De kwaliteit van bewijsvoering bij zelfevaluatie van competentie assessment programma's

Masterthesis Onderwijskunde

Universiteit Utrecht

Juni 2009

Auteur: ing. Richard Spithoven

Studentnummer: 3038440

Begeleider en beoordelaar: Dr. Liesbeth K.J. Baartman

Tweede Beoordelaar: Dr. Jeroen Janssen

1. Samenvatting

Er is een groeiende interesse in zelfevaluatie-instrumenten in het veld van kwaliteitsgarantie in het onderwijs. Daarbij is de rol van bewijsvoering ter onderbouwing van het kwaliteitsoordeel onderbelicht gebleven. In dit onderzoek staat daarom de kwaliteit van bewijsvoering bij de zelfevaluatie van Competentie Assessment Programma's (CAP) centraal. Dit onderzoek is uitgevoerd bij drie faculteiten van een hogeschool in Nederland. Bij één van de faculteiten heeft een training plaatsgevonden waarin aandacht is geschonken aan het verzamelen van bewijzen en argumenten bij zelfevaluatie. Om de kwaliteit van bewijsvoering in kaart te brengen is een codeerschema ontwikkeld dat de kwaliteit beoordeelt vanuit soorten bewijsvoering. Deze soorten bewijsvoering zijn onttrokken aan inzichten uit de argumentatietheorie. Het onderzoek laat zien dat deze training de algehele kwaliteit van bewijsvoering verbeterde gedurende de zelfevaluatie. Daarbij is geen verschil gevonden in de kwaliteit van bewijsvoering tussen verschillende evaluatieaspecten. Bovendien zijn er geen verschillen gevonden in de kwaliteit van bewijsvoering in gesprekken waarbij deelnemers voorafgaand aan de evaluatie *veel consensus* dan wel *in enige mate consensus* laten zien.

2. Introductie

Werknemers van de toekomst moeten kunnen redeneren, problemen oplossen, werken in teams, communiceren, initiatieven nemen en diverse perspectieven op hun werk vormen. Het onderwijs heeft een verandering moeten ondergaan vanuit het industriële tijdperk naar het informatietijdperk (Reigeluth, 1999). Om deze aansluiting van het onderwijs op het werkveld te verbeteren, zijn Nederlandse hogescholen wettelijk gebonden een competentie-gebaseerd curriculum te adopteren. Het doel van dit curriculum is om de benodigde competenties (geïntegreerde gehelen van kennis, vaardigheden en houdingen) van de opleiding beter op de toekomstige werksituatie aan te laten sluiten. Daarnaast verandert de sturing van de overheid in het onderwijs. De regulatie van de overheid vindt tegenwoordig steeds meer plaats door eisen te stellen, waarbij scholen hun eigen beleid maken en bewijzen dat zij aan deze eisen voldoen. Door deze toename van eigen verantwoordelijkheid van onderwijsinstellingen is er volgens van Petegem, Deneire en De Maeyer (2008) een groeiende interesse in zelfevaluatie-instrumenten om de kwaliteit in hoger, middelbaar en lager onderwijs te

stimuleren en te garanderen. Een zelfevaluatie of interne evaluatie is een evaluatie die onder de eigen verantwoordelijkheid van de school valt en door de school zelf wordt uitgevoerd. In het veld van kwaliteitsgarantie op hogescholen neemt zelfevaluatie een steeds belangrijkere plaats in. Veel onderzoek naar zelfevaluatie heeft volgens Fournier en Smith (1993) in het teken gestaan van het oplossen van problemen, beantwoorden van vragen, formuleren van argumenten, het toepassen van resultaten en de opzet en methoden om betrouwbare data te verzamelen. De rol van bewijsvoering bij zelfevaluatie is daarbij onderbelicht gebleven. McNamara en O'Hara (2005) geven daarbij aan dat zonder goede bewijzen er geen waardevolle evaluatie gehouden kan worden. Er is volgens hen gebrek aan duidelijkheid in de bewijzen die scholen moeten tonen om zelfevaluatie te kunnen onderbouwen. In dit onderzoek wordt daarom ingegaan op de kwaliteit van bewijsvoering van hogescholen ter ondersteuning van de zelfevaluatie van Competentie Assessment Programma's (CAP).

2.1 Aanleiding

Het belang van bewijsvoering in de evaluatietheorie is reeds onderschreven door Fournier en Smith (1993). Zelfevaluatie staat volgens hen onder andere in het teken van het genereren van alle bewijsvoeringen die er voor zorgen dat conclusies gerechtvaardigd zijn. Daarbij heeft volgens Kyriakides en Campbell (2004) het verzamelen van empirische data veel aandacht gehad in de evaluatietheorie, omdat het de waarde van de wetenschappelijke rationaliteit benadert. Bewijsvoering tijdens zelfevaluatie is echter verder vorm te geven dan alleen door empirische data als wetenschappelijke rationaliteit. In dit onderzoek zal daarom het begrip bewijsvoering tijdens zelfevaluatie verder uitgediept en toepasbaar worden gemaakt. Hierbij wordt gebruik gemaakt van een zelfevaluatie-instrument dat is ontwikkeld door Baartman, Bastiaens, Kirschner en Van der Vleuten (2006).

Baartman en collega's hebben een zelfevaluatie-instrument ontwikkeld waarmee hogescholen de kwaliteit van hun CAP kunnen beoordelen. Dit instrument past binnen de reeds gesignaleerde traditie van de groeiende verantwoordelijkheid van hogescholen om zelf de kwaliteit van hun onderwijs en assessment te evalueren en te garanderen. Het zelfevaluatie-instrument bestaat uit 12 kwaliteitscriteria en een procedure waarbij deelnemers aan de evaluatie (docenten, studenten en het werkveld) eerst

individueel en daarna gezamenlijk in een groepsdiscussie de kwaliteit van het CAP bepalen en hiervoor bewijzen verzamelen. (Voor meer informatie en de achtergronden wordt verwezen naar het onderzoek van Baartman en collega's uit 2006).

In eerder onderzoek naar het gebruik van het zelfevaluatie-instrument (Baartman, Bastiaens, Kirschner en Van der Vleuten (2007)) zijn twee faculteiten van een hogeschool ondersteund in de evaluatie van hun CAP. Het onderzoek heeft uitgewezen dat de deelnemers hun individuele bewijs over de kwaliteit van het CAP voornamelijk baseren op persoonlijke ervaringen. Deze vorm geldt als een lage vorm van bewijsvoering. Het gehanteerde schema in het onderzoek van Baartman en collega's kwalificeert vormen van bewijsvoering enkel en alleen door te kijken de aan- of afwezigheid van bronnen die gebruikt worden om het bewijs te onderbouwen. In dit onderzoek wordt de kwaliteit van bewijsvoering echter beoordeeld door de te kijken naar de soort bewijsvoering. Het onderzoek vindt plaats op drie faculteiten van een hogeschool in Nederland. Op één van deze faculteiten werd voorafgaand aan de zelfevaluatie een training verzorgd over het verzamelen van bewijzen en argumenten bij zelfevaluatie.

2.2 Onderzoeksvragen

In dit onderzoek staan de volgende onderzoeksvragen centraal:

1. Hoe ziet de kwaliteit van bewijsvoering er op de drie faculteiten uit?
2. Is er een verschil in de mate van kwaliteit van bewijsvoering bij de verschillende faculteiten?
3. Is er een verschil in kwaliteit van bewijsvoering per kwaliteitscriterium van het CAP?
4. Is er een verschil in kwaliteit van bewijsvoering in gesprekken waarbij de deelnemers veel dan wel weinig consensus laten zien?

Bij de tweede onderzoeksvraag wordt de hypothese getoetst, dat op de faculteit waarop voorafgaand aan de zelfevaluatie een training werd verzorgd over het verzamelen van bewijzen en argumenten, de kwaliteit van bewijsvoering hoger zal zijn dan op de faculteit waarop deze training niet heeft plaatsgevonden.

3. Theoretische achtergronden

In dit onderdeel wordt ingegaan op de definitie van argumentatie en verschillende typen gesprekken. Ook worden de kritieken op het gebruik van argumentatietheorie uiteengezet. Het gebruikte schema in eerder onderzoek (Baartman en collega's, 2007) wordt weergegeven in paragraaf 3.3. Om de kwaliteit van bewijsvoering in de context van argumenteren in groepen te kunnen beschrijven, wordt in dit onderzoek gebruik gemaakt van modellen uit de argumentatietheorie. Er wordt in paragraaf 3.3 een overzicht gegeven van de gevonden modellen uit de argumentatietheorie die soorten bewijzen verder uitwerken dan door alleen te kijken naar de aan- of afwezigheid van bewijzen. Daarbij wordt ook ingegaan op begrippen uit het validiteitvraagstuk. De beschreven modellen worden in paragraaf 4.2 uitgewerkt tot een codeerschema om de kwaliteit van bewijsvoering te kunnen beoordelen.

3.1 Definitie van argumentatie en verschillende type gesprekken

De definitie van argumentatie wordt veelal tweeledig opgevat. Kuhn en Udell (2003) geven dit in hun onderzoek duidelijk aan. De uitdrukkingen argument en argumentatie geven volgens de auteurs twee bedoelingen aan. De uitdrukking argument duidt op *product* en argumentatie op *proces*. Een individu zet een argument in elkaar om een bewering te ondersteunen. Het dialogische proces waarbij twee of meer mensen zich verbinden in een debat van tegengestelde beweringen kan worden gezien als argumentatie of argumentatieve discussie (om het te onderscheiden van argument als product). Het argument als product is desalniettemin een eigenschap van argumentatieve discussie. De twee uitdrukkingen van argument en argumentatie zijn daarom gerelateerd. Sampson en Clark (2008) sluiten zich hierbij aan en noemen de uitdrukking argument een kunstproduct (artefact) dat mensen creëren om hun beweringen of uitleg helder te verwoorden en te rechtvaardigen. De uitdrukking argumentatie is het complexe proces van het genereren van deze artefacten. De auteurs geven aan dat deze onderscheiding niet absoluut moet worden gezien. Zij doelen daarmee op de relatie tussen argument en argumentatie. Felton en Kuhn (2002) zien argumentatie als een sociale activiteit waarbij twee of meer mensen toenadering zoeken, argumenten verdedigen en argumenten vergelijken in ondersteunende of tegengestelde posities. Deze definitie van argumentatie is van toepassing in de context van dit onderzoek. De relatie tussen argument en argumentatie komt in deze definitie niet

uitvoerig naar voren. Dat maakt de definitie overzichtelijker. Daarbij is het aannemelijk dat in een groepsdiscussie over de kwaliteit van het CAP de verschillende deelnemers niet alleen tegengestelde posities in zullen nemen. Er zal ook uit gelijke positie toenadering gezocht kunnen worden over de kwaliteit van het CAP. Waar los van het begrip argumentatie over argument wordt gesproken wordt de definitie van Sampson en Clark (2008) aangehouden.

Ook is het mogelijk om vanuit verschillende type gesprekken naar de definitie van argumentatie te kijken. Walton (1989) onderscheidt drie typen van gesprekken verwant aan discussiëren, namelijk overtuigende dialoog, onderzoek en onderhandeling. De context van dit onderzoek heeft overeenkomsten met het tweede soort gesprek dat Walton onderscheidt: het *onderzoek*. Hierbij is het doel dat twee of meer personen kennis uitbreiden op een bepaald gebied. Het *onderzoek* zoekt bewijs, of een bevestiging van een eerder geformuleerde conclusie. De definitie van Felton en Kuhn (2002) dat voor dit onderzoek geselecteerd is vindt met dit type gesprek van Walton (1989) overeenkomst. Met name waar het gaat om toenadering zoeken én het verdedigen en vergelijken van argumenten in ondersteunende of tegengestelde posities. De kenmerken van de groepsdiscussie in dit onderzoek sluiten ook aan op het begrip argumentatieve discussie, zoals Baker en collega's (2007) dit definiëren. Volgens hen is een argumentatieve discussie een proces dat is georiënteerd op *consensus* of *geen consensus*. Hierin worden standpunten aan elkaar gekoppeld en hierdoor kan de mate van acceptatie ten aanzien van een standpunt veranderen. Zij veronderstellen dat het doel van een argumentatieve discussie niet het winnen van een discussie is, maar juist begrip bewerkstelligen bij deelnemers. Ook deze definitie vindt aansluiting bij de geselecteerde definitie van argumentatie van Felton en Kuhn (2002). In de definitie van Baker en collega's (2007) wordt ook over een mate van toenadering gesproken zoals dat in de definitie van Felton en Kuhn (2002) het geval is. Ook Felton en Kuhn hebben in hun onderzoek gekeken naar de indelingen van typen gesprekken. Zij onderscheiden in hun experimenteel onderzoek drie soorten gesprekken: een gesprek waarbij de deelnemers het met elkaar oneens zijn, een gesprek waarbij men het met elkaar eens is en een gespreksvorm waarbij men het met elkaar noch eens noch oneens is. In dit onderzoek wordt in onderzoekvraag vier ingegaan op het verschil in kwaliteit van bewijsvoering in gesprekken waarin deelnemers veel dan wel weinig consensus laten zien. De inhoudelijke vergelijking van definities laat veel overlap zien tussen de

definitie van argumentatie én de verschillende type gesprekken. De driedeling van type gesprekken van Felton en Kuhn vormt de basis voor de indeling van type gesprekken in dit onderzoek. De driedeling wordt niet letterlijk overgenomen omdat er gesproken wordt over een mate van eens zijn. In dit onderzoek gaat het meer om het bereiken van consensus. Waar in dit onderzoek gesproken wordt over type gesprekken wordt daarmee het type *onderzoek* van Walton (1989) bedoeld waarbij de definitie van Baker en collega's (2007) wordt gesplitst naar gesprekken waarin deelnemers *veel consensus, in enige mate consensus* en *geen consensus* laten zien.

3.2 Argumentatietheorieën en de kritieken

Volgens Clark, Sampson, Weinberger en Erkens (2007) is het werk van Toulmin (1958) waarschijnlijk het meest geciteerde raamwerk voor het meten van argumentatiekwaliteit. Toulmins model geeft een overzicht van de kwaliteit van verschillende argumenten, gebaseerd op de aanwezigheid of afwezigheid van verschillende componenten van argumentatie en hun interrelaties. Ook het werk van Erduran, Simon en Osborne (2004) wordt vaak genoemd in onderzoeken naar argumentatiekwaliteit. Erduran en collega's gaan gedetailleerd in op het werk van Toulmin (1958) en geven aan welke structurele componenten van een argument gebruikt kunnen worden om de kwaliteit van argumentatie te classificeren. De kracht van de modellen van Toulmin en Erduran en collega's (2004) is dat ze indicatoren van de sterkte van argumentatie onderscheiden die los staan van het onderwerp, waardoor de modellen in elke context toepasbaar zijn. Het nadeel van de modellen van Toulmin (1958) en Erduran en collega's (2004) is de overkoepelende zienswijze die puur gericht is op de componenten van argumentatie in plaats van op de inhoudelijke juistheid van argumentatie. Veel critici geven aan dat er ook beslissingen over de inhoudelijke juistheid van argumenten genomen moeten worden. Men kan volgens deze critici de kwaliteit niet alleen uitdrukken in de aanwezigheid of afwezigheid van structurele componenten van argumentatie. De definitie van Felton en Kuhn (2002) die voor dit onderzoek is geselecteerd past in deze discussie. Deze definitie past in een trend die Felton en Kuhn beschrijven. Ze zien een verschuiving van de wiskundige modellen van formele logica die argumenten scheiden van de context waarin deze ontstaan naar meer normatieve modellen die gebaseerd zijn op

een sociaal bouwwerk van argumentatie. Deze benadering voorziet in waardevolle informatie over de inhoudelijke juistheid van de argumenten die mensen kunnen genereren.

Als voorbeeld van onderzoek dat oog heeft voor de inhoudelijke juistheid van argumentatie én de componenten van argumentatie gelden de onderzoeken van Sandoval (2003) en Sandoval en Millwood (2005). Een kwaliteitsoordeel wordt door Sandoval en Millwood gedefinieerd als een oordeel over de structuur van een argument én de inhoudelijke adequaatheid. De basis daarvoor is een scheiding tussen twee kwaliteitsaspecten van argumentatie, namelijk de conceptuele kwaliteit van argumentatie (inhoudelijke juistheid) en de epistemologische kwaliteit van argumentatie (componenten van argumentatie). De conceptuele kwaliteit wordt in het onderzoek omschreven als de adequaatheid of juistheid van het inhoudelijke argument. Dit wordt door Sandoval en Millwood vormgegeven door alle mogelijke argumenten over een redelijk gesloten wetenschappelijke discussie (omtrent natuurlijke selectie) te verzamelen en te voorzien van een classificatie naar de adequaatheid of juistheid van het inhoudelijke argument. De epistemologische kwaliteit wordt in het onderzoek omschreven als de structurele juistheid van het argument en wordt vormgegeven door de argumentatietheorieën van Toulmin (1958) en Erduran en collega's (2004) te gebruiken.

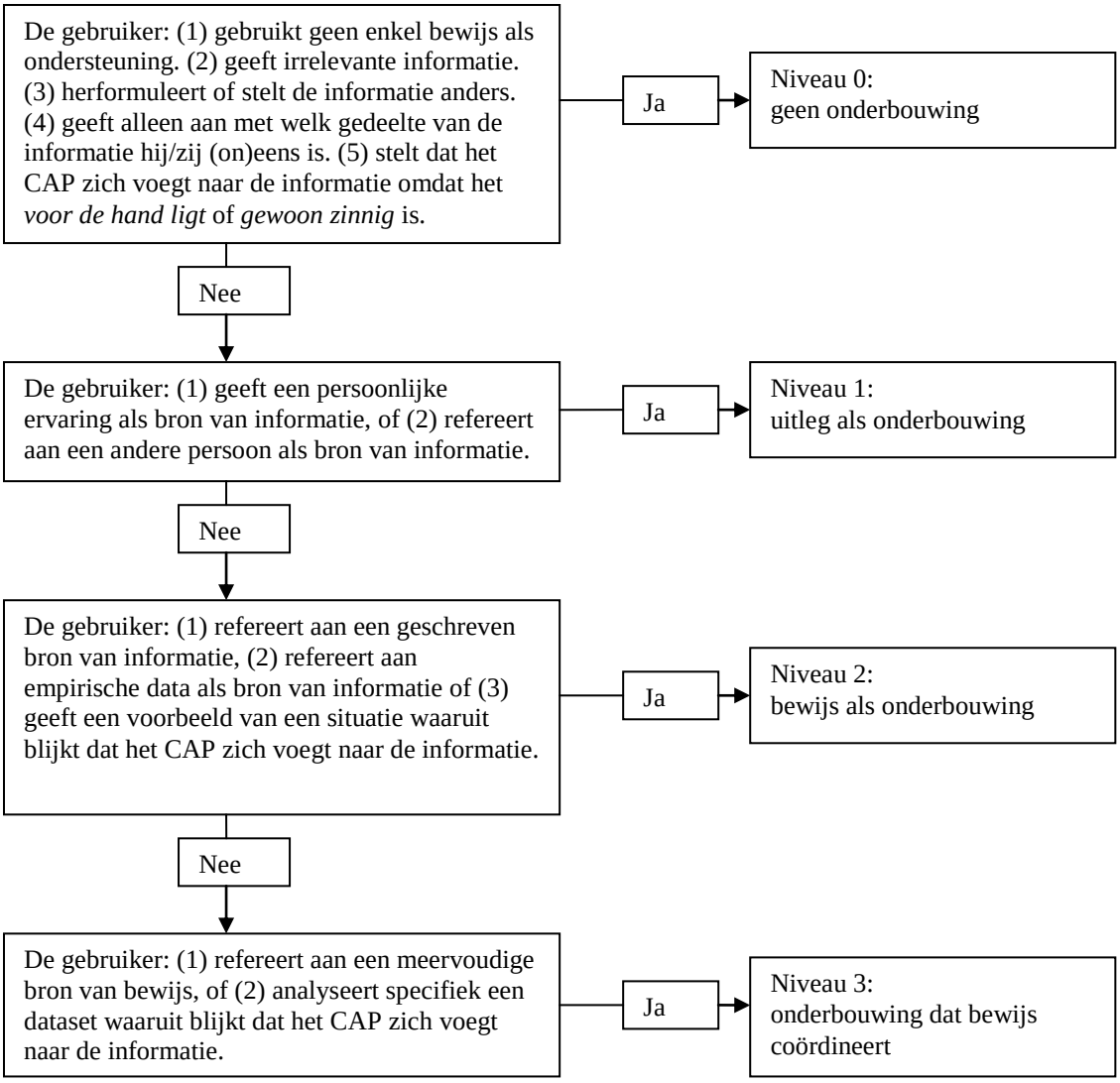
De scheiding van argumentatie naar conceptuele kwaliteit en epistemologische kwaliteit is in de context van het hier beschreven onderzoek niet te realiseren. Het is in dit onderzoek niet mogelijk om de conceptuele kwaliteit uit te werken door alle mogelijke argumenten in de evaluatie van het CAP te verzamelen en te voorzien van een classificatie. Dit omdat er in dit onderzoek geen sprake is van een gesloten wetenschappelijke discussie. Door het open karakter van het onderwerp, is het ondoenlijk om een gesloten classificatie te ontwerpen die op de zelfevaluatie van verschillende CAP's op verschillende onderwijsinstellingen van toepassing is. Daarbij zal een inhoudsexpert moeten bepalen wat de kwaliteit van het inhoudelijke argument is, wat kan leiden tot een arbitrair beoordelingsproces. Daarom wordt de kwaliteit van bewijsvoering hier uitgewerkt door modellen te gebruiken die inhoudelijk naar soorten bewijsvoering kijken en de classificatie daarvan beschrijven. In dit onderzoek wordt daarvoor de term *structurele kwaliteit* gehanteerd. Door vanuit soorten bewijs naar de inhoudelijke juistheid van een bewijsvoering te kijken, ontstaat een meer algemene inhoudelijke blik die bruikbaar is in de contexten van verschillende CAP's op verschillende onderwijsinstellingen. De

discussie omtrent de conceptuele kwaliteit en epistemologische kwaliteit geeft duidelijk het spanningsveld aan in de selectie en het gebruik van kwaliteitsmodellen uit de argumentatietheorie. In de volgende paragraaf wordt eerst het gebruikte model in eerder onderzoek beschreven en worden vervolgens de modellen uit de argumentatietheorie die bruikbaar zijn in dit onderzoek uiteengezet. Gezien de kritieken en de deels inhoudelijke focus in dit onderzoek worden de modellen van Toulmin (1958) en Erduran en collega's (2004) niet gebruikt.

3.3 Modellen voor beoordeling van de kwaliteit van bewijsvoering

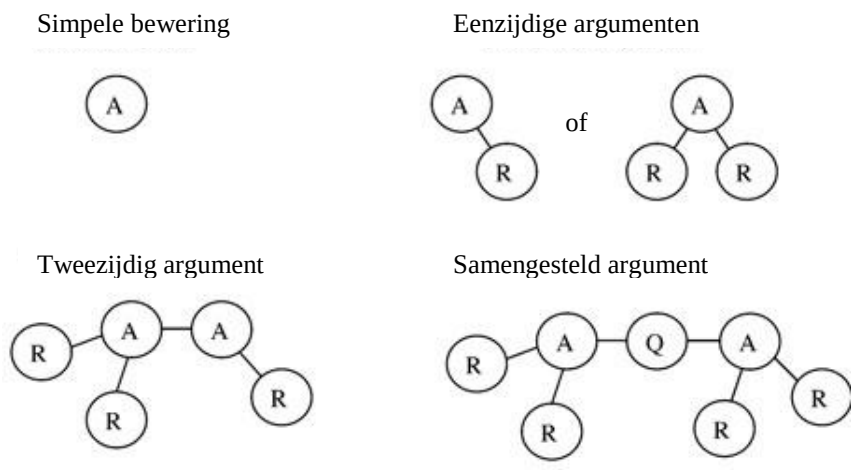
Het oorspronkelijke model van Clark en Sampson (2005) is in aangepaste vorm gebruikt door Baartman, en collega's (2007). In dat laatste onderzoek is het model van Clark en Sampson (2005) omgezet tot een stroomschema om de kwaliteit van bewijsvoering te kunnen kwalificeren in de context van CAP. Dit stroomschema is afgebeeld in figuur 1 op de volgende bladzijde.

Niveau 0 (geen onderbouwing) geeft aan dat de onderbouwing van het bewijs op geen enkele grond gebaseerd is. Met niveau 1 (uitleg als onderbouwing) wordt verwezen naar een persoon of een persoonlijke ervaring. Niveau 2 (bewijs als onderbouwing) wordt gebruikt om te verwijzen naar een geschreven bron van informatie, empirische data of een voorbeeld van een situatie. En als hoogste niveau geldt niveau 3 (onderbouwing dat bewijs coördineert) waarbij men moet denken aan meervoudige bronnen van bewijs of een specifieke analyse van een dataset.



Figuur 1: het stroomschema van Baartman en collega's (2007).

Een model dat hier verder op in gaat, is het model van Schwarz, Neuman, Gil en Ilya (2003). Ze definiëren een argument als een conclusie met minstens een bewering (A: bewering). De bewering kan worden ondersteund met (meervoudige) redeneringen, tegenargumenten en metaverklaringen (R: reden). Argumenten kunnen worden uitgebreid met kwalificaties zoals achtergronden en rechtvaardigingen (Q: kwalificatie). Dit model is in figuur 2 afgebeeld .



Figuur 2: het model van Schwarz en collega's (2003).

Het bereik van het model gaat van een simpele bewering tot een samengesteld argument. Een simpele beweringen bestaat uit een conclusie die niet wordt ondersteund door een rechtvaardiging. Eenzijdige argumenten omvatten alleen een conclusie en een of meer argumenten. Tweezijdige argumenten bevatten argumenten die de conclusie ondersteunen en uitdagen zonder duidelijk in te gaan op een analyse van voor- en tegenargumenten om een probleem op te lossen. Samengestelde argumenten maken gebruik van voor- en tegenargumenten door er een kwalificatie aan te geven.

Hornikx (2008) heeft op een andere wijze onderzoek gedaan naar de kwaliteit van bewijsvoering. Hornikx heeft onderzoek gedaan naar de verwachtingen van mensen over de overtuigingskracht van anekdotische, statistische, causale en expert-bewijzen en heeft deze verwachtingen vergeleken met de feitelijke overtuigingskracht van deze verschillende typen van bewijs. Anekdotisch bewijs is een geval of kwestie en bestaat doorgaans uit een anekdote om een bewering te ondersteunen. Statistisch bewijs wordt gezien als numerieke informatie over een groot aantal gevallen of kwesties. Causaal bewijs bevat een uitleg over waarom een bewering met het bewijs wordt ondersteund. Expertbewijs bestaat uit een bevestiging door een expert. In het onderzoek bleek dat mensen statistisch bewijs het meest overtuigende type bewijs vonden, gevolgd door expert, causaal en als laatste anekdotisch bewijs. In vergelijking met de feitelijke overtuigingskracht van deze verschillende typen van bewijs blijkt dat de verwachting van mensen over het algemeen accuraat is.

De modellen van Baartman en collega's (2007) en Schwarz en collega's (2003) lijken in eerste instantie niet bij de *structurele kwaliteit* thuis te horen, omdat deze de kwaliteit uitdrukken in de aan-

of afwezigheid van componenten van argumentatie. Een meer algemene inhoudelijke blik naar soorten bewijzen lijkt hierbij afwezig. De modellen kunnen echter gebruikt worden doordat ze inhoudelijk verwijzen naar verschillende soorten bewijs in een classificatieraamwerk (Baartman & collega's, 2007) of verwijzen naar het aantal en het soort argumenten dat een bewijs ondersteunt (Schwarz & collega's, 2003). De uitwerking van Schwarz en collega's vult het stroomschema van Baartman en collega's (2007) aan, omdat er wordt ingegaan op het expliciteren van voor- en tegenargumenten. In de context van dit onderzoek is te verwachten dat er voor- en tegenargumenten worden afgewogen om tot een oordeel over de kwaliteit te komen. Het onderzoek van Hornikx (2008) kan eveneens gekoppeld worden aan het stroomschema van Baartman en collega's (2007) en voegt daarmee een meer op inhoud gerichte classificatie toe. Anekdotisch bewijs en causaal bewijs kunnen dienen als extra uitwerking van niveau 1 van het stroomschema. De uitleg als onderbouwing, waarbij verwezen wordt naar een persoon of een persoonlijke ervaring, kan gezien worden als voorbeelden van anekdotisch bewijs. Causaal bewijs kan binnen de conceptuele kwaliteit dienen als een kwalitatief hogere vorm van bewijs binnen niveau 1. Expertbewijs kan binnen niveau 2 van het stroomschema van Baartman en Collega's dienen als een kwalitatief hogere vorm van bewijs. Statistisch bewijs kan dienen als een extra toevoeging binnen niveau 3 van het stroomschema. Ook is er overlap te vinden met het model van Schwarz en collega's (2003) waarbij anekdotisch, causaal en expert-bewijs kunnen worden gezien als inhoudelijke voorbeelden van eenzijdige argumenten. Statistisch bewijs kan daarbij gezien worden als een mogelijkheid om binnen samengestelde bewijzen een nog hoger niveau te onderscheiden, namelijk samengestelde bewijzen vergezeld van statistisch bewijs.

Wat opvalt in alle drie de modellen is dat meervoudige bewijsvoering en bewijsvoering uit verschillende bronnen van een hogere kwaliteit is. Dit wordt onderschreven door inzichten uit validiteitsvraagstukken. Met name Messick (1989) gaat hierbij in op wat constructvaliditeit genoemd wordt. Constructvaliditeit is een onderdeel van het validiteitvraagstuk dat ingaat op de adequaatheid van assessments in het meten van het onderliggende construct (of vaardigheid) welke met het assessment wordt bepaald. In de context van het huidige onderzoek is constructvaliditeit de mate waarin de bewijsvoering in de zelfevaluatie de werkelijke lading van de kwaliteitscriteria in de zelfevaluatie dekt. Messick noemt twee bedreigingen van de validiteit van een dergelijk assessment.

Dit zijn: construct irrelevante variantie en construct onder-representatie. Construct irrelevante variantie kan in het kader van dit onderzoek vertaald worden als de relevantie van bewijs. De bedreiging die daarbij ontstaat, is dat de bewijsvoering de werkelijke lading van het CAP (het onderliggende construct) niet dekt. Er is dan sprake van variatie in de bewijsvoering die is ontstaan door niet relevante denkbeelden in de bewijsvoering te betrekken. Construct onder-representatie kan in het kader van dit onderzoek vertaald worden als compleetheid van bewijs. De bedreiging die daarbij kan ontstaan is dat een bewijsvoering te nauw omschreven is. Dan is het onmogelijk om alle kritieke aspecten van het assessment in de bewijsvoering in beeld te brengen. Door meervoudige bewijsvoering en bewijsvoering uit verschillende bronnen te gebruiken, ontstaat bewijs dat meer relevant en compleet is dan bewijsvoering die enkelvoudig is of uit een enkele bron bestaat. De kans op de beide bedreigingen van constructvaliditeit wordt met dergelijke bewijsvoering verminderd. Vanuit het validiteitvraagstuk is hiermee te motiveren dat meervoudige bewijsvoering en bewijsvoering uit verschillende bronnen van een hogere kwaliteit is dan enkelvoudige bewijsvoering. Ook Moss, Girard en Haniford (2006) hebben vanuit het validiteitvraagstuk naar bewijsvoering gekeken. Zij geven daarbij aan dat bewijsvoering uit verschillende bronnen gecombineerd moet worden om van een goede kwaliteit te kunnen zijn. Vooral het model van Schwarz en collega's (2003) geeft het verschil in kwaliteit van deze vorm van bewijsvoering nadrukkelijk aan.

4. Methode

4.1 Data verzameling

In 2007 en 2008 zijn drie faculteiten van een hogeschool in Nederland ondersteund in hun evaluatieprocedure naar de kwaliteit van het CAP. De evaluatiegroep bestond steeds uit vijf tot zeven deelnemers waaronder docenten, een sectieleider, een lid van de toetsexamencommissie en een stagebegeleider. De evaluatieprocedure bestond uit een individuele evaluatie gevolgd door een groepsdiscussie. In de individuele evaluatie is aan de deelnemers gevraagd een webgebaseerde vragenlijst in te vullen. In deze vragenlijst werd door middel van een aantal indicatoren om een oordeel over het kwaliteitscriterium gevraagd. De deelnemers konden daartoe een score tussen nul en honderd geven en middels een tekstvak bewijsvoering voor hun score aandragen. De gegevens van de deelnemers uit de vragenlijst dienden als basis voor de groepsdiscussie. In deze discussie werd per kwaliteitscriterium door alle deelnemers getracht een oordeel over de kwaliteit van het CAP te geven. Daarbij konden de deelnemers met elkaar in discussie gaan om zo hun bewijsvoering te verduidelijken. Er heeft bij elke faculteit een voorbereidende bijeenkomst plaatsgevonden waarin de focus van het CAP werd bepaald. Op een faculteit is tijdens de voorbereidende bijeenkomst naast het vaststellen van het CAP ook ingegaan op wat kwalitatief goede bewijsvoering is. De deelnemers hebben daarvoor een training gekregen over het verzamelen van bewijzen en argumenten bij zelfevaluatie. Ze werden getraind in het formuleren van onderbouwing op basis van een eigen mening, positief of negatief bewijs, een voorbeeldsituatie of verwijzingen naar geschreven informatie of empirische data. Daarbij werd aangegeven dat de deelnemers in het achterhoofd moesten houden of het een eigen mening, een gedeelde mening of tegengestelde meningen betrof. De discussiegroep omvatte anders dan op de andere faculteit ook een student. Van alle drie de faculteiten zijn de gegevens uit de vragenlijsten en de uitgetypte groepsdiscussies gebruikt voor analyse.

4.2. Codeerschema

Het codeerschema dat in dit onderzoek wordt gebruikt is afgebeeld in tabel 1. De *structurele kwaliteit van bewijsvoering* wordt vormgegeven door de theorieën van Hornikx (2008), Schwarz en collega's

(2003) en het stroomschema van Baartman en collega's (2007) aan elkaar te relateren. In de kolom *code voor structurele bewijsvoering* wordt de uiteindelijke hiërarchie voor structurele kwaliteit weergegeven met de code 0 voor de laagste structurele kwaliteit van bewijsvoering en code 7 voor de hoogste structurele kwaliteit van bewijsvoering.

Tabel 1: Samenvoeging van modellen tot een codeerschema voor structurele kwaliteit van bewijsvoering.

<table>
<tr><th colspan="4">Structurele kwaliteit van bewijsvoering</th></tr>
<tr><td>Hornikx (2008)</td><td>Baartman et al. (2007)</td><td>Schwarz et al. (2003)</td><td>Code voor structurele bewijsvoering:</td></tr>
<tr><td></td><td>Niveau 0:
Geen onder-bouwing</td><td>Simpele bewering</td><td>0: Geen bewijs</td></tr>
<tr><td>1. Anekdotisch bewijs</td><td rowspan="2">Niveau 1:
Uitleg als onderbouwing</td><td rowspan="4">Eenzijdige argumenten</td><td>1PA: Persoonlijk anekdotisch
1AP: Anekdotisch andere persoon</td></tr>
<tr><td>2 Causaal bewijs</td><td>2: Causaal zonder specifiek bewijs</td></tr>
<tr><td rowspan="2"></td><td rowspan="2">Niveau 2:
Bewijs als onder-bouwing</td><td>3GI: Geschreven informatie
3ED: Empirische data
3VS: Voorbeeld-situatie</td></tr>
<tr><td>3. Expertbewijs</td><td>4: Expertbewijs</td></tr>
<tr><td></td><td rowspan="3">Niveau 3:
Onderbouwing dat bewijs coördineert</td><td>Tweezijdige argumenten zonder voor of tegen afwegen</td><td>5: Tweezijdig zonder voor of tegen afwegen</td></tr>
<tr><td></td><td rowspan="2">Samengestelde argumenten met voor of tegen afwegen</td><td>6: Samengesteld met voor of tegen afwegen</td></tr>
<tr><td>4. Statistisch bewijs</td><td>7: Samengesteld met statistisch bewijs</td></tr>
</table>

Code 0 bestaat uit geen onderbouwing voor bewijs zoals een simpele bewering, irrelevante informatie, herformuleringen, aangeven dat men het niet eens is of aangeven dat iets gewoon zo is. Dit is ontleend aan Baartman en collega's (2007) en Schwarz en collega's (2003). Code 1 bestaat uit de

samenvoeging van anekdotisch bewijs van Hornkix (2008) met uitleg als onderbouwing van Baartman en collega's (2007), bestaande uit persoonlijk bewijs en refereren aan een andere persoon als bewijs. Anekdotisch bewijs, waarbij wordt verwezen naar een geval of kwestie om een bewering te ondersteunen, kan zowel uit een persoonlijke ervaring (1PA) alsmede de ervaring van een andere persoon (1AP) bestaan. Beide vormen bestaan uit eenzijdige argumentatie zoals is aangegeven door Schwarz en collega's (2003). Code 2 bestaat uit eenzijdig causaal bewijs. Causaal bewijs is een uitleg waarom een bewering met bewijs wordt ondersteund. Dit is ontleend aan Hornikx (2008) en bestaat uit eenzijdige argumentatie zoals door Schwarz en collega's (2003) is aangegeven. De code 3 wordt gevormd door de drie vormen van bewijs als onderbouwing van Baartman en collega's (2007) te gebruiken. Dit is bewijs op basis van geschreven informatie (3GI), empirische data (3ED) of een voorbeeldsituatie (3VS). Deze drie vormen vallen ook onder het causaal bewijs van Hornikx (2008). De code 4 geldt als extra toevoeging voor bewijs als onderbouwing en is het expertbewijs van Hornikx. Ook deze vormen van bewijsvoering in de code 3 en 4 bestaan uit eenzijdige argumentatie zoals aangegeven door Schwarz en collega's (2003). Code 5 bestaat uit tweezijdige argumenten zonder een afweging van voor- of tegenargumenten en is ontleend aan Schwarz en collega's. Code 6 en 7 bestaan uit samengestelde argumenten met een afweging van voor- of tegenargumenten welke eveneens zijn ontleend aan Schwarz en collega's. In code 7 moet ook aan statistisch bewijs worden gerefereerd zoals aangegeven door Hornikx (2008). Onderbouwing dat bewijs coördineert van Baartman en collega's (2007) is ook terug te vinden onder de codes 5, 6 en 7. Door de theorieën van Schwarz en collega's (2003) en Hornikx (2008) toe te passen, wordt de onderbouwing dat bewijs coördineert nader gespecificeerd en van een verschil in classificatie voorzien.

4.3 Codeerproces

Elke uiting van een deelnemer werd tijdens het codeerproces gecontroleerd door de onderzoeker op de aanwezigheid van een bewijsvoering. Uitingen die niet op de structurele kwaliteit van bewijsvoering ingingen, zoals aangegeven in tabel 1, werden in de analyse niet verder meegenomen. Omdat de deelnemers aan de zelfevaluatie een oordeel over de kwaliteit van een kwaliteitscriterium van het CAP moesten geven, werd voor de structurele kwaliteit gekeken naar het geheel van bewijsvoeringen

binnen een kwaliteitscriterium. Er werd in de uitingen van de groepsdiscussie geen samengestelde bewijsvoering gevonden binnen één enkele uiting van een deelnemer. Dit omdat de samengestelde bewijsvoeringen in de groepsdiscussie ontstonden door discussie en argumentatie. Om dit proces van ontstaan van samengestelde bewijzen te kunnen herleiden tijdens de analyse, werd een gevonden bewijsvoering naast de code voor structurele bewijsvoering ook voorzien van een *identificatiecode*. Deze code kon in een later stadium in verband worden gebracht met een gelijke, nieuwe of samengestelde bewijsvoering. Daarbij werd de *identificatiecode* ook gebruikt om in een later stadium het kwaliteitscriterium terug te kunnen vinden. In zowel de webgebaseerde vragenlijst als de groepsdiscussie werden bij de indicatoren van deze twaalf kwaliteitscriteria bewijzen aangedragen om een oordeel over de kwaliteit van het CAP te geven. De *identificatiecode* start daarom met een nummer van het kwaliteitscriterium gevolgd door een letter om het bewijs te identificeren. In tabel 2 is een voorbeeld gegeven van het codeerproces van een aantal gevonden bewijsvoeringen bij een van de faculteiten.

Tabel 2: Voorbeeld van het codeerproces van de gevonden bewijsvoering van een van de faculteiten.

wie	**Fragment nummer**	**Code voor Structurele kwaliteit**	**Identificatie code**	**Fragment**
1. Geschiktheid voor onderwijsdoelen uit interview				
LF	10	1PA	1d	*1.2 in het CAP wordt voldoende aandacht besteed aan kennis* niet want ik vind onder al die competenties zit niet gewoon kennis dus kun je daar niet aan voldoen
IB	12	3VS	1e	*1.2 in het CAP wordt voldoende aandacht besteed aan kennis* niet want dat als je kijkt naar andere opleidingen of een middelbare school dan is van te voren je gaat dit en dit leren en dus bepaalde kennisbasis die je opbouwt terwijl hier is er aandacht voor kennis maar is er niet bepaald welke kennis dat dan is omdat dat ja, via competenties toch op meerdere manieren in te vullen is
AG	55	5	1s	*1.2 in het CAP wordt voldoende aandacht besteed aan kennis* wel want je kunt niet de volledige kennisbasis toetsen en je kunt diep op een kennisbasis ingaan of minder diep op meerdere kennisbasissen ingaan. **Daarmee is het tegenargument van 1q erbij gekomen. Men is het daar allemaal over eens. Dit bewijs is ontstaan uit 1n (niet volledige kennisbasis toetsen) en 1q (diep op een kennisbasis of minder diep op meerdere kennisbasissen ingaan).**
1. Geschiktheid voor onderwijsdoelen uit webgebaseerde vragenlijst				
JB	--	3VS	1x	*1.2 in het CAP wordt voldoende aandacht besteed aan kennis* wel want er is een volledig competentie-examen met nadruk op kennis-, vaardigheden- en attitude-aspecten in toetsing.

Voorafgaand aan de letterlijke uiting van de deelnemer wordt onder *fragment* het onderwerp dat ter discussie staat cursief aangegeven, teneinde in een later stadium van het coderen eenvoudig naar het kwaliteitscriterium te kunnen herleiden. Onder de *code voor structurele kwaliteit* werd de code uit het codeerschema genoteerd. Onder de *identificatiecode* werd het eigenlijke bewijsstuk genoteerd gekoppeld aan het kwaliteitscriterium. De uiting in fragment nummer 55 dient als voorbeeld van een samengesteld bewijs dat was ontstaan door de voorafgaande bewijsvoeringen met elkaar te relateren. Daarbij was het voor de onderzoeker mogelijk om in het fragment aantekeningen te maken om zo aan te geven uit welke eerder gevonden bewijsvoeringen het bewijs werd samengesteld. Deze aantekeningen zijn in het voorbeeld van tabel 2 vet afgedrukt. Mochten de bewijsvoeringen uit de webgebaseerde vragenlijst niet in de groepsdiscussie aan bod zijn gekomen, dan werden deze bewijsvoeringen los gecodeerd (zie de onderste rij van tabel 2). Om de codes aan de uitingen in het interview toe te kunnen voegen werd gebruik gemaakt van het programma MEPA. Uitingen die niet op bewijsvoeringen uit het codeerschema ingingen werden aangegeven met een streepje (-). Daarnaast werd separaat in het programma Excel per faculteit de tabel van gevonden bewijsvoeringen bijgehouden, naar het voorbeeld in tabel 2.

4.4 Betrouwbaarheid

Om de betrouwbaarheid van het codeerschema te bepalen hebben twee onafhankelijke onderzoekers elk afzonderlijk 200 van totaal 4262 uitingen gecodeerd. Daarbij zijn in het fragment van 200 uitingen door elke onderzoeker dertien bewijsvoeringen aangetroffen. Zeven van de dertien bewijsvoeringen werden daarbij hetzelfde opgevat én op dezelfde uiting van de spreker geselecteerd. De overige zes bewijsvoeringen bleken niet op dezelfde uiting van de spreker te zijn geselecteerd en wisselden van code voor structurele bewijsvoering. Daarbij is een interbeoordelaarsovereenkomst (Cohen’s kappa) gevonden van 0.58. Deze waarde staat voor een gemiddelde interbeoordelaarsovereenkomst. De meeste overeenkomst werd daarbij gevonden op de uitingen waar geen bewijsvoering werd aangetroffen. Deze uitingen werden zoals eerder aangegeven gecodeerd met het streepje (-) en bleken de meeste invloed op de interbeoordelaarsovereenkomst uit te oefenen. Over de overige zes uitingen werd door de onderzoekers gediscussieerd, waarbij bleek dat de codes 1AP, 1PA en 3VS verwarring

opleverden. Over de betekenis van deze codes werd nogmaals onderhandeld, waarna een tweede codeersessie van 200 uitingen volgde. Naar aanleiding van deze tweede codeersessie bleek er minder verwarring over de codes. In deze sessie werden 17 bewijzen gevonden, waarvan er tien op dezelfde uiting van een spreker én met dezelfde code voor structurele kwaliteit werden beoordeeld. De overige zeven codes bleken na een inhoudelijke discussie dezelfde bewijsvoeringen te betreffen, maar op een andere uiting in de data te zijn geselecteerd. Dit bleek te liggen aan de context waarbinnen de onderzoeker de bewijsvoering selecteerde. Hetzelfde soort bewijs werd dan door een onderzoeker geselecteerd, maar werd aan een fragment ervoor of erna gekoppeld. Dat bleek mogelijk omdat in meerdere fragmenten over hetzelfde bewijs werd gesproken. Zo kwam het voor dat een spreker zijn bewijsvoering introduceerde, werd geïnterrumpeerd door een andere spreker om daarna zijn bewijsvoering alsnog af te ronden. De onderzoeker had dan twee fragmenten waaraan de bewijsvoering gekoppeld kon worden. Het lag dan aan de beoordeling van de onderzoeker welk fragment daarvoor geselecteerd werd. Het fenomeen van inhoudelijk gelijke beoordeling op verschillende fragmenten verlaagde de interbeoordelaarsovereenkomst. Tegelijkertijd werd deze verhoogd door het grote percentage uitingen waar geen bewijsvoering werd aangetroffen. Deze twee effecten maakten dat de interbeoordelaarsovereenkomst inhoudelijk niet op het codeerproces aansloot. Dat heeft ertoe geleid dat er geen interbeoordelaarsovereenkomst voor de tweede codeersessie werd berekend. Uit de discussie onder de onderzoekers bleek wel overeenstemming over de manier waarop bewijsvoeringen gecodeerd moesten worden. Door ook de bewijsvoeringen uit de webgebaseerde vragenlijst in het codeerproces op te nemen, werd daarbij gegarandeerd dat bewijsvoeringen die in de groepdiscussie over het hoofd werden gezien of niet aan bod kwamen, ook gecodeerd werden (zie ook de onderste regel van tabel 2).

4.5 Analyse

Voor de beantwoording van de eerste onderzoeksvraag: *Hoe ziet de kwaliteit van bewijsvoering er op de faculteiten uit?* worden de frequenties van de codes voor structurele kwaliteit inzichtelijk gemaakt. Daarmee kunnen uitspraken worden gedaan over het soort bewijs en het voorkomen van het soort bewijs op de onderzochte onderwijsinstellingen. Hiertoe worden alle gevonden bewijsvoeringen, zoals

in tabel 2 aangegeven, per onderwijsinstelling in het programma SPSS versie 12 voor Windows geladen. Vervolgens worden ter beantwoording van de tweede onderzoeksvraag: *Is er een verschil in de mate van kwaliteit van bewijsvoering bij de verschillende faculteiten?* de frequentieverdelingen van de structurele kwaliteit van bewijsvoering per faculteit vergeleken. Daartoe wordt per faculteit een rangscore voor de kwaliteit van bewijsvoering berekend, waarna met een Kruskal-Wallis test verschillen worden onderzocht. De derde onderzoeksvraag: *Is er een verschil in kwaliteit van bewijsvoering per kwaliteitscriterium van het CAP?* wordt beantwoord door voor alle faculteiten een rangscore voor de structurele kwaliteit per criterium te genereren en deze onderling te vergelijken. Vervolgens worden deze rangscores ook per onderwijsinstelling vergeleken om de verschillen per faculteit in kaart te brengen. Dat levert meer inzicht in de criteria die verschillen in kwaliteit van faculteit binnen de twaalf criteria veroorzaken. Deze verschillen worden met behulp van de Kruskal-Wallis test onderzocht. In onderzoeksvraag vier wordt het verschil onderzocht tussen de kwaliteit van bewijsvoering in discussies waarbinnen deelnemers, zoals eerder aangegeven, *veel consensus, in enige mate consensus* en *geen consensus* laten zien. In elk gesprek over een kwaliteitscriterium is het mogelijk om op basis van de webgebaseerde vragenlijst het gesprek in te delen in deze typen. Doordat de deelnemers een score tussen nul en honderd opgeven voor de kwaliteit van een indicator binnen het kwaliteitscriterium, is het mogelijk een frequentieverdeling van de scores per kwaliteitscriterium op te stellen. De vorm van de frequentieverdeling kan staan voor de mate waarin de deelnemers van te voren *veel consensus, in enige mate consensus* en *geen consensus* laten zien. In *Discovering statistics using spss* van Field (2005) wordt uitgelegd hoe men met behulp van de kurtosis en de standaardfout van de kurtosis de vorm van een frequentieverdeling kan omzetten tot een z-score. Daarbij is het mogelijk om, met behulp van betrouwbaarheidsintervallen (95%), vast te stellen wat de kans is dat de deelnemers van te voren *veel consensus* (z-score groter dan 1.96), *in enige mate consensus* (z-score tussen 1.96 en -1,96) en *geen consensus* (z-score kleiner dan -1.96) laten zien. Van de gevonden bewijsvoeringen wordt de frequentieverdeling van de structurele kwaliteit berekend en per type gesprek weergegeven. Vervolgens wordt op basis van de rangscore voor de structurele kwaliteit per kwaliteitscriterium, middels een Mann-Whitney U test, gekeken naar verschillen tussen de type gesprekken.

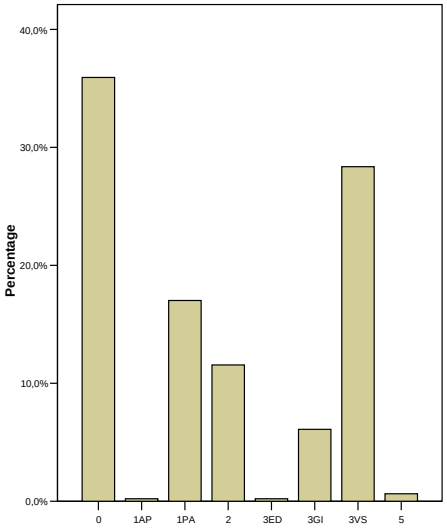
5. Resultaten

Op de volgende pagina is een overzicht gegeven van het soort bewijs en het voorkomen van het bewijs op de verschillende faculteiten. Op faculteit een (zonder training) zijn in totaal 476 verschillende bewijsvoeringen aangetroffen in de webgebaseerde vragenlijst en het groepsinterview. Op deze onderwijsinstelling is de bewijsvoering *geen bewijs* (35,9%) het meest aangetroffen gevolgd door de bewijsvoeringen *voorbeeld situatie* (28,4%) en de *persoonlijke anekdote* (17,0%). In tabel 3 en figuur 3 is een overzicht gegeven van het soort bewijs en het voorkomen van het soort bewijs van faculteit een. Op faculteit twee (met training) zijn in totaal 455 verschillende bewijsvoeringen aangetroffen. Op deze onderwijsinstelling is de bewijsvoering *voorbeeld situatie* (45,1%) het meest aangetroffen gevolgd door de bewijsvoeringen *causaal bewijs* (19,6%) en *geen bewijs* (18,0%). In tabel 4 en figuur 4 is een overzicht gegeven van het soort bewijs en het voorkomen van het soort bewijs van faculteit twee. Op faculteit drie (zonder training) zijn in totaal 374 verschillende bewijsvoeringen aangetroffen. Op deze onderwijsinstelling is de bewijsvoering *geen bewijs* (41,4%) het meest aangetroffen gevolgd door de bewijsvoeringen *persoonlijke anekdote* (24,1%) en *causaal bewijs* (19,0 %). In tabel 5 en figuur 5 is een overzicht gegeven van het soort bewijs en het voorkomen van het soort bewijs van faculteit drie.

Ten einde de drie onderwijsinstellingen onderling te kunnen vergelijken, zijn de verschillende bewijsvoeringen die hetzelfde niveau van bewijsvoering voorstellen samengevoegd. Dit is nodig omdat niet elke soort van bewijsvoering op de onderwijsinstellingen terug te vinden is terwijl deze voor hetzelfde niveau van kwaliteit staan. Dit geldt voor de bewijsvoeringen 1PA (persoonlijk anekdotisch) en 1AP (anekdotisch andere persoon) welke zijn samengevoegd tot één soort bewijsvoering: het anekdotisch bewijs. En voor de bewijsvoeringen 3GI (geschreven informatie), 3ED (empirische data) en 3VS (voorbeeld situatie) welke zijn samengevoegd tot de bewijsvoering: causaal met bewijs. Deze samenvoeging laat per faculteit een frequentieverdeling van de structurele kwaliteit zien welke is afgebeeld in tabel 6.

Tabel 3: Aantallen en soort bewijs op faculteit een.

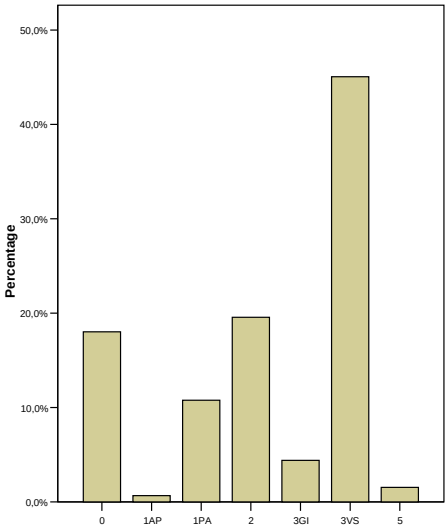
Soort bewijs	Aantal	Percentage
0	171	35,9 %
1AP	1	0,2 %
1PA	81	17,0 %
2	55	11,6 %
3ED	1	0,2 %
3GI	29	6,1 %
3VS	135	28,4 %
5	3	0,6 %
Totaal	476	100,0 %



Figuur 3: Soorten bewijs faculteit een.

Tabel 4: Aantallen en soort bewijs op faculteit twee.

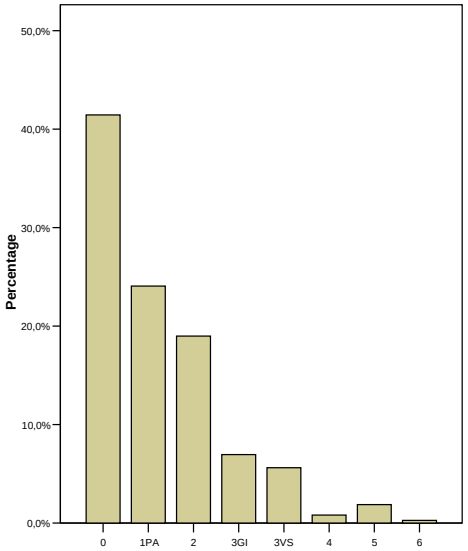
Soort bewijs	Aantal	Percentage
0	82	18,0 %
1AP	3	0,7 %
1PA	49	10,8 %
2	89	19,6 %
3GI	20	4,4 %
3VS	205	45,0 %
5	7	1,5 %
Totaal	455	100,0 %



Figuur 4: Soorten bewijs faculteit twee.

Tabel 5: aantallen en soort bewijs op faculteit drie.

Soort bewijs	Aantal	Percentage
0	155	41,3 %
1PA	90	24,1 %
2	71	19,0 %
3GI	26	7,0 %
3VS	21	5,6 %
4	3	0,8 %
5	7	1,9 %
6	1	0,3 %
Totaal	374	100,0 %



Figuur 5: Soorten bewijs faculteit drie.

Tabel 6: Frequentieverdeling van de structurele kwaliteit van bewijsvoering per faculteit.

Kwaliteit van bewijzen		faculteit 1	faculteit 2	faculteit 3
0	geen bewijs	35,9 %	18,0 %	41,3 %
1	anekdotisch bewijs (PA & AP)	17,2 %	11,5 %	24,1 %
2	causaal zonder specifiek bewijs	11,6 %	19,6 %	19,0 %
3	causaal met bewijs (ED, GI & VS)	34,7 %	49,4 %	12,6 %
4	expert bewijs	0 %	0 %	0,8 %
5	tweezijdig zonder afwegen	0,6 %	1,5 %	1,9 %
6	samengesteld met afwegen	0 %	0 %	0,3 %

Middels een Kruskal-Wallis test is het verschil in structurele kwaliteit van bewijsvoering op de drie onderwijsinstellingen onderzocht. Uit de analyse bleek een significant verschil tussen de drie onderwijsinstellingen (χ^2 = 114,66; df = 2; p < .001). In de rapportage van de resultaten wordt waar mogelijk een frequentieverdeling gegeven en worden bij de Kruskal-Wallis en de Mann-Whitney U testen de rangscores inzichtelijk gemaakt. Andere wijzen van rapporteren (bijvoorbeeld middels weergave van de mediaan of modus) bleken onvoldoende variatie weer te geven om uitslagen inzichtelijk te kunnen interpreteren. In tabel 7 zijn daarom de rangscores als basis voor de bovenstaande Kruskal-Wallis test gegeven voor de drie onderwijsinstellingen waarbij de scores voor de structurele kwaliteit zijn omgezet naar rangscores.

Tabel 7: Rangscores als basis voor de Kruskal-Wallis test voor de structurele kwaliteit van bewijsvoering op de drie faculteiten.

Onderwijsinstelling	Aantal bewijzen	Rangscore
faculteit 1	476	626,32
faculteit 2	455	788,17
faculteit 3	374	522,51

Naar aanleiding van Kruskal-Wallis test is het verschil tussen de faculteiten onderling bepaald door middel van de Mann-Whitney U test. Het verschil tussen faculteit een en twee is significant (U=82396,5; p < .001). Het verschil tussen faculteit een en drie blijkt eveneens significant (U=75818; p < .001). Evenals het verschil tussen faculteit twee en drie (U=49476,5; p < .001). Waarbij op faculteit twee de hoogste kwaliteit van bewijsvoering is gevonden, gevolgd door faculteit een en faculteit drie.

In tabel 8 is een overzicht gegeven van de rangscores voor de structurele kwaliteit van bewijsvoering per kwaliteitscriterium. Deze scores gelden voor alle drie de onderwijsinstellingen en zijn van hoog naar laag gerangschikt. Er is gekozen voor een rapportage naar rangscores omdat een overzicht van frequentieverdelingen van de structurele kwaliteit per kwaliteitscriterium leidt tot een onoverzichtelijke hoeveelheid grafieken en tabellen voor alle drie de onderwijsinstellingen.

Tabel 8: De rangscores voor de structurele kwaliteit van bewijsvoering per kwaliteitscriterium voor alle drie de onderwijsinstellingen.

Kwaliteitscriterium	Aantal bewijzen	Rangscore
authenticiteit	110	725,96
doel	114	725,92
zelfbeoordeling	98	703,86
herhaalbaarheid	106	702,06
tijd en kosten	113	677,69
vergelijkbaarheid	124	662,88
cognitief complex	78	634,91
onderwijs gevolgen	125	626,90
redelijkheid	139	603,25
aanvaardbaarheid	82	602,73
transparantie	97	597,21
betekenisvol	119	573,89

Middels de Kruskal-Wallis test is het verschil tussen de rangscores over de verschillende kwaliteitscriteria onderzocht. Uit de analyse bleek een significant verschil tussen de verschillende kwaliteitscriteria (χ^2 = 26,64; df = 11; p = .005). In tabel 9, 10 en 11 zijn de rangscores voor de structurele kwaliteit van bewijsvoering per kwaliteitscriterium gegeven van respectievelijk faculteit een, twee en drie. Deze rangscores zijn van hoog naar laag gerangschikt.

Tabel 9: faculteit een.

Kwaliteitscriterium	Aantal bewijzen	Rangscore
zelfbeoordeling	33	290,83
authenticiteit	42	288,63
doel	51	285,80
betekenisvol	39	272,38
onderwijs gevolgen	40	251,61
tijd en kosten	36	241,94
cognitief complex	33	226,64
redelijkheid	63	217,82
vergelijkbaarheid	49	199,54
aanvaardbaarheid	29	192,28
herhaalbaarheid	29	188,36
transparantie	32	181,72

Tabel 10: faculteit twee.

Kwaliteitscriterium	Aantal bewijzen	Rangscore
herhaalbaarheid	42	260,04
zelfbeoordeling	33	256,83
authenticiteit	36	251,39
doel	26	248,52
vergelijkbaarheid	42	248,31
transparantie	37	222,46
aanvaardbaarheid	34	221,82
tijd en kosten	52	221,19
redelijkheid	46	214,50
cognitief complex	29	212,41
onderwijs gevolgen	44	203,69
betekenisvol	34	180,53

Tabel 11: faculteit drie.

Kwaliteitscriterium	Aantal bewijzen	Rangscore
herhaalbaarheid	35	236,04
vergelijkbaarheid	33	235,59
doel	37	220,69
tijd en kosten	25	194,04
transparantie	28	190,29
authenticiteit	32	183,66
aanvaardbaarheid	19	179,84
cognitief complex	16	177,84
onderwijs gevolgen	41	174,01
redelijkheid	30	162,97
zelfbeoordeling	32	158,53
betekenisvol	46	141,49

Middels de Kruskal-Wallis test is voor alle drie de onderwijsinstellingen het verschil tussen de rangscores over de verschillende kwaliteitscriteria onderzocht. Voor faculteit een bleek een significant verschil tussen de verschillende kwaliteitscriteria (χ^2 = 41,22; df = 11; p < .001). Voor faculteit twee bleek geen significant verschil (χ^2 = 16,15; df = 11; p = .136). Voor faculteit drie bleek een significant verschil tussen de verschillende kwaliteitscriteria (χ^2 = 33,40; df = 11; p < .001).

Wat betreft de laatste onderzoeksvraag is in tabel 12 het overzicht gegeven van de verschillende type gesprekken waarbij de deelnemers in de webgebaseerde vragenlijst voorafgaand aan het groepsinterview *veel consensus, in enige mate consensus* en *geen consensus* laten zien. Voor alle drie de onderwijsinstellingen is per criterium, middels de z-score voor de kurtosis, bepaald in welke categorie het type gesprek te plaatsen is.

Tabel 12: Verschillende type gesprekken voorafgaand aan het groepsinterview.

Kwaliteitscriterium	faculteit een		faculteit twee		faculteit drie	
	Z-score kurtosis	Type gesprek	Z-score kurtosis	Type gesprek	Z-score kurtosis	Type gesprek
doel	.94	Enige mate	2,53	Veel	-.96	Enige mate
vergelijkbaarheid	6.35	Veel	2,80	Veel	1.58	Enige mate
herhaalbaarheid	.28	Enige mate	-1,04	Enige mate	-.26	Enige mate
aanvaardbaarheid	-.06	Enige mate	1,22	Enige mate	.19	Enige mate
transparantie	1.93	Enige mate	0,24	Enige mate	.12	Enige mate
redelijkheid	2.41	Veel	3,30	Veel	.34	Enige mate
authenticiteit	2,61	Veel	2,82	Veel	.81	Enige mate
cognitief complex	-0,93	Enige mate	-0,29	Enige mate	5.9	Veel
betekenisvol	-0,97	Enige mate	3,10	Veel	-.58	Enige mate
zelfbeoordeling	-0,00	Enige mate	5,48	Veel	2.56	Veel
tijd en kosten	-1,01	Enige mate	-0,02	Enige mate	.05	Enige mate
onderwijs gevolgen	0,50	Enige mate	2,77	Veel	.08	Enige mate

Op basis van de z-score voor de kurtosis, is op basis van de webgebaseerde vragenlijst geen gesprek gevonden dat te plaatsen is onder het type *geen consensus*. Opvallend is dat het type gesprek waarin van te voren *veel consensus* is te onderscheiden, het meeste voorkomt op faculteit twee. In tabel 13 zijn de frequentieverdelingen van de structurele kwaliteit van bewijsvoering van de alle drie de onderwijsinstellingen verdeeld naar het type gesprek. Uit de Mann-Whitney U test blijkt geen significant verschil te bestaan voor de structurele kwaliteit tussen het type gesprek (U=188560; p = .06). In tabel 14 zijn de rangscores als basis voor de Mann-Whitney U test gegeven.

Tabel 13: De frequentieverdelingen van de structurele kwaliteit van bewijsvoering van de alle drie de onderwijsinstellingen verdeeld naar het type gesprek.

kwaliteit van bewijzen	Type gesprek	
	veel consensus	enige mate consensus
0 geen bewijs	30,3 %	31,9 %
1 anekdotisch bewijs (PA & AP)	14,7 %	18,6 %
2 causaal zonder specifiek bewijs	16,0 %	16,8 %
3 causaal met bewijs (ED, GI & VS)	37,0 %	31,4 %
4 expert bewijs	0,4 %	0,1 %
5 tweezijdig zonder afwegen	1,6 %	1,1 %
6 samengesteld met afwegen	0 %	0,1 %

Tabel 14: Rangscores voor de Mann-Whitney U test voor het type gesprek en de structurele kwaliteit.

Type gesprek	Aantal bewijzen	Rangscore
Veel consensus	495	667,07
Enige mate consensus	810	638,29

6. Discussie en conclusie

Het doel van dit onderzoek is het bepalen van de kwaliteit van bewijsvoering bij de zelfevaluatie van drie faculteiten van een hogeschool. Daarbij worden verschillen onderzocht in de kwaliteit van bewijsvoering tussen de kwaliteitscriteria van het CAP en gesprekken waarin de deelnemers *veel consensus, in enige mate consensus* en *geen consensus* laten zien. Ook wordt het effect op de kwaliteit van bewijsvoering van een training onderzocht, die ingaat op het verzamelen van bewijzen en argumenten in een bijeenkomst voorafgaand aan de zelfevaluatie. Daar waar in de conclusies en aanbevelingen wordt gesproken over bewijsvoering, wordt de bewijsvoering bedoeld zoals aangegeven in het codeerschema (tabel 1).

Voor de beantwoording van onderzoeksvraag één: *Hoe ziet de kwaliteit van bewijsvoering er op de drie faculteiten uit?* zijn de frequentieverdelingen van de kwaliteit van bewijsvoering per faculteit vergeleken. Het expertbewijs en de tweezijdige of samengestelde bewijzen komen op de onderwijsinstellingen niet of nauwelijks voor. Het samengestelde bewijs met een verwijzing naar statistisch bewijs is op geen enkele onderwijsinstelling aangetroffen. Het anekdotisch bewijs waarbij wordt verwezen naar een geval of kwestie, wordt het meeste aangetroffen in de vorm van de persoonlijke anekdote. Dit komt overeen met de conclusie van Baartman en collega's (2007) uit het eerdere onderzoek waarin is geconstateerd dat de deelnemers hun individuele bewijs over de kwaliteit van het CAP voornamelijk op persoonlijke ervaringen baseren. Dit onderzoek laat echter zien dat de deelnemers van faculteit één en twee zich in grotere mate baseren op de voorbeeldsituatie. De voorbeeldsituatie komt namelijk het meeste voor binnen de bewijzen geschreven informatie (3GI), empirische data (3ED) en de voorbeeldsituatie (3VS). Dit is voor faculteit twee te verklaren door de training waarin ook de voorbeeldsituatie is besproken. Voor faculteit één kan dit verklaard worden door het codeerschema. Er is gekeken vanuit de verschillende bewijsvoeringen, in plaats van de aan- of afwezigheid van bronnen voor de bewijsvoering zoals in het onderzoek van Baartman en collega's. Dat kan leiden tot een andere, meer inhoudelijke blik, bij het beoordelen van de kwaliteit van bewijsvoering. Het grootste aandeel van de bewijsvoering is op faculteit één en drie de bewijsvoering geen bewijs (code 0). Dit is te verklaren doordat faculteit één en drie geen training hebben gekregen en

lijken, op het eerste gezicht, daardoor met een lagere kwaliteit van bewijsvoering de zelfevaluatieprocedure te hebben doorlopen.

Bij onderzoeksvraag twee: *Is er een verschil in de mate van kwaliteit van bewijsvoering bij de verschillende faculteiten?* kan worden geconcludeerd dat op faculteit twee kwalitatief betere bewijsvoering wordt aangedragen. Op deze faculteit is het minste aantal keer *geen bewijs* geleverd en bestaat het kwalitatief betere bewijs voornamelijk uit meer voorbeeldsituaties dan op de andere faculteiten. Ook uit de Kruskal-Wallis test bleek een verschil in kwaliteit van bewijsvoering in het voordeel van faculteit twee. Hiermee wordt de hypothese bevestigd dat op faculteit twee de kwaliteit van bewijsvoering hoger is dan op faculteit één en drie. De kwaliteit van bewijsvoering van de deelnemers wordt verhoogd door in de training in te gaan op het verzamelen van bewijzen en argumenten.

Bij onderzoeksvraag drie: *Is er een verschil in kwaliteit van bewijsvoering per kwaliteitscriterium van het CAP?* blijkt een verschil te bestaan in de kwaliteit van bewijsvoering in de verschillende kwaliteitscriteria. Het overzicht van de rangscores, van hoog naar laag per faculteit, voegt daar echter voor de kwaliteit van bewijsvoering niets aan toe. Op de twee faculteiten waar verschillen zijn gevonden spreken de overzichten van rangscores elkaar tegen. Met enige voorzichtigheid kan men stellen dat voor de kwaliteitscriteria *doel* en *authenticiteit* de bewijsvoeringen hoger zijn. Deze conclusie kan echter niet worden getrokken, omdat de resultaten in de vergelijking van de faculteiten onderling elkaar op faculteit één en drie tegenspreken. Dit is te verklaren aan de hand van het feit dat de context van het CAP op elke onderwijsinstelling zó specifiek is, dat dergelijke vergelijkingen niet te maken zijn. Op elke faculteit is een andere CAP met zijn eigen specifieke context geëvalueerd. In vervolgonderzoek, met een grotere populatie faculteiten, kan dit breder worden onderzocht. Ook kan daarbij onderzocht worden of faculteiten, met een vergelijkbare context voor het CAP, vergelijkbare resultaten opleveren.

Voor de beantwoording van onderzoeksvraag 4: *Is er een verschil in kwaliteit van bewijsvoering in gesprekken waarbij de deelnemers veel dan wel weinig consensus laten zien?* zijn geen verschillen gevonden voor de kwaliteit van bewijsvoering. Dit kan wederom verklaard worden door de specifieke context van elk CAP op de onderwijsinstellingen. Opvallend is dat het type gesprek waarin de

deelnemers voorafgaand aan de groepsdiscussie *geen consensus* laten zien, niet voor komt op de faculteiten. Dit kan te verklaren zijn door het type gesprek zoals is aangegeven door walton (1989). Zoals reeds is aangegeven is het type gesprek in de context van dit onderzoek te plaatsten onder het *onderzoek*. Hierbij is het doel dat twee of meer personen kennis uitbreiden op een bepaald gebied. Het *onderzoek* zoekt bewijs, of een bevestiging van een eerder geformuleerde conclusie. Omdat het zoeken naar toenadering en uitbreiding duidelijk aanwezig is in de context van deze vorm van zelfevaluatie kan dat leiden tot het ontbreken van het type *geen consensus*. Op faculteit twee komt het type gesprek waarin de deelnemers veel consensus laten zien vaker voor dan op faculteit één en drie. Wellicht is er een effect waarneembaar tussen het invullen van de webgebaseerde vragenlijst en een training in het verzamelen van bewijzen en argumenten. Door de training zijn de deelnemers zich meer bewust zijn van bewijzen en argumenten, wat kan leiden tot een meer doordachte onderbouwing van het oordeel wat leidt tot meer consensus. In vervolgonderzoek kan dit effect worden onderzocht.

In vervolgonderzoek wordt aangeraden de training te blijven hanteren. Deze training beperkt zich in de huidige vorm tot de bewijzen: eigen mening, positief of negatief bewijs, een voorbeeldsituatie of verwijzingen naar geschreven informatie of empirische data. In trainingen bij vervolgonderzoek kunnen ook de overige bewijsvoeringen uit het codeerschema worden behandeld. Men kan daarbij denken aan de bewijsvoeringen: geschreven informatie, empirische data, expertbewijs en de tweezijdige of samengestelde bewijzen, zodat ook deze kwaliteit van bewijsvoering meer aandacht kan krijgen in de groepsdiscussies.

Het codeerschema zorgt voor een heldere analyse van soorten bewijs en de kwaliteit daarvan. Vooralsnog is de interbeoordelaarsovereenkomst van het codeerschema laag gebleken. Dit komt door de fenomenen van verlaging door inhoudelijk gelijke beoordeling op verschillende fragmenten en verhoging door op de uitingen waar geen bewijsvoering werd aangetroffen. Hierdoor sluit het bepalen van de interbeoordelaarsovereenkomst niet aan op het inhoudelijke codeerproces. In vervolgonderzoek zou een methode gevonden kunnen worden waarmee het mogelijk kan zijn om het bepalen van de interbeoordelaarsovereenkomst op het inhoudelijke codeerproces aan te laten sluiten. Men moet daarbij uitgaan van een methode waardoor het bewijs op hetzelfde moment in de data geselecteerd wordt. Men kan hierbij denken aan regels zoals bijvoorbeeld het selecteren van het bewijs in de uiting

waarin het bewijs totaal is uitgesproken door een spreker (ongeacht onderbrekingen van andere sprekers). Dit kan leiden tot meer overeenstemming op het fragment in de data waarop onafhankelijke onderzoekers een bewijsvoering selecteren.

Als gevolg van de onderzoeksmethode zijn de resultaten van het onderzoek op dit moment nog niet te generaliseren naar een totale populatie faculteiten. Het verdient daarom de aanbeveling het aantal faculteiten in het onderzoek te vergroten, waardoor een mogelijke generalisatie plaats kan vinden. De verwachting dat het codeerschema ook voor de bewijsvoeringen expertbewijs en de tweezijdige of samengestelde bewijzen dienst kan doen, kan daarbij tevens worden getest.

Bij de kritieken op het gebruik van argumentatietheorieën is aangegeven dat veel critici van mening zijn dat er beslissingen over de inhoudelijke juistheid genomen moeten worden. Men kan niet alleen volstaan met het uitdrukken van kwaliteit in de aan- of afwezigheid van structurele componenten van argumentatie. Door meer vanuit verschillende type bewijzen te kijken is een codeerschema ontstaan dat meer oog heeft voor de inhoudelijke juistheid. Dit is echter niet op de wijze zoals door Sandoval en Millwood (2005) wordt voorgesteld. In vervolgonderzoek zou men alle onderscheiden bewijzen uit dit onderzoek kunnen voorzien van een classificatie naar inhoudelijke juistheid. Vervolgens kan deze classificatie gebruikt worden voor nieuwe data om de classificatie te testen en deze uit te breiden. Door deze classificatie naast het codeerschema uit het huidige onderzoek te leggen, ontstaat een meer inhoudelijke analyse zoals Sandoval en Millwood dat hebben bedoeld. Een dergelijke aanpak kan de kritieken op het gebruik van argumentatietheorie sterker weerleggen. Deze aanpak past in de trend van verschuiving naar modellen die gebaseerd zijn op een sociaal bouwwerk van argumentatie die door Felton en Kuhn (2002) is beschreven.

Literatuurlijst

Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A. & Van der Vleuten, C. P. M. (2006). The wheel of competency assessment: presenting quality criteria for competency assessment programs. *Studies in Educational Evaluation, 32,* 153-170.

Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A. & Van der Vleuten, C. P. M. (2007). Determing the quality of competence assessment programs: a self-evaluation procedure. *Studies in Educational Evaluation, 33,* 258-281.

Baker, M., Andriessen, J., Lund, K., Van Amelsvoort, M. & Quignard, M. (2007). Rainbow: A framework for analyzing computer-mediated pedagogical debates. *International Journal of Computer Supported Collaborative Learning* (2007).

Clark, D.B. & Sampson, V., Weinberger, A. & Erkens, G. (2007). Analytic frameworks for assessing dialogic argumentation in online learning environments. *Educational psychological review, 19,* 343-374.

Clark, D.B., Sampson, V. (2005). Analyzing the quality of argumentation supported by personally-seeded discussions. In T. Koschman, T. Chan & D.D. Suthers (Eds.), *Computer-supported collaborative learning 2005: The next 10 years!* (pp. 76-85). Taipei, Taiwan: Erlbaum.

Erduran, S., Simon, S. & Osborne, J. (2004). TAPping into argumentation: developments in the application of Toulmin's argument pattern for studying science discourse. *Science Education, 88,* 915-933.

Felton, M. K. & Kuhn, D. (2002). The development of argumentative discourse skill. *Discourse Processes, 32*, 135-154.

Field, A. (2005). Discovering statistics using SPSS. Londen: Sage Publications ltd.

Fournier, D & Smith, N.L. (1993). Clarifying the merits or argument in evaluation practice. *Evaluation and Program Planning, 16*, 315-323.

Hornikx, J. (2008). Comparing the actual and expected persuasiveness of evidence types: how good are lay people at selecting persuasive evidence? *Argumentation, 22*, 555-569.

Kuhn, D. & Udell, W. (2003). The development of argument skills. *Child Development, 74 (5),* 1245-1260.

Kyriakides, L. & Campbell, R.J. (2004). School self-evaluation and school improvement: a critique of values and procedures. *Studies in educational evaluation, 30,* 23-36.

McNamara, G. & O'Hara, J. (2005). Internal review and self-evaluation - the chosen route to school improvement in ireland? *Studies in educational evaluation, 31 (4),* 267-282.

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

Moss, P.A., Girard, B.J. & Haniford. L.C. (2006). Validity in Educational Assessment. In J. Green & A. Luke (Eds.), *Review of Research in Education 2006* (pp.109-162). Sage Publications.

Reigeluth, C.M. (1999). What is instructional design theory and how is it changing? In C.M. Reigeluth (Ed.), *instructional design theories and models, volume II* (pp.5-29) Mahwah, NJ, LEA.

Sampson, V. & Clark, D.B. (2008). Assessment of the ways students generate arguments in science education: current perspectives and recommendations for future directions. *Science Education, 92,* 447-472.

Sandoval, W. A. (2003). Conceptual and epistemic aspects of students' scientific explanations. *Journal of the Learning Sciences, 12,* 5-51.

Sandoval, W.A. & Millwood, K.A. (2005). The quality of students' use of evidence in written scientific explanations. *Cognition and instruction, 23 (1),* 23-55.

Schwarz, B. B., Neuman, Y., Gil, J. & Ilya, M. (2003). Construction of collective and individual knowledge in argumentative activity. *Journal of the Learning Sciences, 12 (2),* 219-256.

Toulmin, S. (1958). The uses of argument. Cambridge: Cambridge University Press.

Van Petegem, P., Deneire, A. & De Maeyer, S. (2008). Evaluation and participation in secondary education: Designing and validating a self-evaluation instrument for teachers to solicit feedback from pupils. *Studies in Educational Evaluation, 34,* 136-144.

Walton, D. N. (1989). Dialogue theory for critical thinking. *Argumentation, 3,* 169–184.