data zijn toegekende studiepunten door interne assessoren ($N1_{int.}$ = 19) en externe assessoren ($N1_{ext.}$ = 9) die momenteel nog assessments afnemen, $N1_{tot.}$ = 28. Om te bepalen in hoeverre er sprake is van subjectiviteit bij de beoordelingen van de assessments door de assessoren, wordt er gekeken naar de beoordelingen tussen beide groepen assessoren. Hierbij zal uitgegaan worden van het inhoudelijk deel van de beoordeling, aangezien het overige deel van het beoordelingsformulier van de Hogeschool uit het al dan niet aanwezig zijn van procedurele onderdelen in het assessmentdossier bestaat. De subjectiviteit zal gemeten worden met de *Mann-Whitney U-test*. Voor deze toets is gekozen, omdat de data onafhankelijk van elkaar zijn, de groepen a select gevormd zijn en het om een kleine steekproef gaat. Daarnaast zal er waarschijnlijk geen sprake van een normaalverdeling zijn. De subjectiviteit zal per toets worden vastgesteld wanneer er een significant verschil bestaat tussen de uitkomsten ($p < .05$). Hier zal sprake zijn van toetsend onderzoek.

3.3.2. Vooronderzoek deel 2

In het tweede deel van het onderzoek wordt de assessoren gevraagd op twee momenten te noteren hoeveel studiepunten zij de student op dat moment zouden toekennen. Deze momenten zijn na het bekijken van het assessmentdossier en na het voeren van het assessmentgesprek (Zie Figuur 1). Met de data die hieruit voortkomen, zal bepaald worden wanneer de subjectiviteit de grootste rol speelt bij de interne en externe assessoren afzonderlijk (met andere woorden: wanneer vormt de assessor zijn mening?). Aan dit deel van het vooronderzoek werken 13 interne assessoren ($N2_{int.} = 13$) en 5 externe assessoren ($N2_{ext.} = 5$) mee, waarbij $N2_{tot.} = 18$ geldt.

Zowel de beide voorlopige toekenningen van moment 1 en van moment 2 van de assessoren worden met elkaar vergeleken, als de voorlopige toekenning van moment 2 en de uiteindelijke beoordeling. Deze twee toetsen worden binnen de groep interne assessoren en binnen de groep externe assessoren uitgevoerd. De toets die voor deze vier vergelijkingen gebruikt wordt, is de *Wilcoxon signed ranks test*. Voor deze toets is gekozen, omdat de data afhankelijk van elkaar zijn, de groepen a select gevormd zijn en het om een kleine steekproef gaat, en er zal waarschijnlijk geen sprake zijn van een normaalverdeling. De subjectiviteit zal per toets worden vastgesteld wanneer er een significant verschil bestaat tussen de uitkomsten ($p < .05$). Hier zal sprake zijn van toetsend onderzoek.

De voorlopig toegekende beoordelingen door de interne en externe assessoren worden eveneens met elkaar vergeleken. Dit gebeurt zowel voor de voorlopig toegekende beoordelingen van moment 1 als van moment 2. Voor beide vergelijkingen wordt de *Mann-Whitney U test* gebruikt, aangezien de data onafhankelijk van elkaar zijn, de groepen a select gevormd zijn en er sprake is van een kleine steekproef. Verwacht wordt dat er geen sprake zal zijn van een normaalverdeling. De subjectiviteit zal per toets eveneens worden vastgesteld wanneer er een significant verschil bestaat tussen de uitkomsten ($p < .05$). Wederom zal er sprake zijn van toetsend onderzoek.



Figuur 1. *Momenten van dataverzameling*

3.3.3. Interventie

De data uit beide delen van het vooronderzoek vormen de basis voor de interventie die is opgesteld. Deze interventie bestaat, afgezien van een kleine aanvulling, uit het inhoudelijk onderdeel van het huidige beoordelingsformulier van Hogeschool Domstad. Bij dit onderzoek is gekozen alleen het inhoudelijk deel van het beoordelingsformulier van de Hogeschool in het onderzoek mee te nemen, aangezien hierin bij de beoordeling van het assessmentdossier en assessmentgesprek de grootste interpretatievrijheid van de assessor over de standaarden in beheersingsniveaus bestaat. Het overige deel van het beoordelingsformulier van de Hogeschool bestaat uit het al dan niet aanwezig zijn van procedurele onderdelen in het assessmentdossier. De aanvulling bevat een gespecificeerd onderdeel, waardoor de assessments van alle studenten in aanmerking komen voor het onderzoek.

De assessorenparen uit de controle groep ($N3_{contr.\ paren} = 12$) maken bij de beoordeling gebruik van gebruikelijke twee puntsschaal; per onderdeel kunnen zij een onvoldoende dan wel een voldoende toekennen. De assessorenparen uit de experimentele groep ($N3_{exp.\ paren} = 7$) passen een vijf puntsschaal toe; per onderdeel kunnen zij 1, 2, 3, 4 of 5 punten toekennen (Zie Appendix 1, Tabel 1 en 2). Beide puntsschalen zijn van toepassing op alle onderdelen in het inhoudelijk deel van het beoordelingsformulier dat momenteel gebruikt wordt op Hogeschool Domstad. Er is sprake van gelijke afstanden tussen de punten in het beoordelingsformulier van de experimentele groep. Het assessment wordt door deze opzet op de huidige inhoudelijke punten beoordeeld, alleen het aantal standaarden in beheersingsniveaus verschilt. Voor deze tweedeling is gekozen om het verschil tussen de huidige situatie, de controle groep, en de experimentele groep zo eenduidig mogelijk te laten zijn, waardoor de uitkomsten geheel te wijten zijn aan

het verschil in de twee- en vijf puntsschaal. De formulieren worden door de assessoren individueel onafhankelijk van elkaar zowel na het bekijken van het assessmentdossier als na het voeren van het assessmentgesprek gescoord (Zie Figuur 1).

Met behulp van de vier beoordelingen (moment 1 en moment 2 van beide assessoren) zullen de assessoren een eindoordeel vormen dat wordt omgezet in respectievelijk 0, 20 of 30 studiepunten. Aan de hand van de beoordelingscriteria zullen de assessoren zoals gebruikelijk bepalen op welk vlak een aanvulling nodig is, indien de student 20 studiepunten behaald heeft.

Ook bij de interventie worden zowel de beide voorlopige toekenningen van moment 1 en van moment 2 door de assessoren uit de controle groep met elkaar vergeleken, als de voorlopige toekenning van moment 2 en de uiteindelijke beoordeling. Deze toetsen worden ook uitgevoerd met de data van de assessoren uit de experimentele groep. De toets die voor deze vier vergelijkingen gebruikt wordt, is de *Wilcoxon signed ranks test*. Voor deze toets is gekozen, omdat de data afhankelijk van elkaar zijn, de groepen a select gevormd zijn en het om een kleine steekproef gaat, en er zal waarschijnlijk geen sprake zijn van een normaalverdeling. De subjectiviteit zal per toets worden vastgesteld wanneer er een significant verschil bestaat tussen de uitkomsten ($p < .05$). Ook hier zal sprake zijn van toetsend onderzoek.

Daarnaast worden zowel de voorlopige toekenningen van moment 1 tussen de controle en experimentele groep met elkaar vergeleken, als de voorlopige toekenningen van moment 2 tussen de controle en experimentele groep. Voor beide vergelijkingen wordt de *Mann-Whitney U test* gebruikt, aangezien de data onafhankelijk van elkaar zijn, de groepen a select gevormd zijn en er sprake is van een kleine steekproef. Verwacht wordt dat er geen sprake zal zijn van een normaalverdeling. De subjectiviteit zal per toets eveneens worden vastgesteld wanneer er een significant verschil bestaat tussen de uitkomsten ($p < .05$). Wederom zal er sprake zijn van toetsend onderzoek.

Uit de literatuur blijkt en in het vooronderzoek wordt verondersteld dat assessoren gebruik maken van een zekere interpretatievrijheid bij de beoordeling van de competenties. Deze komen voort uit vooroordelen van assessoren, de basiskennis van assessoren en het beeld dat de individuele assessor heeft van het ideale assessment.

Om een duidelijker beeld te krijgen van de interpretaties van de standaarden en de kwantitatieve data uit het vooronderzoek en om de interventie te kunnen onderbouwen, zullen assessorenparen van zowel

de controle groep ($N3_{HDMcontr.}$ = 10) als van de experimentele groep ($N3_{HDMexp.}$ = 6) na het voeren van het assessmentgesprek de beoordelingsformulieren individueel invullen met behulp van de Hardop-denk-methode. De onderzoeker zal een geluidsopname maken van de onderbouwing van de keuzes binnen de beoordeling, zodat achteraf een beeld kan worden geschetst van de individuele waarden, de interpretatie, die aan de gebruikte standaarden gehecht worden. Ook het gesprek dat de assessoren samen voeren om tot een eindbeoordeling te komen, zal worden opgenomen. De transcriptie van deze geluidsopnames zal volgens open, axiale en selectieve codering worden geanalyseerd en zal op die wijze een complementerende en additionele functie hebben bij de uitkomsten van het kwantitatieve deel van het onderzoek.

3.4. Instrumenten

In het eerste deel van het vooronderzoek wordt de significantie van de beoordelingen tussen de groep interne assessoren en de groep externe assessoren bepaald. De significantie bepaalt of er sprake is van subjectiviteit bij het beoordelen van de assessments.

In het tweede deel van het vooronderzoek wordt de significantie bepaald tussen de toekenningen van studiepunten na het lezen van het assessmentdossier (moment 1) en na het voeren van het assessmentgesprek (moment 2) en tussen de toekenningen van na het assessmentgesprek en de beoordeling (Zie Figuur 1). Dit gebeurt zowel van de groep interne assessoren als van de groep externe assessoren. Daarnaast worden de voorlopige toekenningen van studiepunten van zowel moment 1 als van moment 2 tussen de groepen met interne en externe assessoren vergeleken. De significanties bepalen of er sprake is van subjectiviteit bij het beoordelen van de assessments en het moment dat de assessor zijn mening vormt.

Bij de interventie wordt de significantie bepaald tussen de voorlopige toekenningen van studiepunten op moment 1 en op moment 2 en tussen de voorlopige toekenningen van moment 2 en de beoordeling. Dit gebeurt zowel van de controle groep als van de experimentele groep. Het verschil tussen de controle en de experimentele groep richt zich uitsluitend op de puntsschalen in het inhoudelijk deel van het huidige beoordelingsformulier. Daarnaast worden de voorlopige toekenningen van studiepunten van zowel moment 1 als van moment 2 tussen de controle en de experimentele groep met elkaar vergeleken. De significanties bepalen of er sprake is van subjectiviteit bij het beoordelen van de assessments, het moment

dat de assessor zijn mening vormt en de mate van invloed dat het beoordelingsformulier (Zie Appendix 1, Tabel 1 en 2) op de beoordeling heeft. Vervolgens zal een vijftal assessorenparen uit de controle groep en een drietal assessorenparen uit de experimentele groep deelnemen aan de Hardop-denk-methode. Hierbij verwoorden zij hun overwegingen die zij maken tijdens het invullen van het beoordelingsformulier, nadat zij het assessmentgesprek met de student hebben gevoerd, waardoor de onderbouwing van de keuze voor een bepaalde beoordeling duidelijk wordt.

3.5. Analyse

De data uit het vooronderzoek zullen allen kwantitatief van aard zijn. Van de studiepunten uit de beoordelingen, gegeven door de interne en externe assessoren in het eerste deel van het vooronderzoek, wordt de significatie bepaald tussen de interne en externe groep assessoren. Dit gebeurt met de *Mann-Whitney U test*, aangezien de data onafhankelijk van elkaar zijn, de groepen a select gevormd zijn en het om een kleine steekproef gaat. Daarnaast zal er waarschijnlijk geen sprake van een normaalverdeling zijn.

Van de data uit het tweede deel van het vooronderzoek, wordt de significantie bepaald tussen de momenten van het toekennen van de voorlopige studiepunten en tussen de interne en externe groep assessoren. De toets die voor de eerste vergelijkingen –uitsluitend gericht op het moment– gebruikt wordt, is de *Wilcoxon signed ranks test*. Voor deze toets is gekozen, omdat de data afhankelijk van elkaar zijn, de groepen a select gevormd zijn en het om een kleine steekproef gaat. Daarnaast zal er waarschijnlijk geen sprake van een normaalverdeling zijn. De toets die voor de laatste twee vergelijkingen –tussen de interne en externe assessoren– gebruikt wordt, is de *Mann-Whitney U test*. Voor deze toets is gekozen omdat de data onafhankelijk van elkaar zijn, de groepen a select gevormd zijn en het om een kleine steekproef gaat. Daarnaast is de verwachting dat er wederom geen sprake zal zijn van een normaalverdeling.

Van de data van de interventie, gegeven door de assessoren uit de controle en de experimentele groep, wordt de significantie bepaald tussen de momenten van het toekennen van de voorlopige studiepunten en tussen de assessoren uit de controle en experimentele groep. Bij deze vergelijkingen wordt ook een onderscheid gemaakt tussen de interne en externe assessoren, zowel binnen de controle groep als de experimentele groep. De toets die voor de eerste vergelijkingen –uitsluitend gericht op het moment van

voorlopige toekenning– gebruikt wordt, is eveneens de *Wilcoxon signed ranks test*. Voor deze toets is wederom gekozen, omdat de data afhankelijk van elkaar zijn, de groepen a select gevormd zijn en het om een kleine steekproef gaat. Daarnaast zal er waarschijnlijk geen sprake van een normaalverdeling zijn. De toets die voor de laatste twee vergelijkingen –tussen alle assessoren uit de controle groep en alle assessoren uit de experimentele groep– gebruikt wordt, is de *Mann-Whitney U test*. Voor deze toets is gekozen omdat de data onafhankelijk van elkaar zijn, de groepen a select gevormd zijn en het om een kleine steekproef gaat. Daarnaast is de verwachting dat er wederom geen sprake zal zijn van een normaalverdeling.

De transcriptie van de geluidsopnamen, gemaakt ten behoeve van de Hardop-denk-methode, zullen volgens open, axiale en selectieve codering geanalyseerd worden. De coderingen zullen vervolgens een onderbouwing zijn van de interpretaties die bij de verschillende assessoren bestaan achter de standaarden van de beoordelingen.

Vanzelfsprekend worden de gegevens vertrouwelijk en anoniem behandeld en is achteraf niet te bepalen welke interne of externe assessor welke beoordelingen heeft gegeven. Met behulp van het softwareprogramma SPSS wordt gekeken of er bij de gemiddelde beoordelingen van de interne of externe assessoren of de assessoren uit de controle of experimentele groep sprake is van een significant verschil. Aan de hand hiervan wordt vastgesteld of er sprake is van subjectiviteit bij de beoordelingsprocessen.

4. Resultaten

4.1. Vooronderzoek deel 1

Om te kijken of er een significant verschil bestaat tussen de beoordelingen van de assessments door interne of door externe assessoren in het eerste deel van het vooronderzoek, is er gebruik gemaakt van de *Mann-Whitney U test*. Uit de vergelijking van de data blijkt geen verschil tussen de interne en de externe assessoren te bestaan ($U(19, 9) = 5.3$, $p = .43$). Hoewel de hypothese een lagere beoordeling door de interne assessoren voorspelde, geven de interne en externe assessoren geen verschillende beoordeling ($M1_{int.} = 16.0$; $M1_{ext.} = 17.3$; $SD_{int.} = 5.8$; $SD_{ext.} = 4.9$).

4.2. Vooronderzoek deel 2

Voor de vergelijking tussen de voorlopige toekenningen van studiepunten van moment 1 en van moment 2 is gekozen voor de *Wilcoxon signed ranks test*, aangezien de data afhankelijk van elkaar zijn. Bij zowel de interne assessoren als bij de externe assessoren is het verschil tussen het voorlopig toegekende aantal studiepunten van moment 1 en van moment 2 significant. Bij de interne assessoren geldt ($U(34, 47) = 14.6$, $p = .01$). In de lijn der verwachting geven de interne assessoren gemiddeld na het voeren van het assessmentgesprek een hogere voorlopige beoordeling dan wanneer zij uitsluitend het assessmentdossier bekeken hebben ($M2_{int.;\ mom.1} = 15.6$; $M2_{int.;\ mom.\ 2} = 18.7$; $SD_{int.;\ mom.1} = 6.1$; $SD_{int.;\ mom.2} = 3.4$). Bij de externe assessoren geldt ($U(11, 11) = 13.2$, $p = .01$). Ook de externe assessoren geven dus, zoals verwacht werd, gemiddeld na het voeren van het assessmentgesprek een hogere voorlopige beoordeling dan zij aanvankelijk na het lezen van het assessmentdossier in gedachten hadden ($M2_{ext.;\ mom.\ 1} = 12.7$; $M2_{ex.;\ mom.\ 2} = 18.2$; $SD_{ext.;\ mom.\ 1} = 4.7$; $SD_{ext.;\ mom.\ 2} = 4.0$).

Ook de voorlopige toekenning van studiepunten van moment 2 en de uiteindelijke definitieve beoordeling zijn met elkaar vergeleken. Om te kijken of er significante verschillen bestaan bij de interne of externe assessoren, is gebruik gemaakt van de *Wilcoxon signed ranks test*, aangezien ook deze data afhankelijk van elkaar zijn. Uit de vergelijking van de interne assessoren blijkt een significant verschil te bestaan tussen de voorlopig toegekende studiepunten van moment 2 en de uiteindelijke beoordeling ($U(47, 47) = 8.1$, $p = .03$). De interne assessoren geven, zoals verwacht, gemiddeld bij de uiteindelijke beoordeling een lagere beoordeling dan de voorlopige toekenning van studiepunten na het voeren van het assessmentgesprek ($M2_{int.;\ mom.\ 2} = 18.7$; $M2_{int.;\ beoord.} = 17.4$; $SD_{int.;\ mom.\ 2} = 3.4$; $SD_{int.;\ beoord.} = 4.4$).

Dezelfde vergelijking, tussen de voorlopige toekenning van studiepunten van moment 2 en de uiteindelijke definitieve beoordeling, is ook bij de externe assessoren gemaakt, waarbij geen verschil is waargenomen ($U(11, 11) = 2.1$, $p = 1.00$). Hoewel de hypothese uitging van verschillende beoordelingen veranderen de externe assessoren gemiddeld na het assessmentgesprek niet van mening ($M2_{ext.;\ mom.\ 2} = 18.7$; $M2_{ext.;\ beoord.} = 17.4$; $SD_{ext.;\ mom.\ 2} = 4.0$; $SD_{ext.;\ beoord.} = 4.0$).

Tot nu toe zijn in het tweede deel van het vooronderzoek slechts vergelijkingen gemaakt tussen de verschillende momenten en binnen dezelfde groep deelnemers; enerzijds de interne assessoren, anderzijds de externe assessoren.

Voor de vergelijking van de voorlopige toekenning van studiepunten van moment 1 tussen de interne en externe assessoren, is gebruik gemaakt van de *Mann-Whitney U test*, aangezien hier geen sprake is van afhankelijkheid. Uit de data blijkt geen verschil te bestaan tussen de voorlopig toegekende studiepunten van moment 1 door de interne versus externe assessoren ($U(34, 11) = 7.3$, $p = .10$). De interne assessoren lijken, tegen de voorspelling van een lagere beoordeling in, gemiddeld juist een hogere beoordeling te geven dan de externe assessoren, maar er is geen sprake van een verschil ($M2_{int.;\ mom.\ 1} = 15.6$; $M2_{ext.;\ mom.\ 1} = 12.7$; $SD_{int.;\ mom.\ 1} = 6.1$; $SD_{ext.;\ mom.\ 1} = 3.6$).

Dezelfde toets is uitgevoerd op voorlopig toegekende studiepunten van moment 2 door de interne en externe assessoren. Uit de data blijkt eveneens geen verschil te bestaan tussen de voorlopig toegekende beoordelingen na het voeren van het assessmentgesprek door de interne versus externe assessoren ($U(47, 11) = 4.3$, $p = .56$). Hoewel de hypothese uitging van een lagere toekenning van studiepunten door de interne assessoren, geven zij gemiddeld geen hogere of lagere score na het voeren van het assessmentgesprek dan de externe assessoren ($M2_{int.;\ mom.\ 2} = 18.7$; $M2_{ext.;\ mom.\ 2} = 18.2$; $SD_{int.;\ mom.\ 2} = 3.6$; $SD_{ext.;\ mom.\ 2} = 6.1$).

4.3. Interventie

Er is eerst gekeken naar het verschil tussen de voorlopige toekenningen van studiepunten van moment 1 en van moment 2 door de interne en externe assessoren uit de controle groep (gebruikmakend van de twee puntsschaal in beheersingsniveaus) afzonderlijk. Voor beide vergelijkingen is gekozen voor de *Wilcoxon signed ranks test*, aangezien de data afhankelijk van elkaar zijn. Uit de vergelijking van de voorlopige toekenningen van studiepunten van moment 1 en van moment 2 door de interne assessoren uit de controle groep, bleek geen verschil te bestaan ($U(10, 15) = 4.1$, $p = .08$). Deze uitkomst neigt wel naar significantie. Tegen de voorspelling van de hypothese in, geven de interne assessoren uit de controle groep gemiddeld na het voeren van het assessmentgesprek geen hogere of lagere beoordeling dan wanneer zij uitsluitend het assessmentdossier bekeken hebben ($M3_{contr.;\ int.;\ mom.\ 1} = 15.0$; $M3_{contr.;\ int.;\ mom.\ 2} = 18.0$; $SD_{contr.;\ int.;\ mom.\ 1} = 5.3$;

$SD_{contr.;\ int.;\ mom.\ 2}$ = 4.1). Uit de vergelijking tussen de voorlopige toekenningen van studiepunten van moment 1 en van moment 2 door de externe assessoren, bleek geen verschil te bestaan ($U(11, 11) = 6.6$, $p = .32$). Hoewel de hypothese een verschillende beoordeling voorspelde, geven de externe assessoren gemiddeld na het voeren van het assessmentgesprek geen hogere of lagere beoordeling dan wanneer zij uitsluitend het assessmentdossier bekeken hebben ($M3_{contr.;\ ext.;\ mom.\ 1}$ = 16.4; $M3_{contr.;\ ext.;\ mom.\ 2}$ = 17.3; $SD_{contr.;\ ext.;\ mom.\ 1}$ = 6.7; $SD_{contr.;\ ext.;\ mom.\ 2}$ = 4.7).

Voor dezelfde vergelijkingen bij de experimentele groep (gebruikmakend van de vijf puntsschaal in beheersingsniveaus) is eveneens gekozen voor de *Wilcoxon signed ranks test*, aangezien ook deze data afhankelijk van elkaar zijn. Bij de experimentele groep zijn de vergelijkingen voor de interne en externe assessoren eveneens afzonderlijk van elkaar uitgevoerd. Uit de vergelijking tussen de voorlopige toekenningen van studiepunten van moment 1 en van moment 2 door de interne assessoren, bleek geen verschil te bestaan ($U(5, 5) = 6.6$, $p = .32$). De interne assessoren uit de experimentele groep geven gemiddeld na het voeren van het assessmentgesprek, tegen de verwachting in, geen verschillende beoordeling dan wanneer zij uitsluitend het assessmentdossier bekeken hebben ($M3_{exp.;\ int.;\ mom.\ 1}$ = 14.0; $M3_{exp.;\ int.;\ mom.\ 2}$ = 16.0; $SD_{exp.;\ int.;\ mom.\ 1}$ =5.5; $SD_{exp.;\ int.;\ mom.\ 2}$ = 5.5). Uit de vergelijking tussen de voorlopige toekenningen van studiepunten van moment 1 en van moment 2 door de externe assessoren uit de experimentele groep, bleek eveneens geen verschil te bestaan ($U(4, 5) = 6.6$, $p = .32$). Tegen de verwachting in geven de externe assessoren uit de experimentele groep gemiddeld na het voeren van het assessmentgesprek eveneens geen verschillende beoordeling dan wanneer zij uitsluitend het assessmentdossier bekeken hebben ($M3_{exp.;\ ext.;\ mom.\ 1}$ = 12.5; $M3_{exp.;\ ext.;\ mom.\ 2}$ = 14.0; $SD_{exp.;\ ext.;\ mom.\ 1}$ = 5.0; $SD_{exp.;\ ext.;\ mom.\ 2}$ = 5.5). Zowel de interne als de externe assessoren gaven in het tweede deel van het vooronderzoek, zoals verwacht, een significant hogere beoordeling bij moment 2 dan bij moment 1. Bij de interventie zijn in deze vergelijking echter geen verschillen gemeten.

Bij de interventie is eveneens gekeken naar de verschillen tussen de voorlopige toekenning van studiepunten van moment 2 en de uiteindelijke definitieve beoordeling. Deze vergelijkingen zijn zowel voor de controle groep, als voor de experimentele groep gemaakt. Daarnaast is binnen de controle en de experimentele groep ook een onderscheid gemaakt tussen de interne en de externe assessoren. Aangezien

de data binnen de vier vergelijkingen afhankelijk van elkaar zijn, is er vier keer gekozen voor de *Wilcoxon signed ranks test*.

Uit de vergelijking tussen de voorlopige toekenning van studiepunten van moment 2 en de uiteindelijke definitieve beoordeling gegeven door de interne assessoren uit de controle groep, bleek geen verschil te bestaan ($U(15, 15) = 12.1$, $p = 1.00$). Tegen de verwachting in, geven de interne assessoren uit de controle groep gemiddeld na het voeren van het assessmentgesprek geen verschillende beoordeling dan de uiteindelijke beoordeling ($M3_{contr.;\ int.;\ mom.\ 2} = 18.0$; $M3_{contr.;\ int.;\ beoord.} = 18.0$; $SD_{contr.;\ int.;\ mom.\ 2} = 4.1$; $SD_{contr.;\ int.;\ beoord.} = 4.1$). In tegenstelling tot de interne assessoren in het tweede deel van het vooronderzoek; zij gaven zoals verwacht significant lagere beoordelingen bij de uiteindelijke beoordeling dan bij de voorlopige toekenning van studiepunten bij moment 2. Uit dezelfde vergelijking, tussen de toekenning van studiepunten van moment 2 en de uiteindelijke definitieve beoordeling, maar dan gegeven door de externe assessoren uit de controle groep, bleek eveneens geen verschil te bestaan ($U(11, 11) = 11.9$, $p = 1.00$). De externe assessoren uit de controle groep geven gemiddeld na het voeren van het assessmentgesprek geen verschillende beoordeling dan de uiteindelijke beoordeling ($M3_{contr.;\ ext.;\ mom.\ 2} = 17.3$; $M3_{contr.;\ ext.;\ beoord.} = 17.3$; $SD_{contr.;\ ext.;\ mom.\ 2} = 4.7$; $SD_{contr.;\ ext.;\ beoord.} = 4.7$).

Dezelfde vergelijkingen zijn gemaakt voor de interne en externe assessoren afzonderlijk uit de experimentele groep. Uit de vergelijking tussen de voorlopige toekenning van studiepunten van moment 2 en de uiteindelijke definitieve beoordeling gegeven door de interne assessoren uit de experimentele groep, bleek geen verschil te bestaan ($U(5, 5) = 12.1$, $p = 1.00$). De interne assessoren uit de experimentele groep geven, tegen de verwachting in, gemiddeld na het voeren van het assessmentgesprek geen verschillende beoordeling dan de uiteindelijke beoordeling ($M3_{exp.;\ int.;\ mom.\ 2} = 16.0$; $M3_{exp.;\ int.;\ beoord.} = 16.0$; $SD_{exp.;\ int.;\ mom.\ 2} = 5.5$; $SD_{exp.;\ int.;\ beoord.} = 5.5$). In tegenstelling tot de interne assessoren in het tweede deel van het vooronderzoek; zij gaven zoals verwacht toch significant lagere beoordelingen bij de uiteindelijke beoordeling dan bij moment 2. Uit dezelfde vergelijking, tussen de voorlopige toekenning van studiepunten van moment 2 en de uiteindelijke definitieve beoordeling, maar dan gegeven door de externe assessoren uit de experimentele groep, bleek eveneens geen verschil te bestaan ($U(5, 5) = 4.9$, $p = .16$). De externe assessoren uit de experimentele groep geven gemiddeld na het voeren van het assessmentgesprek geen

hogere of lagere beoordeling dan de uiteindelijke beoordeling ($M3_{exp.;\ ext.;\ mom.\ 2}$ = 14.0; $M3_{exp.;\ ext.;\ beoord.}$ = 10.0; $SD_{exp.;\ ext.;\ mom.\ 2}$ = 5.5; $SD_{exp.;\ ext.;\ beoord.}$ = 10.0). Dit werd niet verwacht.

Vervolgens is gekeken naar of er een verschil bestaat tussen de controle en de experimentele groep van de voorlopige toekenning van studiepunten van moment 1 en van moment 2 en bij het geven van de uiteindelijke beoordeling. Aangezien de data binnen de drie vergelijkingen onafhankelijk van elkaar zijn, is er alle keren gekozen voor de *Mann-Whitney U test*.

Bij de vergelijking tussen de controle groep en de experimentele groep bij moment 1, bleek geen verschil te bestaan ($U(21, 9) = 7.3$, $p = .30$). Hoewel de hypothese een verschillende beoordeling voorspelde, geven de assessoren uit de controle groep gemiddeld na het bekijken van het assessmentdossier geen hogere of lagere beoordeling dan de assessoren uit de experimentele groep ($M3_{contr.;\ mom.\ 1}$ = 15.7; $M3_{exp.;\ mom.\ 1}$ = 13.3; $SD_{contr.;\ mom.\ 1}$ = 6.0; $SD_{exp.;\ mom.\ 1}$ = 5.0).

Bij de vergelijking tussen de controle groep en de experimentele groep van moment 2, bleek eveneens geen verschil te bestaan ($U(26, 10) = 4.3$, $p = .12$). De assessoren uit de controle groep geven, tegen de voorspelling van een verschillende beoordeling in, gemiddeld na het bekijken van het assessmentdossier geen hogere of lagere beoordeling dan de assessoren uit de experimentele groep ($M3_{contr.;\ mom.\ 2}$ = 17.7; $M3_{exp.;\ mom.\ 2}$ = 15.0; $SD_{contr.;\ mom.\ 2}$ = 4.3; $SD_{exp.;\ mom.\ 2}$ = 5.3).

Bij de vergelijking tussen de controle groep en de experimentele groep bij de beoordeling, bleek wel een significant verschil te bestaan ($U(26, 10) = 1.9$, $p = .03$). De assessoren uit de controle groep geven, zoals verwacht, gemiddeld bij de beoordeling een hogere uiteindelijke definitieve beoordeling dan de assessoren uit de experimentele groep ($M3_{contr.;\ beoord.}$ = 17.7; $M3_{exp.;\ beoord.}$ = 13.0; $SD_{contr.;\ beoord.}$ = 4.3; $SD_{exp.;\ beoord.}$ = 8.2). Dit komt voornamelijk doordat de controle groep hun mening niet meer verandert tussen het moment dat het assessmentgesprek gevoerd is en de uiteindelijke beoordeling, door de assessoren gezamenlijk bepaald. De experimentele groep verandert op dat moment nog wel van mening, zij stellen deze naar beneden bij.

4.3.1. Aanvullende resultaten uit de Hardop-denk-methode

Uit de transcripties (Zie Appendix 2, Tabel 3) blijkt dat assessoren tijdens de Hardop-denk-methode en tijdens het overleg tussen beide assessoren voor de definitieve beoordeling de meeste tijd besteden aan vier aspecten; (1) het verantwoordingsdocument, (2) de competenties, (3) de reflectie en (4) de koppeling tussen theorie en praktijk. Deze vier aspecten zijn met behulp van open, axiale en selectieve codering geanalyseerd en daaruit blijkt dat de assessoren veel waarde hechten aan de koppeling tussen de theorie en de persoonlijke praktijk van de student en dat deze koppeling geheel geïntegreerd moet zijn in de ontwikkeling van de student. Daarnaast prefereren assessoren het als studenten zowel op abstract niveau naar hun eigen ontwikkeling kunnen kijken, als dat zij de concrete praktijksituaties niet uit het oog verliezen.

De overwegingen die assessoren tijdens het invullen van de beoordelingsformulieren maken, zijn grofweg in te delen in vier categorieën. (1) *De eigen vraagstelling*; besteden de assessoren wel genoeg tijd aan de onderdelen, besteden zij aandacht aan *alle* onderdelen en geven zij de student de kans om onduidelijkheden toe te lichten en daarmee te bewijzen de competentie te beheersen? (2) *De beoordeling van het assessment*; de plaats die het gesprek inneemt in de beoordeling, uitgaande van het beheersen van de competentie totdat het tegendeel bewezen is en is het assessment volledig of meer dan volledig? (3) *Een mogelijke aanvulling*; het aanpassen of verbeteren van een oud bewijs of het toevoegen van nieuw bewijs op voldoende niveau. Tot slot (4) *overige opmerkingen*.

Tegen de verwachting in nemen de assessoren hun eigen vraagstelling en aandachtspunten onder de loep en nemen zij dit mee in de uiteindelijke beoordeling.

> *"Soms raak je minder aan bij een gesprek, maar ga je meer de diepte in."*
>
> *"Je kan niet alles aan bod laten komen."*
>
> *"Daar hebben we het ook niet over gehad."*
>
> *"We hebben er een aantal keer specifiek naar gevraagd."*
>
> *"Echt veel naar gevraagd, ook op andere manieren."*
>
> *"Hebben we in het gesprek daar wel genoeg op doorgevraagd?"*

Uit de resultaten van de Hardop-denk-methode blijkt, zoals verwacht, dat de assessoren zich bewust zijn van de impact die een beoordeling op de studieloopbaan en ontwikkeling van de student heeft. Dit is terug te zien in de citaten van de assessoren.

"Ze mag het ook toelichten in het gesprek, dus dan is het wel voldoende."

"*Ik heb geen aanleiding om op grond van het gesprek te zeggen dit wordt een O* [onvoldoende]*."*

"*Als het daarop vast blijft zitten, dan zou ik als assessor ingrijpen; een interventie afdwingen."*

"*Wel 30 punten, maar met een aanbeveling. Je kan altijd nog wat leren ook al is het voldoende."*

"*Wat ze moest aantonen zat er in, dus wat dat betreft moet je zeggen 30 punten."*

"*Gesprek was verhelderend, een toegevoegde waarde."*

"*Twijfel heeft ze weggenomen."*

Vanuit de beoordeling wordt vervolgens gekeken naar een mogelijke aanvulling. Uit de resultaten blijkt dat assessoren verschillend denken over de invulling van de aanvulling.

"*Nu heeft ze de kans het onder begeleiding te doen."*

"*Ik ben op zoek naar 'Wat werkt het beste?', 'Hoe help je haar het meeste verder?'"*

"*Bewijzen achteraf aanpassen, dat levert vaak niet zo veel op."*

"*Wat ze van mij ook mag doen, is dat ze gaat reflecteren op wat er gebeurd is tussen haar en haar mentor of tussen haar en de kinderen. En dat ze daar die theorie op los laat."*

De geciteerde uitspraken geven een goed beeld over bepaalde onderwerpen, daarnaast zijn er een aantal uitspraken van assessoren die waardevol zijn. De uitspraken hebben geen direct verband met het doel van het onderzoek, maar geven een beeld van hoe de assessoren zelf tegen de beoordelingen aankijken en hoe zij dus in het beoordelingsproces staan.

"*Voor mij een lastige, omdat ik daar helemaal geen zicht op heb, omdat dat mijn leergebied niet is."*

"*Dat is ook mijn vak, dus daar kijk ik een beetje extra naar."*

"*Als tweede assessor leg ik dat in eerste instantie bij de eerste assessor, hij heeft het duidelijkste beeld van het dossier op papier."*

"*Die dossiers zijn gigantisch groot."*

"*Je voelt dat het oprecht is."*

"*Dat is kennelijk iets dat ik anders zie."*

"*Bewijzen van voldoende niveau."* [assessor citeert]*: "Ik heb onvoldoende, jij hebt voldoende. Doen we hem* [Bewijzen zijn van voldoende niveau] *dan op voldoende niveau en dan competentie 7 onvoldoende?"*

"Dit is absoluut voldoende." [assessor omcirkelt 3]

"*Dat was gewoon goed … Dat is gewoon een 5."*

"*En bronnen dat was wat mij betreft oké."* [assessor omcirkelt 5]

"*Misschien ben ik wat streng."*

"*Ik ben iets coulanter."*

5. Discussie en conclusie

In het onderzoek is gekeken naar de subjectiviteit van assessmentbeoordelingen en mogelijke invloeden op de beoordelingen. Hierbij is een onderscheid gemaakt tussen de achtergrond van de interne versus externe assessoren, het moment dat de assessoren hun mening vormen en de waarde die de assessoren hechten aan de standaarden in beheersingsniveaus. Met behulp van de Hardop-denk-methode zijn de overwegingen van de assessoren tijdens het beoordelen en bij het gesprek waarbij de assessoren samen tot een definitieve beoordeling kwamen, duidelijk geworden. Wanneer er significante verschillen in de resultaten zichtbaar worden, kan er sprake zijn van subjectiviteit in de beoordeling. Vervolgens is er gekeken wat de oorzaak van deze subjectiviteit is.

Uit de resultaten van het eerste deel van het vooronderzoek blijkt dat de achtergrond van de interne en externe assessoren geen invloed heeft op de beoordelingen. In het tweede deel van het vooronderzoek is te zien dat veel assessoren na het voeren van het assessmentgesprek een ander oordeel hebben dan dat zij met uitsluitend het bekijken van het assessmentdossier gehad zouden hebben. Zowel bij de interne als de externe assessoren is een significant verschil gemeten. Het voeren van het assessmentgesprek heeft dus wel degelijk invloed op de beoordeling van het totale assessment. Frappant is dat alleen de interne assessoren significant van oordeel wijzigen na het voeren van het assessmentgesprek en de uiteindelijke definitieve beoordeling, dat in overleg met de andere assessor bepaald wordt. Een mogelijke verklaring hiervoor zou kunnen zijn dat interne assessoren zich door externe assessoren laten beïnvloeden, aangezien de externe assessoren zicht hebben op de praktijk en daarmee de capaciteiten van de student kunnen afzetten tegen de realiteit. Dit is echter niet op te maken uit de data van de Hardop-denk-methode. Anderzijds kunnen interne assessoren, die over het algemeen vaker assessments afnemen dan externe assessoren, flexibeler zijn geworden in de vorming hun oordeel. Mogelijkerwijs vertrouwen zij meer op het assessmentgesprek dan op het hun voorgelegde assessmentdossier. Gezien de omvang van dit dossier, kunnen interne assessoren eerder geneigd zijn studenten (onbewust) het voordeel van de twijfel te gunnen indien uit het assessmentgesprek naar voren lijkt te komen dat de student bepaalde aspecten wel heeft behandeld, die in eerste instantie door de assessor over het hoofd waren gezien. Aangezien interne assessoren de studenten ook al van tevoren kennen, weten zij beter dan externe assessoren wie zij op deze manier kunnen

bevoordelen. Het lijkt er op dat de bedreiging dat competentiebeoordelingen subjectief zijn (Luken, 2004) hier opspeelt.

Uit de interventie blijkt maar één significant verschil te bestaan; assessoren uit de experimentele groep geven significant lagere uiteindelijke definitieve beoordelingen dan zij na het voeren van het assessmentgesprek aanvankelijk bedacht hadden te geven. Dit is opmerkelijk, want bij het tweede deel van het vooronderzoek bestaan drie significante verschillen; tussen moment 1 en moment 2 bij zowel interne als externe assessoren en tussen moment 2 en de beoordeling bij interne assessoren. Mogelijk heeft de opzet van het vooronderzoek waarin de assessoren voor het eerst gevraagd werden op meerdere momenten een voorlopige toekenning van studiepunten te geven (in plaats van uitsluitend aan het eind van het assessment een definitieve beoordeling te geven) geleid tot significante verschillen tussen moment 1 en moment 2. De externe assessoren hebben door het gebruik van het gewijzigde beoordelingsformulier (een vijf puntsschaal in plaats van de twee puntsschaal in beheersingsniveau per competentie) wellicht bij moment 1 al een beter oordeel over het assessmentdossier kunnen vormen, waardoor er bij de interventie minder significante verschillen zijn ontstaan. Het opmerkelijke significante verschil tussen de voorlopige toekenning van studiepunten van moment 2 en de definitieve beoordeling door de assessoren uit de experimentele groep, kan verklaard worden door het feit dat de assessoren door de vijf puntsschaal een genuanceerder en daardoor kritischer beeld kunnen vormen en vervolgens in het gezamenlijk overleg eerder tot een lager oordeel komen dan dat zij aanvankelijk zelf hadden toegekend.

Met name uit de resultaten van de Hardop-denk-methode blijkt dat assessoren na het bekijken van het assessmentdossier meestal een duidelijk beeld hebben van de student. Assessoren bekijken het assessmentdossier aan de hand van verschillende onderdelen van het beoordelingsformulier. Bij het invullen van het beoordelingsformulier op moment 2, ten tijde van de Hardop-denk-methode, wordt dit beeld eveneens geschetst. Het beeld is over het algemeen gericht op de (in)competentie van de student, toch kan het beeld ook twijfels bevatten. Het gesprek wordt door de assessoren benut om het beeld te bevestigen, dan wel twijfels weg te nemen. Uit de uitspraken van assessoren blijkt dat zij in eerste instantie uitgaan van de competentheid van de student.

Los van het lage aantal significante verschillen in de interventie is het toch opmerkelijk dat de assessoren na het bekijken van het assessmentdossier en na het voeren van het assessmentgesprek zelden

grote meningsverschillen hebben. De achtergronden en werkervaring van de assessoren spelen nauwelijks tot geen rol. In slechts één situatie was er duidelijk sprake van onenigheid over de beoordeling van een onderdeel van het assessment. De assessoren hebben een compromis bereikt door te zoeken naar een soortgelijk onderdeel, waarmee de meningsverschillen gelijk getrokken zijn. In dit specifieke geval ging het om twee nauw verwante beoordelingsaspecten, waarvan er uiteindelijk één op 'voldoende' werd gescoord en de ander als 'onvoldoende' werd beoordeeld. Uit dit gegeven kan geconcludeerd worden dat de bedreiging van subjectieve competentiebeoordelingen (Luken, 2004) toch ongegrond is.

Het lage aantal significante verschillen is te verklaren door de *doorzichtigheid* (Klarus, 1998) van de *procedure* (Messick, 1989) van de assessmentbeoordelingen bij Hogeschool Domstad. Eveneens blijkt uit de transcripties van de Hardop-denk-methode dat de assessoren voor het grootste deel hetzelfde beeld hebben van de gehanteerde beoordelingscriteria en de beoordelingsmethode, in tegenstelling tot waarvoor Luken (2004) waarschuwde.

Een andere verklaring voor deze resultaten betreft het aantal assessoren dat deelgenomen heeft aan het onderzoek, dit is lager dan vooraf verondersteld werd. Dit komt mogelijk door de werkdruk die ten tijde van het afnemen van assessments hoger is dan buiten deze periodes, daarnaast is er een aantal assessments niet doorgegaan wegens terugtrekking door de student. Deze situatie heeft invloed op de conclusies die uit de resultaten getrokken kunnen worden; een kleinere onderzoeksgroep bemoeilijkt het verkrijgen van significante resultaten aanzienlijk. Ten slotte was het gebruik van de Hardop-denk-methode nieuw voor de assessoren en heeft dit mogelijk ook invloed gehad op de uitkomsten van dit onderzoek. Aangezien assessoren voor het eerst in aanraking kwamen met een onderzoeksmethode waarbij zij al hun gedachten die betrekking hadden tot de assessments moesten verwoorden, is er een zekere kans dat assessoren onbewust niet alles hebben gezegd dat zij dachten en dat daardoor informatie verloren is gegaan (Ericsson & Simon, 1984).

Gestoeld op de door Luken (2004) gesignaleerde problemen werd verwacht dat er een aantal significante verschillen in de beoordelingen zichtbaar zouden worden. Deze verschillen zouden zich met name richten op de inhoud van de competenties en de directe koppeling met de praktijk. Uit de data van de Hardop-denk-

methode is gebleken dat hier geen verschillen bestonden die doorgewerkt hebben in de beoordelingen. Vanuit de voorwaarde *doorzichtigheid* voor validiteit van Klarus (1998), het facet *proces* van Messick (1989) en de data van het vooronderzoek werd verwacht dat de beoordelingen tussen moment 1 en moment 2 door zowel de interne als de externe assessoren ook in de interventie significant zouden verschillen, dit was echter niet het geval; hier was uitsluitend in het vooronderzoek sprake van. Volgens Eraut (1994) zou er geen significant verschil zichtbaar zijn tussen moment 2 en de uiteindelijke beoordeling bij de interne assessoren, dit was echter in het vooronderzoek wel het geval. In de interventie bleek er bij die vergelijking geen significant verschil te bestaan.

Hoewel er in dit onderzoek geprobeerd is de data op een zo betrouwbaar en valide mogelijke wijze te verzamelen, zijn er een aantal aspecten die mogelijk invloed hebben gehad op de verkregen data. Voor zover de opzet van het onderzoek dat toeliet, zijn de assessoren a select verdeeld over de controle en de experimentele groep. Ericsson en Simon (1984) benadrukken dat het verstandig is om voor het gebruik van de Hardop-denk-methode, de deelnemers deze werkwijze te laten oefenen. Dit is wegens het tijdsbestek waarin het onderzoek is uitgevoerd niet gebeurd, waardoor assessoren tijdens het onderzoek voor het eerst in aanraking kwamen met de Hardop-denk-methode. Mogelijk heeft de methode hierdoor minder informatie opgeleverd dan wenselijk is. (Gebrek aan) tijd was op een ander vlak een factor van betekenis; voor assessments is er een bepaalde hoeveelheid tijd gereserveerd, helaas kon dit roostertechnisch niet verruimd worden ten behoeve van het onderzoek.

Afgezien van het geringe aantal (significante) verschillen die bij de vergelijking van de onderzochte factoren zijn gebleken, is het toch nuttig om regelmatig de objectiviteit van de beoordelingsmethode te toetsen. De meest logische manier om de objectiviteit van de beoordelingen van de assessments vast te stellen, is kruislings beoordelen. Dit houdt in dat een student door twee assessorparen beoordeeld wordt en dat daarnaast de assessorparen ook van samenstelling wisselen. Vervolgens wordt er gekeken naar het aantal studiepunten dat de student in de verschillende assessorsamenstellingen toegekend krijgt. Een andere wijze om de objectiviteit van beoordelingen van assessments vast te stellen is dat de assessoren worden beoordeeld tijdens het beoordelen. Van der Schaaf en collega's (2008) hanteren in hun onderzoek weer een

andere aanpak om een indicatie voor de validiteit te krijgen. Zij analyseren de inhoud van de docentportfolio's kwalitatief en vergelijken de uitkomsten daarvan met de beoordelingen van de leerlingen en de beoordelaars van de docenten.

Helaas waren deze manieren van meten om de objectiviteit vast te stellen bij dit onderzoek niet mogelijk, aangezien de extra tijdsdruk die hiermee bij de studenten en de assessoren komt te liggen te hoog zou zijn. Ook het tijdsbestek waarin dit onderzoek uitgevoerd is, bood geen ruimte aan bovenstaande uitwerkingen. In de toekomst is het a select uitvoeren van kruislings beoordelen en het beoordelen van assessoren een goede wijze om de objectiviteit te waarborgen.

Naast het vaststellen van de objectiviteit van het beoordelen is het ook interessant om te weten welke standaarden, de twee- of de vijf puntsschaal in beheersingsniveaus, door de assessoren geprefereerd wordt. Daarnaast kan onderzocht worden of studenten een verschil merken of een voorkeur hebben voor een van beide standaarden in beheersingsniveaus.

Uit de resultaten van dit onderzoek blijkt dat geen van de deelhypothesen geheel ondersteund wordt door uitsluitend significante verschillen in de vergelijkingen. Hieruit kan de conclusie getrokken worden dat geen van de deelhypothesen blindelings aangenomen kan worden en daarmee wordt de nulhypothese aangenomen en de hypothese verworpen. Doordat er wel significante verschillen zijn gebleken, moet verder onderzoek uitwijzen of het terecht is dat alle (deel)hypothesen verworpen dienen te worden. Hieruit voortvloeiend kan met deze resultaten wel gesteld worden dat de problemen, voorzien door Luken (2004), niet van toepassing waren bij deze beoordelingsmethode. Daarentegen kan met grote zekerheid aangenomen worden dat in deze beoordelingsmethode aan de voorwaarden om de validiteit te waarborgen (Klarus, 1998) en de validiteitfacetten van Messick (1989) voldaan is. Uit de resultaten van dit onderzoek kan niet worden vastgesteld dat de assessmentbeoordelingen aan Hogeschool Domstad subjectief verlopen of dat de assessoren de beoordelingsmethode, bestaande uit een assessmentdossier en een assessmentgesprek, op ongelijke wijze toepassen.

Om een completer beeld van de assessmentbeoordelingen te krijgen, is verder onderzoek nodig. Ten eerste kan de inhoud van de assessments door onafhankelijke onderzoekers beoordeeld worden, zodat deze vergeleken kunnen worden met de beoordelingen die de assessoren bij het betreffende assessment

geven. Ten tweede kunnen de assessoren tijdens het gehele proces van de beoordeling gevolgd en beoordeeld worden; met behulp van de Hardop-denk-methode zal veel informatie verkregen worden. Een derde mogelijkheid om het onderzoek op dit terrein voort te zetten, is het uitvoeren van kruislingse beoordelingen, waarbij de assessorenparen in verschillende samenstellingen hetzelfde assessment beoordelen. Een voorwaarde voor deze uitbreidingen is wel dat er genoeg tijd per assessment ingeruimd wordt, zodat dit geen invloed op de kwaliteit zal hebben. Ten vierde is het mogelijk in het vervolgonderzoek specifiek te richten op de invloed van de eerste en de tweede beoordelaar op de beoordeling. In dit onderzoek is geen aanwijzing geweest dat de eerste dan wel de tweede assessor de mening van de andere assessor overschaduwde. Toch is inzicht op dit terrein in de toekomst interessant, aangezien hiermee de toegevoegde waarde van twee assessoren, in plaats van één assessor, bepaald kan worden. Een laatste mogelijkheid is dat onderzocht worden welke standaarden in beheersingsniveaus geprefereerd worden door de assessoren en door de studenten.

Literatuurlijst

Cluitmans, J., & Klarus, R. (2005). Competentiebeoordeling: een pleidooi voor congruentie. *Tijdschrift voor Hoger Onderwijs, 23, 4*, 221-238.

Crooks, T. J., & Kane, M. T. (1996). Threats to the valid use of assessments. *Assessment in Education: Principles, Policy & Practice, 3, 3*, 265-286.

Elshout-Mohr, M., & Oostdam, R. (2001). *Assessment van competenties in een dynamisch curriculum*. Amsterdam: SCO-Kohnstamm Instituut.

Eraut, M. (1994). *Developing professional knowledge and competence*. London and New York: Routledge Falmer.

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

Gipps, C. V. (1994). *Beyond testing. Towards a theory of educational assessment*. London and New York: Routledge Falmer.

Gonczi, A. (1994). Competency based assessment in the professions in Australia. *Assessment in Education, 1, 1*, 27-45.

Habermas, J. (1990). *Moral consciousness and communicative action* (C. Lenhardt, & S. W. Nicholsen, Trans.). Cambridge: MIT Press. (Original work published xxxx).

Klarus, R. (1998). *Competenties Beoordelen*. Den Bosch / Nijmegen: CINOP/KUN. (dissertatie).

Kuijpers, M. A. C. T. (2003). *Loopbaanontwikkeling: Onderzoek naar 'Competenties'*. Proefschrift. Universiteit Twente. Enschede: Twente University Press.

Luken, T. P. (2004). Zijn competenties meetbaar? Dilemma en uitweg bij het werkbaar maken van het competentiebegrip. *Tijdschrift voor Hoger Onderwijs, 22, 1*, 38-53.

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

Moss, P. A., & Schutz, A. (2001). Educational Standards, Assessment, and the Search for Consensus. *American Educational Research Journal, 39, 1*, 37-70.

Myers, C. B., & Myers, S. M. (2007). Assessing assessment: The effects of two exam formats on course achievement and evaluation. *Innovative Higher Education, 31*, 227-236.

Robson, C. (2002). *Real world research*. Malden, Oxford and Carlton: Blackwell Publishing.

Stiggins, R. J., & Bridgeford, N. J. (1985). The ecology of classroom assessment. *Journal of Educational Measurement, 22, 4*, 271-286.

Stokking, K. M., Van der Schaaf, M. F., Jaspers, J., & Erkens, G. (2004). Teachers' assessment of students' research skills. *British Educational Research Journal, 30, 1*, 93-116.

Van Merriënboer, J. J. G., Van der Klink, M. R., & Hendriks, M. (2002). *Competenties: Van complicaties tot compromis; Over schuifjes en begrenzers; Een studie in opdracht van de Onderwijsraad*. Den Haag: Onderwijsraad.

Van der Schaaf, M. F., Stokking K. M., & Verloop, N. (2008). De validiteit van het beoordelen van docentcognities en docentgedrag in docentbeoordelingen. *Pedagogische Studiën, 85*, 222-239.

Appendix 1

Tabel 1

Interventie controle groep

Verantwoordingsdocument bestaande uit onderwijsvisie, beroepsrollen en levensbeschouwelijke identiteit (werkconcept)	V	O
Bewijzen voor de competenties zijn van voldoende niveau; per kern drie goede bewijzen (twee bewijzen per kern bij comp. 5 en 6)	V	O
Waarvan competentie 1 (interpersoonlijk competent)	**V**	**O**
Waarvan competentie 2 (pedagogisch competent)	**V**	**O**
Waarvan competentie 3 (didactisch competent	**V**	**O**
Bewijzen zijn geïntegreerd aanwezig voor competentie 7 inclusief verslaglegging eigen leerproces (waaronder POP)	V	O
Feedback vanuit O&P en POW, inclusief zichtbare en voldoende verwerking door student, is aanwezig (één leergebied per assessment)	V	O
ICT is op voldoende niveau en geïntegreerd ingezet bij de competenties en leergebieden	V	O
Bewijzen zijn voorzien van reflectie van de student, er is voldoende diepgang in de reflecties (Korthagen)	V	O
Bewijzen zijn authentiek en relevant voor de betreffende competentie	V	O
Koppeling tussen theorie en praktijk waaronder het toepassen van de cyclus van Kolb en verslaglegging 5 feedbackvragen	V	O
Bronnenoverzicht volgens richtlijnen op de portal	V	O
Aantal ECTS 0 20 30		

Tabel 2

Interventie experimentele groep

	O				V
Verantwoordingsdocument bestaande uit onderwijsvisie, beroepsrollen en levensbeschouwelijke identiteit (werkconcept)	1	2	3	4	5
Bewijzen voor de competenties zijn van voldoende niveau; per kern drie goede bewijzen (twee bewijzen per kern bij comp. 5 en 6)	1	2	3	4	5
Waarvan competentie 1 (interpersoonlijk competent)	**1**	**2**	**3**	**4**	**5**
Waarvan competentie 2 (pedagogisch competent)	**1**	**2**	**3**	**4**	**5**
Waarvan competentie 3 (didactisch competent)	**1**	**2**	**3**	**4**	**5**
Bewijzen zijn geïntegreerd aanwezig voor competentie 7 inclusief verslaglegging eigen leerproces (waaronder POP)	1	2	3	4	5
Feedback vanuit O&P en POW, inclusief zichtbare en voldoende verwerking door student, is aanwezig (één leergebied per assessment)	1	2	3	4	5
ICT is op voldoende niveau en geïntegreerd ingezet bij de competenties en leergebieden	1	2	3	4	5
Bewijzen zijn voorzien van reflectie van de student, er is voldoende diepgang in de reflecties (Korthagen)	1	2	3	4	5
Bewijzen zijn authentiek en relevant voor de betreffende competentie	1	2	3	4	5
Koppeling tussen theorie en praktijk waaronder het toepassen van de cyclus van Kolb en verslaglegging 5 feedbackvragen	1	2	3	4	5
Bronnenoverzicht volgens richtlijnen op de portal	1	2	3	4	5
Aantal ECTS 0 20 30					

Appendix 2

Tabel 3

Resultaat coderingsproces

Begrippen	Open coding	Axiale coding	Selectieve coding
Verantwoordingsdocument	Duidelijke visie		Theoretische onderbouwing van
	Tonen doorgemaakte ontwikkeling		concrete ontwikkeling
	Beeld van ontwikkelingskansen	Ontwikkeling	
	Zelfbewust		
	Helikopterview	Theorie	
	Theoretische onderbouwing visie		
	Koppeling levensbeschouwelijke identiteit en theorie	Persoonlijke	
	Praktijksituaties aan visie gekoppeld	praktijksituaties	
	Persoonlijk beeld van beroepsrollen		
	Helemaal uitgewerkt		
Competenties	Tonen proces		
	Abstract niveau		
	Echte bespreking competenties		
	Betrokkenheid		
	Reflectie	Proces op abstract	
	Aantonen meeropbrengst	niveau	
	Sterkte- zwakte analyse		
	Goede onderbouwing keuzes		
	Juiste koppeling naar kern		
	Heldere praktijkvoorbeelden bij kernen		
	Pendel theorie – praktijk	Theorie integreren in	
	Theorie toegepast	eigen praktijk	
	Theorie bewust inzetten		
	Koppeling doelen aan praktijk		
	Leiderschapsstijlen		
	Kennis leerlijnen		
	Vakoverstijgend	Ongebonden	
	Loslaten methode		
	Onderwijsresultaten meten		
	Aangeven meerwaarde van groepsoverzichten		
	Aangeven meerwaarde van groepsplannen		
	Levendige voorbeelden		

Tabel 3 (voortzetting)

Resultaat coderingsproces

Begrippen	Open coding	Axiale coding	Selectieve coding
Reflectie	Plan van aanpak leerproces		
	Heldere persoonlijke leerdoelen		
	Sterkte- zwakte analyse	Beschrijving leerproces	
	Beschrijving leerproces		
	Doorlopende ontwikkelingslijn		
	Afwegingen tonen		
	Slotanalyse		
	Bewijzen voorzien van reflectie	Reflecteren op concrete	
	Reflecteren in het gesprek	aspecten	
	Feedback van SLB'er		
	Concreet maken		
	Diepgang		
Koppeling theorie – praktijk	Theorie geïntegreerd in lespraktijk		
	Onderzoekende houding		
	Handelen theoretisch onderbouwd	Theorie geïntegreerd in	
	5 leervragen van Kolb	les- en leerproces	
	Kennis van termen		
	Natuurlijke opbouw		
	Diepgang		