Criteria voor het valide examineren van mondelinge examens

Een Delphi studie

Masterthesis Onderwijskundig Ontwerp en Advisering
Universiteit Utrecht

Januari 2012

Auteur:	Dorinde van Loopik
Studentnummer:	3231658
Begeleider:	Dr. M. F. van der Schaaf (Marieke)
Tweede beoordelaar:	Dr. F. J. Prins (Frans)
Opdrachtgever:	College voor Examens

Inhoudsopgave

Criteria voor het valide examineren van mondelinge examens

Een Delphi studie

Dorinde van Loopik

Samenvatting

Er bestaan veel twijfels in de literatuur over de kwaliteit van mondelinge assessments. Een advies dat veel genoemd wordt om de kwaliteit van deze complexe assessmentvorm te verbeteren is het nauwkeurig selecteren, trainen en monitoren van examinatoren. De eerste stap hiervoor is het identificeren van de hoofdtaken van deze examinatoren. In dit onderzoek is dit gedaan door binnen de context van de mondelinge staatsexamens in Nederland, welke uitgevoerd worden onder verantwoordelijkheid van het College voor Examens, een lijst te construeren met criteria voor het valide examineren van deze mondelinge examens. Middels een Delphi studie is getracht consensus en support te vinden voor deze lijst met criteria onder een panel van experts (n = 29). Dit panel heeft elk van de achttien criteria beoordeeld op de relevantie, haalbaarheid en formulering. Na drie rondes was er over het algemeen voldoende support en consensus over de criteria, slechts bij enkele criteria was relatief weinig consensus over de haalbaarheid van het desbetreffende criterium. De definitieve lijst met criteria kan als uitgangspunt gebruikt worden bij het selecteren, trainen en monitoren van examinatoren van mondelinge college-examens. Tevens geven de resultaten van dit onderzoek aanknopingspunten voor het College voor Examens om intern over in discussie te gaan. Ten slotte biedt dit onderzoek een voorbeeld en mogelijk aanzet voor het creëren van een algemeen geldende lijst met kwaliteitscriteria voor mondelinge assessments.

Keywords: Delphi studie; mondelinge examens; validiteit; kwaliteitscriteria

1. Inleiding

1.1 Introductie

Mondelinge assessments worden al sinds lange tijd gebruikt in het medisch onderwijs, maar ook regelmatig bij opleidingen met een gering aantal studenten (Joughin, 1998; Van Berkel & Bak, 2006; Memon, Joughin & Memon, 2010). Bij mondelinge assessments reageert de student verbaal op de assessmenttaak (Joughin, 1998). Een dergelijk assessment kan echter op veel verschillende manier vorm gegeven worden. Volgens Joughin (1998) zijn er zes dimensies waarop mondelinge assessments kunnen verschillen, te weten (1) het primaire type inhoud, (2) de mate van interactie, (3) de mate van authenticiteit, (4) de mate van structuur, (5) het type assessor en (6) de mate waarin het assessment volledig mondeling is ('oraliteit'). De eerste dimensie kan onderverdeeld worden in vier mogelijke opties. Het primaire type inhoud welke getoetst wordt tijdens een mondelings assessment kan volgens Joughin (1998) zijn: kennis en begrip, probleemoplossend vermogen, interpersoonlijke competenties of persoonlijke kwaliteiten. Ook bij de vijfde dimensie zijn er meerdere opties mogelijk. Joughin (1998) stelt dat er drie type assessoren te onderscheiden zijn, te weten: de leerling zelf (*self-assessment*), een medeleerling (*peer assessment*) of een examinator. Bij de overige vier dimensies is er sprake van een continuüm, hetgeen wil zeggen dat bij een bepaald mondeling assessment in meer of mindere mate sprake is van interactie, authenticiteit, structuur en oraliteit. Waarbij de laatste dimensie duidt op het continuüm tussen een assessment welke volledig mondeling is en een assessment waarbij het mondelinge onderdeel secundair is aan een ander onderdeel, bijvoorbeeld bij de presentatie of verdediging van een geschreven essay.

Mondelinge assessments kunnen dus sterk van elkaar verschillen. In het algemeen worden mondelinge assessments echter als een complexe assessmentvorm ervaren en worden er in de literatuur diverse twijfels wat betreft de kwaliteit van mondelinge assessments benoemd (Ang-Aw & Chuen Meng Goh, 2011; Kehm, 2001; Memon, Joughin & Memon, 2010; Wakeford, Southgate & Wass, 1995). Voordat deze twijfels aan de kwaliteit van mondelinge assessments besproken worden, zullen in deze inleiding allereerst kwaliteitscriteria voor assessments in het algemeen behandeld worden. Vervolgens zal de kwaliteit van mondelinge assessments aan de hand van deze kwaliteitscriteria besproken worden, hetgeen resulteert in de vraagstelling die centraal staat in deze thesis. Ten slotte wordt in deze inleiding de specifieke context, waarbinnen deze vraagstelling onderzocht zal worden, toegelicht.

1.2 Kwaliteitscriteria voor (mondelinge) assessments

Psychometrische en edumetrische kwaliteitscriteria

Vanuit de psychometrische traditie, met veelal summatieve, gestandaardiseerde toetsen, wordt de kwaliteit van assessments voornamelijk geëvalueerd aan de hand van de kwaliteitscriteria validiteit en betrouwbaarheid. Met andere woorden, een assessment moet, onafhankelijk van tijdstip of assessor (betrouwbaarheid), datgene meten wat wordt beoogd te meten (validiteit) en niet meer (*construct-irrelevant variance*) of niet minder (*construct underrepresentation*) (Gipps, 1994; Heller, Sheingold & Myford, 1998; Messick, 1995). Met de komst van nieuwe (formatieve) assessment vormen, zoals bijvoorbeeld portfolio en peer assessments, ontstond de discussie of de kwaliteit van deze assessments enkel vastgesteld kon en moest worden aan de hand van deze traditionele, psychometrische kwaliteitscriteria (Gipps, 1994; Dierick & Dochy, 2001). Deze nieuwe vormen van assessment zijn over het algemeen authentieker en daardoor in mindere mate gestandaardiseerd, in vergelijking met traditionele gestandaardiseerde toetsen. Deze assessments zijn niet bedoeld om studenten met elkaar te vergelijken (*norm-referenced*), zoals in de klassieke testtheorie, maar om de ontwikkeling in competentie van een student te toetsen (*criterion-referenced* or *self-referenced*) (Baartman, Bastiaens, Kirschner & van der Vleuten, 2007; Gipps, 2004). Validiteit en betrouwbaarheid alleen geven daarom volgens sommige auteurs een te beperkt beeld van de kwaliteit van deze nieuwe assessment vormen. Zij betogen dat er bij assessments ook rekening gehouden dient te worden met edumetrische kwaliteitscriteria, zoals de transparantie, effectiviteit en authenticiteit van een assessment (Gipps, 1994; Dierick & Dochy, 2001; Stokking, Van der Schaaf, Jaspers & Erkens, 2004).

Mijns inziens zijn zowel de psychometrische als de edumetrische kwaliteitscriteria belangrijke kwaliteitseisen waar assessments aan moeten voldoen en zouden ze aanvullend aan elkaar toegepast moeten worden om de kwaliteit van assessments vast te stellen. Er blijkt echter dat maar weinig auteurs beide type kwaliteitscriteria hebben gecombineerd in één raamwerk van kwaliteitseisen (Poldner, Simons, Wijgaards & van der Schaaf, 2011). Door middel van een literatuurstudie vonden Poldner en collega`s vier raamwerken die psychometrische en edumetrische kwaliteitscriteria bevatten, te weten het raamwerk van Linn, Baker en Dunbar (1991), het raamwerk van Stokking en van der Schaaf (Stokking, van der Schaaf, Japsers & Erkens, 2004; Van der Schaaf & Stokking, 2008), het raamwerk van Birenbaum (2007) en het raamwerk van Baartman (Baartman, Bastiaens, Kirschner & van der Vleuten, 2006, 2007). Poldner en collega`s (2001) hebben in hun onderzoek deze vier verschillende raamwerken vergeleken met Messick`s conceptuele kader van construct validiteit (Messick, 1995). Messick (1995) beschouwt construct validiteit namelijk als een holistisch begrip waaronder aspecten vallen die ook zowel psychometrisch als edumetrisch van aard zijn (Messick, 1995; Poldner et al., 2011). In zijn conceptuele kader is construct validiteit onder te verdelen in zes aspecten, te weten een inhoudsaspect, een substantief aspect, een structureel aspect, generaliseerbaarheid, een extern aspect en een consequentieel aspect. Poldner en collega`s (2011)

hebben deze aspecten aangevuld met betrouwbaarheid en utiliteit, aspecten welke volgens Messick (1995) ook belangrijke kwaliteitcriteria zijn voor assessments.

Op basis van een vergelijking tussen de hierboven genoemde vier raamwerken van kwaliteitscriteria en de aspecten van construct validiteit van Messick (1995) hebben Poldner en collega`s (2011) een raamwerk van *integrated assessment quality criteria* ontwikkeld, zie tabel 1. De betekenis van de afzonderlijke kwaliteitscriteria in dit geïntegreerde raamwerk vinden dus hun oorsprong in de vier raamwerken van Linn en collega`s, Stokking en van der Schaaf, Birenbaum en Baartman en collega`s. Vanuit deze raamwerken en het geïntegreerde raamwerk zullen hieronder de verschillende kwaliteitscriteria per aspect van Messick (1995) kort toegelicht worden.

Tabel 1

Geïntegreerde kwaliteitscriteria voor assessments ingedeeld naar aspecten van construct validiteit

Aspecten van construct validiteit Messick (1995)	Geïntegreerde kwaliteitscriteria voor assessments Poldner en collega`s (2011)
Inhoudsaspect	Trouw aan inhoud
Substantief aspect	Cognitieve complexiteit
Structureel aspect	Scoring
	Specificiteit en ondubbelzinnige interpretatie
Generaliseerbaarheid	Generaliseerbaarheid
Betrouwbaarheid	Betrouwbaarheid
Extern aspect	Externe structuur
Consequentieel aspect	Consequenties voor het onderwijs
	Geschikt voor doel
	Geschikt voor zelf assessment
	Betekenisvol
Utiliteit	Objectiviteit
	Perceptie en gezindheid ten opzichte van het assessment
	Transparantie
	Gelijkheid
	Praktisch nut: werkbaarheid & efficiëntie

Geintegreerde kwaliteitscriteria voor assessments

Zoals te zien is in tabel 1, behoort het kwaliteitscriterium *trouw aan inhoud* van Poldner en collega`s (2011) dus bij het inhoudsaspect van Messick (1995). Dit kwaliteitscriterium gaat, net zoals het inhoudsaspect van Messick (1995), over het belang van de relevantie, representativiteit en authenciteit van de inhoud die getoetst wordt in een assessment. Het kwaliteitscriterium *cognitieve complexiteit*

stelt dat de bedoelde cognitieve processen ook daadwerkelijk getoetst worden in het assessment (substantieve aspect) (Poldner et al., 2011). Het structurele aspect van Messick stelt dat de manier van scoring de structuur van het construct dat getoetst wordt moet weerspiegelen. De bijbehorende kwaliteitscriteria van Poldner en collega`s (2011) gaan dan ook in op de manier van *scoring* en op het belang van *specificiteit en ondubbelzinnige interpretaties*. Met dit laatste kwaliteitscriterium wordt bedoeld dat het assessment adequaat verschillen tussen meer en minder vaardige studenten moet kunnen meten en dat de resultaten niet verklaard mogen worden door andere factoren dan de kennis en vaardigheden die het assessment beoogd te meten, de eerder genoemde *construct-irrelevant variance* (Poldner et al., 2011). Het vierde aspect van Messick (1995), en tegelijkertijd ook het kwaliteitscriterium van Poldner en collega`s (2011), is *generaliseerbaarheid*. In beide gevallen wordt hiermee verwezen naar het belang dat de resultaten van een assessment te generaliseren zijn naar een breder scala aan taken, condities of populaties (Messick, 1995; Poldner et al., 2011). Het aspect en kwaliteitscriterium *betrouwbaarheid* stelt dat het resultaat van een assessment onafhankelijk van het tijdstip of de assessor van het assessment moet zijn. Het externe aspect van Messick (1995) en het kwaliteitscriterium *externe structuur* stellen beide dat het resultaat van het assessment consistent moet zijn met andere assessments die hetzelfde construct meten (convergerend bewijs) en zich tevens moet onderscheiden van assessments die een ander construct beogen te meten (discriminerend bewijs) (Messick, 1995; Poldner et al., 2011).

De laatste twee aspecten met bijbehorende kwaliteitscriteria zijn meer van edumetrische aard (Poldner et al., 2011). Bij Messick`s (1995) consequentieel aspect behoren bijvoorbeeld de kwaliteitscriteria *consequenties voor het onderwijs*, *geschiktheid voor doel*, *geschiktheid voor zelf assessment* en *betekenisvol* (Poldner et al., 2011). Al deze kwaliteitscriteria gaan in op de effecten van het resultaat van het assessment voor het onderwijs. Zo moeten de resultaten geschikt zijn voor het doel van het assessment, dus moet het assessment aansluiten op de leerdoelen en de gegeven instructie. Daarbij achten Poldner en collega`s (2011) het van belang dat het assessment zelf-assessment en zelf-regulerend leren stimuleert. Dit kan bijvoorbeeld als een assessment aangeeft wat de zwakke punten van de lerende zijn (Baartman et al., 2006). Het bij dit aspect laatst genoemde kwaliteitscriterium stelt dat het assessment zowel voor de lerende, als de docent zin- en betekenisvol moet zijn. Dus dat beide partijen zinvolle informatie verkrijgen door middel van het assessment (Poldner et al., 2011). Onder het laatste aspect, de utiliteit van een assessment, scharen Poldner en collega`s (2011) zes kwaliteitscriteria. Ten eerste *objectiviteit* waarmee het belang van duidelijke procedures en criteria bij een assessment wordt aangegeven. Het tweede kwaliteitscriterium, *perceptie en gezindheid ten opzichte van het assessment*, stelt dat het belangrijk is dat de deelnemers van een assessment het doel van het assessment begrijpen, anticiperen op het type assessment en gemotiveerd zijn om hun best te doen voor dit assessment (Birenbaum, 2007). *Transparantie* stelt dat alle procedures, criteria, normen en consequenties vooraf duidelijk moeten zijn voor alle betrokkenen (Poldner et al., 2011). Dat alle deelnemers de dezelfde mogelijkheden moeten krijgen om het

assessment te kunnen voorbereiden en gelijk getoetst en behandeld moeten worden, wordt bedoeld met het kwaliteitscriterium *gelijkheid*. Ten slotte geeft het kwaliteitscriterium *praktisch nut* aan dat het assessment *werkbaar* en *efficiënt* moet zijn.

Kwaliteit van mondelinge assessments

Reeds in de introductie is genoemd dat er veel twijfels omtrent de kwaliteit van mondelinge assessments heersen in de literatuur. De hierboven behandelde geïntegreerde kwaliteitscriteria voor assessments van Poldner en collega`s (2011) bieden een kader waarbinnen de kwaliteit van mondelinge assessments besproken kan worden.

Allereerst kan gesteld worden dat de mate waarin een mondeling assessment voldoet aan het kwaliteitscriterium *trouw aan inhoud* onder andere afhangt van de mate van standaardisatie van het assessment. Mondelinge assessments zijn over het algemeen in veel mindere mate gestandaardiseerd dan traditionele schriftelijke (meerkeuze) toetsen. Daarbij kan de mate van standaardisatie in een mondeling assessment onder andere afhangen van de, reeds in de introductie besproken, dimensies interactie en structuur. Zo is het aannemelijk dat een mondeling assessment minder gestandaardiseerd is als de mate van interactie toeneemt en de mate van structuur afneemt. In een dergelijk mondeling examen met minder standaardisatie zal een grotere diversiteit zijn aan gestelde vragen (Joughin, 1998). Hierbij kan het moeilijker te verantwoorden zijn dat alle gestelde vragen tezamen representatief zijn voor de inhoud (Wakeford, Southgate & Wass, 1995; Memon, Joughin & Memon, 2010). Echter, een mondeling assessment kan meer authentiek zijn in vergelijking met een schriftelijk assessment (de derde dimensie genoemd in de introductie). In dat geval komt deze authenciteit juist ten goede komen aan het kwaliteitscriterium *trouw aan inhoud* (Messick, 1995).

Volgens Wass, Wakeford, Neighbour & van der Vleuten (2003) is een mondeling assessment heel geschikt om de bedoelde (eventueel complexere) cognitieve processen ook daadwerkelijk te toetsen, gezien de mogelijkheid om in een mondeling assessment door te vragen. Dit zou het tweede kwaliteitscriterium, *cognitieve complexiteit*, ten goede komen. De vraag is echter of in een mondeling assessment ook daadwerkelijk doorgevraagd wordt. Er wordt namelijk ook gesignaleerd dat in mondelinge assessments voornamelijk feitenkennis getoetst wordt, mogelijk doordat er over het algemeen beperkte tijd is voor een mondeling assessment (Schuwirth & van der Vleuten, 1996; Wakeford, Southgate & Wass, 1995) .

Onder het structureel aspect van Messick (1995) vallen de kwaliteitscriteria *scoring* en *specificiteit en ondubbelzinnige interpretatie*. Bij deze kwaliteitscriteria is het van belang dat er heldere procedures beschreven zijn voor een nauwkeurige scoring, zodat de eerder genoemde valkuil *construct-irrelevant variance* zoveel mogelijk voorkomen kan worden. Bij mondeling assessments kan een beoordelaar bijvoorbeeld, in tegenstelling tot schriftelijk toetsen, de spreekvaardigheid van een lerende mee laten wegen in de beoordeling, terwijl dit niet onderdeel is van de te toetsen stof.

Het kwaliteitscriterium *generaliseerbaarheid* draagt er zorg voor dat de inhoud van een assessment representatief is, zodat het mogelijk is om de resultaten op de assessment taken te generaliseren naar het construct wat getoetst wordt (Messick, 1995). Generaliseerbaarheid neemt daarom toe wanneer er een grotere steekproef uit de te toetsen inhoud wordt getrokken (Baartman, 2007). Jayawickramarajah (1985) betoogt echter dat zelfs onder de beste omstandigheden een mondeling assessment in dezelfde tijd een minder grote steekproef uit de inhoud kan toetsen dan een schriftelijke toets.

Over de *betrouwbaarheid* van mondelinge assessments zijn in het algemeen veel twijfels. Zoals eerder genoemd, zijn mondelinge assessments over het algemeen in veel mindere mate gestandaardiseerd dan traditionele schriftelijke (meerkeuze) toetsen. Minder standaardisatie betekent bijvoorbeeld meer diversiteit in de gestelde vragen aan kandidaten. Hierdoor kan de betrouwbaarheid van het assessment in mindere mate gewaarborgd en verantwoord worden: het kan per beoordelaar (interbeoordelaarsbetrouwbaarheid) of per examen bij dezelfde beoordelaar (intrabeoordelaarsbetrouwbaarheid) verschillen welke vragen gesteld worden. Daar komt bij dat bij mondelinge assessments over het algemeen meer examinatoren betrokken zijn, in vergelijking met schriftelijke toetsen (gezien de hoeveelheid tijd die meer nodig is bij mondeling assessments), en het ook daardoor moeilijker is om een voldoende interbeoordelaarsbetrouwbaarheid te bereiken (Wass, Wakeford, Neighbour & Van der Vleuten, 2003). Verder geven Memon, Joughin en Memon (2010) terecht aan dat de betrouwbaarheid van mondelinge assessments ook omlaag gaat doordat deze assessments vaak gepaard gaan met subjectieve beoordelingen. Door de zeer beperkte standaardisatie en het feit dat beoordelaar en kandidaat *face to face* een dialoog voeren (in tegenstelling tot schriftelijke toetsvormen), is het aannemelijk dat een beoordelaar zich sneller laat leiden door bijvoorbeeld het uiterlijk of de etniciteit van een student (Wakeford, Southgate & Wass, 1995). Deze *observer bias*, tast de betrouwbaarheid aan en maakt de beoordeling bovendien niet fair. Hetzelfde geldt voor het zogenoemde ‘halo’-effect: het antwoord dat een student geeft op een vraag heeft invloed op de beoordeling van de volgende vraag of een eerder afgenomen mondeling assessment heeft invloed op de beoordeling van een volgend mondeling assessment. Na bijvoorbeeld twee zwakke studenten, kan een mondeling assessment met een matige student opeens heel goed lijken (Van Berkel & Bax, 2006; Wakeford, Southgate & Wass, 1995; Yaphe & Street, 2003).

De mate waarin een mondeling assessment voldoet aan het kwaliteitscriterium *externe structuur* kan vastgesteld worden door het mondeling assessment te vergelijken met andere assessments die hetzelfde beoogt te meten. Hierover zijn echter weinig tot geen empirische gegevens te vinden. Echter, wanneer de betrouwbaarheid van een mondeling assessment laag is (zie voorgaande alinea), zal de mate waarin het resultaat van een mondeling assessment overeenkomt met een resultaat van een ander assessment ook laag zijn.

De *consequenties voor het onderwijs* en de mate waarin een mondeling assessment *betekenisvol* is en *geschikt is voor het doel van het assessment* en *zelf-assessment*, hangt natuurlijk sterk af van de context waarbinnen een mondeling assessment wordt afgenomen. Er kan echter wel een onderscheid gemaakt worden tussen verschillende type mondelinge assessments. Zo benoemt Joughin (1998) enerzijds het type mondeling assessment waarin de communicatie- en/of taalvaardigheden van de student centraal staan en gemeten worden, anderzijds het type waarin de student getoetst wordt op de inhoud van hetgeen hij of zij mondeling verwoordt. Het is aannemelijk dat het mondeling toetsen van communicatie- en/of taalvaardigheden meer betekenisvol en geschikt voor zelf-assessment is, dan wanneer mondeling inhoud getoetst wordt waarbij de talige kant van de inhoud niet behoort tot het object van beoordeling.

Ten slotte kunnen de kwaliteitscriteria behorende bij utiliteit besproken worden in relatie tot mondelinge assessments. Ten eerste kan over de *objectiviteit* van en *gelijkheid* bij mondelinge assessments gesteld worden dat deze over het algemeen minder is dan bij schriftelijke assessments. Het is namelijk aannemelijk dat eigenschappen van studenten, zoals sekse, uiterlijk of sociaaleconomische status, moeilijker te negeren zijn bij een mondeling assessment, in vergelijking met een schriftelijke toets (de eerder genoemde *observer bias*). In het onderzoek van Yaphe en Street (2003) laten examinatoren zich bij de beoordeling van een mondeling assessment bijvoorbeeld beïnvloeden door nervositeit van de student. Verder bleek dat examinatoren in enkele gevallen al binnen de eerste dertig seconden, van het veertig minuten durende examen, een besluit hadden genomen over hun beoordeling van de student (Yaphe & Street, 2003). Dit tast de gelijkheid en objectiviteit van het assessment aan. Wat betreft het kwaliteitscriterium *transparantie* geven Van Berkel en Bax (2006) aan dat studenten over het algemeen minder bekend zijn met mondelinge assessmentvormen en dat dit onduidelijkheid kan geven over wat er van hen verwacht wordt. Dit komt ook de *perceptie en gezindheid ten opzichte van een assessment* niet ten goede. Daarbij komt de reeds vaker genoemde over het algemeen verminderde standaardisatie van mondelinge assessments. Deze beide gegevens maken een mondelinge assessment in minder mate transparant. Dit kan extra stress en spanning veroorzaken bij de studenten (Van Berkel & Bax, 2006; Memon, Joughin & Memon, 2010). Ten slotte zijn er ook bedenkingen te noemen wat betreft het *praktisch nut, de werkbaarheid en efficiëntie* van een mondeling assessment. Dit type assessment is namelijk, zeker in vergelijking met gestandaardiseerde schriftelijke toetsen, een zeer tijdsintensieve assessmentvorm. Zeker bij grote groepen studenten is het meestal efficiënter om gebruik te maken van schriftelijke toetsen (Van Berkel & Bax, 2006).

Kwaliteit van mondelinge assessments vergroten

Om te waarborgen dat mondelinge assessments wel voldoen aan de hierboven beschreven kwaliteitseisen, worden in de literatuur verschillende suggesties gedaan. In veel artikelen over mondeling assessments wordt geadviseerd om, zeker als het gaat om summatieve *high-stakes* assessments, meer standaardisatie aan te brengen in de assessments (Wakeford, Southgate & Wass, 1995; Joughin, 1998; Crisostomo, 2011). Wakeford, Southgate en Wass (1995) stellen bijvoorbeeld voor om enkele identieke vragen bij verschillende kandidaten te stellen om zo een betere vergelijking tussen kandidaten mogelijk te maken, hetgeen de betrouwbaarheid ten goede komt. Ook adviseren Wakeford en collega`s (1995) om elke gestelde vraag of aan bod gekomen onderwerp afzonderlijk te beoordelen. De uiteindelijke beoordeling is gebaseerd op deze gewogen afzonderlijke beoordelingen. Dit kan als hulpmiddel dienen om alles wat aan bod is gekomen tijdens een mondeling examen evenwichtig mee te nemen in de beoordeling, hetgeen het structureel aspect van constructvaliditeit ten goede komt.

Het advies dat echter het meest genoemd wordt om de kwaliteit van mondelinge toetsen te verbeteren is het nauwkeurig selecteren, trainen en monitoren van de beoordelaars in het afnemen en beoordelen van de desbetreffende mondelinge toetsen (Davis & Karunathilake, 2005; Heller, Scheingold & Myford, 1998; Joughin, 1998; Memon, Joughin & Memon, 2010; Muñoz & Álvarez, 2003; Wakeford, Southgate & Wass, 1995; Wass, Wakeford, Neighbour & Van der Vleuten, 2003; Yaphe & Street, 2003). Een meta analyse van Woehr en Arthur (2003) laat daarbij zien dat door assessor training de validiteit van een mondeling assessment verbetert, waarbij een langere training ook meer effect heeft in vergelijking met een kortere.

Een eerste stap naar training voor examinatoren van mondeling toetsen, is het identificeren van de hoofdtaken van examinatoren (Wakeford, Southgate & Wass, 1995). Als duidelijk is wat de taken van examinatoren bij een bepaald mondeling examen zijn, lettend op de dimensies en factoren waarin mondelinge examens kunnen verschillen, kunnen examinatoren op basis hiervan geselecteerd en (herhaaldelijk) getraind worden. Dit zal de kwaliteit van de mondelinge examens verbeteren.

1.3 Vraagstelling

Uit de voorgaande inleiding is gebleken dat mondeling toetsen een complexe toetsvorm is. Het identificeren van de hoofdtaken van examinatoren, om op basis daarvan examinatoren te selecteren en te trainen, is een eerste stap om de kwaliteit van mondelinge toetsen te vergroten. Tot op heden is er weinig literatuur over de hoofdtaken van examinatoren. Er zijn wel diverse richtlijnen voor toetsontwikkelaars of toetsgebruikers bij schriftelijke (gestandaardiseerde) toetsen, zoals het *COTAN Beoordelingssysteem voor de kwaliteit van tests* (2010) of de *Code of fair testing practices in education* (revised) (2004), echter voor mondelinge toetsen zijn er nog geen dergelijke richtlijnen die op grote schaal gehanteerd worden. In dit onderzoek zal daarom getracht worden een lijst op te stellen met criteria welke examinatoren in acht moeten houden voor het valide mondeling examineren.

Zoals in de introductie reeds aangegeven is er niet één type mondeling examen, maar kunnen deze toetsen op onder andere de zes besproken dimensies van Joughin (1998) verschillen. Het is aannemelijk dat per context de hoofdtaken van examinatoren verschillen. Het type mondeling examen dat in dit onderzoek centraal staat zijn de mondelinge college-examens die onderdeel uitmaken van de staatsexamens in het Nederlands voorgezet onderwijs (en welke in de volgende paragraaf nader toegelicht zullen worden). De mondelinge examens binnen deze context waren bij aanvang van dit onderzoek nog nooit eerder structureel onderzocht. Daarbij is ook nooit nauwkeurig vastgesteld wat de hoofdtaken zijn van de examinatoren van deze college-examens. Centraal in dit onderzoek staat daarom de vraag: *Welke criteria dienen examinatoren in acht te houden voor het valide afnemen en beoordelen van mondelinge college-examens in het voortgezet onderwijs?* Waarbij validiteit verwijst naar de eerder besproken holistische benadering van (construct) validiteit van Messick (1995), waaronder ook betrouwbaarheid en edumetrische kwaliteitseisen vallen.

Inzicht in deze criteria voor het valide examineren van mondelinge college-examens biedt daarbij hopelijk ook aanknopingspunten voor belangrijke criteria voor examinatoren die binnen een andere context gebruik maken van mondelinge assessments.

1.4 Context: Mondelinge college-examens in het voortgezet onderwijs

De context van dit onderzoek zijn de mondelinge staatsexamens in het voortgezet onderwijs, welke al sinds het einde van de negentiende eeuw bestaan in het Nederlandse onderwijs. Middels een staatsexamen had men toentertijd de mogelijkheid om in één of enkele vakken een diploma te behalen, aanvullend op een eerder behaald diploma. Volgens de manager staatsexamens voortgezet onderwijs van het College voor Examens, is toen vanwege praktische overwegingen er voor gekozen om deze staatsexamens mondeling af te nemen (Manager staatsexamens VO, mondelinge communicatie, 15 maart 2011). De keuze was dus niet, of in zeer mindere mate, gebaseerd op kwaliteitsoverwegingen.

Vandaag de dag bestaan de staatsexamens nog steeds in het voortgezet onderwijs en vallen sinds 2009 onder toezicht van het zelfstandig bestuursorgaan het *College voor Examens* (College voor Examens, 2010b). Deze staatsexamens vormen een gelijkwaardig alternatief om examen in het voortgezet onderwijs (vmbo, havo of vwo) af te leggen. De staatexamens worden niet door een school georganiseerd, maar worden centraal georganiseerd door het College voor Examens (College voor Examens, 2010b). Deze examens zijn bedoeld voor leerlingen van niet-bekostigde scholen (ofwel particuliere scholen), leerlingen van scholen voor voorgezet speciaal onderwijs (vso) en leerlingen die zich individueel op een examen voorbereiden. Voorbeelden van leerlingen uit deze laatste groep zijn militairen, hoogbegaafde leerlingen, ouderen, gedetineerden of topsporters. In 2010 namen 4000 leerlingen deel aan de staatsexamens (College voor Examens, n.d.).

Staatsexamens bestaan allereerst uit een centraal schriftelijk examen, welke identiek is aan de centrale examens op reguliere scholen voor voortgezet onderwijs en ook op hetzelfde moment wordt afgenomen (in mei). Medio juli vindt dan het tweede deel plaats, het college-examen. Dit is de pendant van het schoolexamen (College voor Examens, n.d.) en bestaat bij de meeste vakken uit een mondeling examen van veertig minuten. Bij praktische vakken zoals muziek en tekenen, bestaat het college-examen uit een praktisch examen. Echter, in deze thesis zal de focus liggen op de mondelinge college-examens bij de taal, zaak- en exacte vakken.

Deze mondelinge college-examens worden afgenomen door twee examinatoren, waarbij één examinator de vragen stelt en de ander de gestelde vragen en de antwoorden van de leerling protocolleert (op 'het protocol'). De examinatoren beschikken over een lesbevoegdheid in het vak dat zij examineren en hebben recente ervaring met het lesgeven aan examenklassen (College voor Examens, 2010a). Tot op heden is er geen (verplichte) training voor de examinatoren en hoeven zij daarnaast ook geen onderwijs of ervaring in het afnemen van mondelinge examens te hebben (Manager staatsexamens VO, mondelinge communicatie, 15 maart 2011).

De mondelinge college-examens van de zaak- en exacte vakken starten over het algemeen met een casus met bijbehorende vragen, welke een leerling 20 minuten voor aanvang van het examen ontvangt om te kunnen bestuderen voorafgaand aan het examen (Dienst Uitvoering Onderwijs, n.d.). De examinator heeft de verantwoordelijkheid een casus te verzorgen en hierbij vragen te construeren. Bij de taalvakken wordt meestal gestart met het bespreken van de door de leerling gelezen literatuur in de desbetreffende taal. De overige tijd van een examen wordt besteed aan het mondeling toetsen van (een selectie van) de overige examenstof. Deze examenstof, met bijbehorende eindtermen, is vastgelegd in de zogenaamde 'vakinformatie', welke openbaar is (Dienst Uitvoering Onderwijs, n.d). Ook hierbij krijgt de examinator de ruimte om zelf vragen te bedenken en te stellen. Wanneer het examen na veertig minuten is afgelopen, verlaat de kandidaat de ruimte en stellen de examinatoren samen, met behulp van het geschreven protocol, het eindcijfer vast. De examinatoren zijn (in duo`s) dus verantwoordelijk voor zowel de constructie (welke onderwerpen en vragen worden bijvoorbeeld behandeld), de afname, als de beoordeling van het examen, waarbij ze erg veel vrijheid krijgen in de manier waarop ze dit doen (Manager staatsexamens VO, mondelinge communicatie, 15 maart 2011).

Aan de hand van de eerder besproken dimensies van Joughin (1998), kan dus gesteld worden dat de mondelinge college-examens een heel open structuur hebben: er is weinig gestandaardiseerd. Daarbij is er sprake van veel interactie tussen de examinator en de kandidaat, het examen neemt de vorm aan van een gesprek met vragen en antwoorden. Wat betreft de dimensies oraliteit en authenticiteit kan gesteld worden dat het examen volledig mondeling is, waarbij er sprake is van weinig authenticiteit (het examen vindt niet in de professionele praktijkcontext van het vak plaats, maar altijd in een gedecontextualiseerde ruimte). Tenslotte zijn de mondelinge college-examens voornamelijk gericht op het toetsen van kennis en begrip, wat gedaan wordt door een externe ('authority-based') examinator.

2. Methode

2.1 Onderzoeksopzet

Om in kaart te brengen welke criteria examinatoren in acht dienen te houden voor het valide afnemen en beoordelen van mondelinge college-examens is een Delphi studie als methode gebruikt. Daarvoor werd eerst een voorstudie uitgevoerd naar onder andere de context van de mondelinge college-examens. Op basis van deze voorstudie en op basis van de literatuur werd vervolgens een eerste lijst met criteria opgesteld. In deze paragraaf zullen deze voorstudie, lijst met criteria en Delphi studie verder besproken worden.

Voorstudie

Voorafgaand aan de constructie van de lijst met criteria, zijn drie interviews afgenomen met zeer ervaren examinatoren, waaronder de oud-voorzitter van de mondelinge college-examens. Het doel van deze interviews was om een gedetailleerder beeld te verkrijgen van de gang van zaken bij de college-examens. Tevens is er gevraagd naar de mening van de respondent over een eerste opzet van de lijst met criteria en de keuzes die gemaakt zijn in het kader van de Delphi studie, zoals de samenstelling van het panel. De informatie en kennis die vergaard is via deze interviews is gebruikt om de lijst met criteria goed aan te passen aan de specifieke context van mondelinge college-examens. Daarnaast is gebruik gemaakt van enkele praktische adviezen die gegeven zijn door de geïnterviewde experts, bijvoorbeeld over de communicatie met de examinatoren.

Ten tweede zijn de uitnodiging voor de Delphi studie en de lijst met criteria voorgelegd aan de opdrachtgever (het College voor Examens) en de oud-voorzitter van de mondelinge college-examens. Doel hiervan was om te controleren of deze documenten voldoende duidelijk, begrijpelijk en juist geformuleerd waren voor de deelnemers. Op basis van hun reacties werden enkele kleine wijzigingen doorgevoerd.

Constructie van de lijst met criteria

De *Code of fair testing practices in education* (revised) (2004) heeft als uitgangspunt gediend voor het construeren van de lijst met criteria. Deze *Code* is opgesteld door de *Joint Committee on Testing Practices* en is consistent met de algemeen geldende *Standards for Educational and Psychological Testing* (1999) van de American Educational Research Association [AREA], American Psychological Association [APA] en National Council on Measurement in Education [NCME] (*Code of fair testing practices in education* (revised), 2004). De code is gebaseerd op zowel psychometrische als edumetrische kwaliteitscriteria, zoals de naam (*'fair testing practices'*) al doet vermoeden.

De *Code* bestaat uit 4 onderdelen: 1) het ontwikkelen en selecteren van geschikte toetsen, 2) het afnemen en scoren van toetsen, 3) het rapporteren en interpreteren van toets resultaten en 4) het informeren van diegene die de toets hebben gemaakt. Daarbij wordt per onderdeel onderscheid gemaakt in criteria die van belang zijn voor de ontwikkelaars van toetsen (*test developers*) en voor de gebruikers van toetsen (*test users*). Aangezien onderdeel één en vier van de *Code* niet van toepassing zijn op de context van mondelinge college-examens, zijn deze buiten beschouwing gelaten. Het valt namelijk niet onder de verantwoordelijkheid van examinatoren om een bepaalde toets te selecteren of om de kandidaten te informeren over hun resultaten. De toetsvorm staat bij de staatsexamens vast en de kandidaten informeren wordt door de locatievoorzitters gedaan en niet door de examinatoren.

De criteria behorende bij de andere twee onderdelen van de *Code*, afname en beoordeling, zijn overgenomen (indien van toepassing) en aangepast aan de context van de mondelinge college-examens en op basis van de literatuur over mondeling toetsen. Criterium 17 is bijvoorbeeld gebaseerd op het vijfde criterium voor *test users* uit het derde onderdeel (het rapporteren en interpreteren van toets resultaten) van de *Code*. De *Code* stelt in dit criterium dat gebruikers van toetsen niet één toets resultaat mogen gebruiken om een beslissing te nemen over de persoon die getoetst wordt. Binnen de context van de mondelinge college-examens is dit te realiseren door de beide examinatoren onafhankelijk van elkaar hetzelfde mondelinge examen te laten beoordelen. Dit wordt vanuit de literatuur sterk geadviseerd (Wakeford, Southgate & Wass, 1995) en draag bij aan het in de inleiding besproken kwaliteitscriterium *betrouwbaarheid*.

Over het algemeen vallen examinatoren van mondelinge college-examens in de categorie *test users*. De criteria die bij deze categorie horen zijn dan ook grotendeels overgenomen en aangepast aan de context van examinatoren van mondelinge college-examens. De taken van *test developers* vallen voornamelijk onder de verantwoordelijkheid van managers en kernteamleden van het College voor Examens. Echter, op bepaalde punten zijn examinatoren ook *test developers* en daarom zijn deze criteria niet op voorhand geheel buiten beschouwing gelaten. Bijvoorbeeld criterium 9 (zie figuur 1), welke ingaat op woordkeuze en taalgebruik bij het stellen van vragen, is van toepassing op examinatoren omdat zij zelf de vragen ontwikkelen en stellen.

Ten slotte zijn, naar aanleiding van de voorstudie, nog twee algemene criteria toegevoegd (criteria 1 en 2, zie figuur 1) aan de lijst. Dit zijn voorwaarden voor het examineren van mondelinge college-examens.

Algemene criteria

1. Examinatoren zijn experts in hun vak en kennen de vakinformatie voor hun vak volledig.
2. Examinatoren hebben affiniteit en recente werkervaring met de doelgroep die zij toetsen.

Criteria bij het afnemen

3. Examinatoren zorgen er voor dat tijdens het examen de verschillende te toetsen domeinen (zoals vastgesteld in de vakinformatie) in voldoende mate aan bod komen.
4. Examinatoren hanteren passende, aangepaste procedures voor kandidaten met handicaps, leerproblemen of andere problemen.
5. Examinatoren stellen, aan het begin van een mondeling examen, kandidaten op hun gemak en laten ze wennen aan de manier van toetsen.
6. Examinatoren protocolleren elke vraag afzonderlijk, zorgvuldig en objectief.
7. Examinatoren geven elke kandidaat dezelfde kansen en zijn onbevooroordeeld.
8. Examinatoren gaan vertrouwelijk om met de gegevens van de kandidaat en zijn/haar examen.
9. Examinatoren formuleren hun vragen duidelijk. De vragen moeten begrijpelijk zijn voor de kandidaat (op zijn/haar niveau) en geen mogelijk aanstootgevende woorden bevatten.

Criteria bij het beoordelen

10. Examinatoren beoordelen kandidaten op basis van de eisen die worden gesteld aan desbetreffende vak en niveau volgens de vakinformatie.
11. Examinatoren zijn zich bewust van mogelijke subjectiviteit van een mondeling examen. Zij laten zich niet beïnvloeden door niet relevante kenmerken van de kandidaat, zoals uiterlijk of nervositeit.
12. Examinatoren beoordelen elk aan bod gekomen onderwerp afzonderlijk. De uiteindelijke beoordeling is het totaal van deze, eventueel gewogen, afzonderlijke beoordelingen.
13. De examinator die de vragen stelt en de examinator die protocolleert beoordelen het examen eerst onafhankelijk van elkaar en komen daarna door middel van overleg tot een gezamenlijk eindcijfer.
14. Examinatoren verwoorden hun beoordeling duidelijk.

Figuur 1: Lijst met criteria (eerste versie) op basis van de *Code*.

Delphi studie

Op basis van de voorstudie en op basis van literatuur (waaronder de *Code*) werd dus een lijst met criteria geconstrueerd. Echter, de vraag is of deze criteria – gedestilleerd uit onderzoeken die plaatsvonden in andere contexten – wel geheel aansluiten bij de specifieke context van de college-examens, zoals hierboven geschetst. Welke criteria examinatoren moeten bezitten zal namelijk ten dele afhangen van de context waarin deze examinatoren opereren. In de context van de college-examens heersen wellicht andere waarden, sociaal-culturele normen en doelen, dan in de context waarin eerder onderzoek is gedaan naar criteria van examinatoren. Vanuit deze verschillen in waarden, normen en doelen zullen andere ideeën over de vraag wat belangrijke criteria zijn bij het examineren voortvloeien. Daarom zal deze eerste lijst met criteria getoetst moeten worden aan de context van de

mondelinge college-examens. Dit is allereerst dus al gedaan door de voorstudie, maar voor breder gedragen support voor deze lijst met criteria is gebruik gemaakt van een conventionele Delphi studie. Deze methode kan volgens Delbecq, Van de Ven en Gustafson (1975) onder andere gebruikt worden om onderliggende assumpties of informatie te verkennen die leiden tot verschillende oordelen. In dit onderzoek kunnen dus middels een Delphi studie de verschillende assumpties van examinatoren, over de vraag wat belangrijke criteria zijn, in kaart gebracht worden.

De Delphi methode was oorspronkelijk bedoeld om technologische prognoses in kaart te brengen. Tegenwoordig zijn er veel meer andere toepassingen van Delphi studies. Vooral in de gezondheidszorg wordt regelmatig gebruik gemaakt van deze survey methode (Delbecq, Van de Ven & Gustafson, 1975; Greatorex & Dexter, 2000). Een Delphi studie beoogt namelijk consensus en steun onder een panel experts te vinden (Linstone & Turoff, 1975). In dit onderzoek bestaat dit panel uit experts op het gebied van de mondelinge college-examens. Deze experts zullen in meerdere, opeenvolgende rondes vragen over de lijst met criteria beantwoorden. Op basis hiervan kan de lijst met criteria worden verbeterd en kan er dus consensus en steun gevonden worden voor deze criteria onder de verschillende experts (Linstone & Turoff, 1975).

Echter, er zijn ook enkele lastige kwesties die bij het gebruik van de Delphi methode besproken moeten worden. Onderzoek van Van der Schaaf en Stokking (2011) laat bijvoorbeeld zien dat het kan zijn dat er statistisch gezien veel consensus is bereikt, maar dat desondanks de onderliggende assumpties en voorkeuren van panelleden nog sterk kunnen verschillen. Een andere lastige overweging is de mate van structuur in de vragenlijsten. Linstone & Turoff (1975) waarschuwen voor te veel structuur, waardoor deelnemers geen mogelijkheid wordt geboden om andere perspectieven of ideeën te benoemen. Preble (1983) waarschuwt daarentegen voor het omgekeerde: mogelijke communicatie problemen door te weinig structuur in vragenlijsten. Tenslotte zijn er geen wetenschappelijke richtlijnen voor het selecteren van experts, de grootte van het panel en ook niet over de vraag wanneer consensus bereikt is (Keeney, Hasson & McKenna, 2005). Deze laatste kwesties zullen ook in het vervolg van deze methodesectie besproken worden.

2.2 Deelnemers

Panelselectie

Bij een Delphi studie wordt dus gewerkt met een panel aan experts, maar wanneer is iemand een expert? Volgens Clayton (1997) is een expert iemand die beschikt over de benodigde kennis en ervaringen om deel te nemen aan een Delphi studie. Echter, wie bepaalt wanneer iemand over voldoende kennis en ervaringen beschikt? Hiervoor moeten criteria worden opgesteld (Delbecq, Van de Ven & Gustafson, 1975; Keeney, Hasson & McKenna, 2005). Deze kunnen per type expert verschillen, zo maken Gorney en Ness (2000) bijvoorbeeld onderscheid tussen implementators, beoefenaars en consumenten. Dit is een nuttig onderscheid voor de selectie van experts in dit

onderzoek. Leden van het kernteam 'staatsexamens voortgezet onderwijs' en locatievoorzitters verwezenlijken elk jaar weer de mondelinge college-examens. Zij hebben veel ervaring met en kennis over het mondeling examineren dat dan plaatsvindt. De tweede groep experts, de zogenaamde beoefenaars, zijn natuurlijk de examinatoren zelf. In overleg met de manager staatsexamens voortgezet onderwijs en de voormalig algemeen voorzitter van de mondelinge college-examens is vastgesteld dat examinatoren, die tevens vakvoorzitter zijn, over het algemeen de meest bekwame examinatoren zijn. Deze vakvoorzitters zullen daarom worden benaderd om deel te nemen aan dit onderzoek. Tenslotte is er de groep consumenten, de examenkandidaten. In dit onderzoek zullen examenkandidaten die vorig jaar (2010) hebben deelgenomen aan de mondelinge college-examens benaderd worden als experts. Selectiecriterium hierbij is dat de kandidaat het gehele staatsexamenprogramma heeft gevolgd en dus voor vrijwel alle vakken een mondeling college-examen heeft afgelegd.

Zoals eerder genoemd zijn er echter geen universeel geaccepteerde richtlijnen voor de hoeveelheid deelnemers en de selectie van deze deelnemers (Keeney, Hassen & McKenna, 2005). Volgens Turoff (1975) is een Delphi methode geschikt voor 10 tot 50 deelnemers. Gomey en Ness (2000) zijn van mening dat dertig deelnemers de optimale groepsgrootte is voor het genereren van ideeën in een Delphi studie. Keeney, Hasson en McKenna (2005) noemen echter diverse Delphi studies met meer dan 100 deelnemers. Deze auteurs waarschuwen ook voor het feit dat er bij een Delphi studie het risico bestaat dat panelleden stoppen gedurende de studie. Vanwege dit serieuze risico werden alle kernteamleden (N = 10), alle locatievoorzitters (N = 3) en alle vakvoorzitters (N = 85) uitgenodigd om deel te nemen. Van deze totale groep van 98 medewerkers, gaven 44 mensen aan deel te willen nemen aan de Delphi studie (van deze 44 mensen hebben 2 mensen geen enkele vragenlijst ingevuld). Tevens werden er oud-staatsexamenkandidaten uitgenodigd. Er is een aselecte steekproef getrokken (n = 100) uit alle kandidaten die een mondeling staatsexamen hebben afgelegd in het schooljaar 2009-2010. Om deze groep extra te stimuleren om deel te nemen, werd hun als dank voor deelname een boekenbon ter waarde van 10 euro aangeboden. Helaas kwamen van deze groep maar 18 positieve reacties op de uitnodiging om deel te nemen.

Beschrijving deelnemers

In het totaal hebben 60 verschillende deelnemers 1 of meerdere rondes geparticipeerd in de Delphi studie. Zoals verwacht is er sprake geweest van uitval: maar 29 deelnemers hebben alle drie de rondes deelgenomen. Deze groep van 29 deelnemers bestond uit 22 medewerkers van de mondelinge college-examens en 7 oud-staatsexamenkandidaten. Hieronder zullen beide subgroepen apart beschreven worden.

Medewerkers mondelinge college-examens. Deze groep bestond uit 17 mannen en 5 vrouwen. Deze deelnemers waren tussen de 49 en 66 jaar oud, met een gemiddelde leeftijd van 58 jaar (SD = 5.03). 20 van de 22 deelnemers waren examinator, de andere twee waren locatievoorzitters en namen

geen examens (meer) af. Van de examinatoren gaf het merendeel aan dat ze de functie van vakvoorzitter vervullen (*n* = 15). Gemiddeld waren deze medewerkers reeds 15-20 jaar (*SD* = 1.90) betrokken bij de mondelinge college-examens en hadden ze 30-35 jaar (*SD* = 1.67) ervaring als docent in het voortgezet onderwijs. Van de groep van 20 examinatoren waren er 11 examinatoren betrokken bij taalvakken, 3 examinatoren bij zaakvakken en 6 examinatoren bij exacte vakken. De meerderheid van deze examinatoren examineerden op havo/vwo-niveau (*n* = 13), de overige 7 examinatoren examineerden op vmbo-niveau.

Oud-staatsexamenkandidaten. De groep oud-staatsexamenkandidaten bestond uit 3 mannen en 4 vrouwen. Deze deelnemers waren tussen de 19 en 24 jaar oud, met een gemiddelde leeftijd van 21 jaar (*SD* = 1.95). Het aantal vakken waarin deze deelnemers een mondeling college-examen hebben afgelegd is gemiddeld 4,57 (*SD* = 3,82). Eén respondent had deze examens afgelegd op het niveau vmbo gl-tl, twee deelnemers op havo-niveau en vier op vwo-niveau.

2.3 Vragenlijsten

Per Delphi ronde is een vragenlijst ontwikkeld op basis van de lijst met criteria. Per criterium werd gevraagd in hoeverre de expert dit criterium: a) relevant, b) haalbaar en c) correct geformuleerd vond binnen de context van de mondelinge college-examens. Deze drie vragen konden beantwoord worden op een 7-puntsschaal (*Likert-type Scale*) van helemaal mee eens tot helemaal mee oneens. Tevens had de expert per criterium de mogelijkheid om een toelichting te geven bij het antwoord en/of een suggestie voor verbetering te doen voor (formulering van) het criterium. Tenslotte werd in de eerste vragenlijst (tijdens de eerste Delphi ronde) gevraagd naar enkele achtergrondvariabelen zoals geslacht, leeftijd, jaren ervaring met/als examinator en in welke vakken. Alle vragenlijsten zijn anoniem afgenomen, hetgeen de invloed van één of enkele dominante meningen beperkt (Preble, 1983; Smith & Simpson, 1995).

2.4 Procedure

Na het uitnodigen van de geselecteerde panelleden, is naar de mensen die wilden deelnemen digitaal de eerste vragenlijst verstuurd. Bij deze eerste vragenlijst werd de belangrijkste achtergrondinformatie over de Delphi studie toegelicht en instructie gegeven hoe de vragenlijst ingevuld moest worden. De panelleden hadden een week de tijd om de vragenlijst in te vullen en te retourneren (het digitaal ‘inleveren’ van de vragenlijst). Op basis van de kwantitatieve en kwalitatieve resultaten van de eerste vragenlijst, is een tweede vragenlijst voor de tweede Delphi ronde ontwikkeld. Ook deze tweede vragenlijst werd digitaal verspreid. Tevens ontvingen panelleden een korte samenvatting van de resultaten van de eerste ronde en de aanpassingen die aangebracht zijn in de lijst met criteria na aanleiding van deze resultaten (de zogenaamde ‘feedback’). Voor de tweede vragenlijst hadden panelleden weer een week de tijd om deze in te vullen. Op basis van de resultaten uit deze tweede

ronde is nogmaals de lijst met criteria aangepast. In de daar op volgende derde ronde ontvingen de panelleden weer een vragenlijst over de aangepaste lijst met criteria en een terugkoppeling van de tweede ronde.

Het aantal rondes van een Delphi studie wordt bepaald op basis van de mate van consensus. Als – volgens vooraf opgestelde vuistregels – gesteld kan worden dat er consensus is bereikt, is de Delphi studie ten einde (Keeney, Hasson & McKenna, 2005; Smith & Simpson, 1995). In dit onderzoek wordt in principe uitgegaan van drie rondes, aangezien is gebleken dat dit meestal voldoende is voor het bereiken van consensus (Van der Schaaf & Stokking, 2011). De gehanteerde criteria voor consensus worden in de volgende paragraaf besproken.

2.5 Analyse

Kwantitatieve analyse

Zoals eerder vermeld, bevatten de vragenlijsten 7-punts schalen, welke daarmee te benaderen zijn als interval meetniveau. Allereerst zijn per item gemiddelden berekend. Het gemiddelde representeert per item de mening van het panel: in hoeverre is dit criterium relevant/haalbaar/correct geformuleerd binnen de context van mondelinge college-examens? Vastgesteld werd dat een gemiddelde van meer dan 5.5 (op een schaal van 1 tot en met 7) een hoge mate van support representeert. Om de mate van consensus vast te stellen zijn de richtlijnen van De Loe (1995) gehanteerd. Hij hanteert de maten *high* (70% van alle scores in één categorie of 80% van alle scores in twee aan elkaar grenzende categorieën), *medium* (60% van alle scores in één categorie of 70% van alle scores in twee aan elkaar grenzende categorieën) en *low* (50% van alle scores in één categorie of 60% van alle scores in twee aan elkaar grenzende categorieën). De laatste maat, *none*, is voor items waarbij minder dan 60% van alle scores in twee aangrenzende categorieën valt. Berekende percentages zijn afgerond naar het meest nabije vijftal, dus bijvoorbeeld 68% werd 70% en 67% werd 65%. Aangezien er gestreefd werd naar consensus én support, is er bij het vaststellen van consensus gekeken naar percentages van de twee hoogste categorieën (6 en 7 van de *likert-scale*). Indien uit de percentages van deze twee hoogste categorieën blijkt dat de consensus bijvoorbeeld *medium* of *high* is, is ook zeker dat de support ook voldoende is ($M > 5.5$). Stel namelijk dat de categorieën 2 en 3 samen 80% van de scores bevatten, dan zou –volgens de richtlijnen van De Loe – de overeenstemming *high* zijn. Echter, de mate van support zal dan niet voldoende zijn volgens het hierboven gestelde criterium ($M > 5.5$).Daarom is alleen gekeken naar de (cumulatieve) percentages van de categorieën 6 en 7.

Ten tweede zijn er per criterium per ronde betrouwbaarheidsanalyses uitgevoerd. Met behulp van Cronbach`s alpha is een inschatting gemaakt van de interne consistentie van de drie vragen (relevantie, haalbaarheid en formulering) per criterium. Op basis van de COTAN-richtlijnen (2010) kan gesteld worden dat deze drie vragen samen een betrouwbare schaal vormen wanneer Cronbach`s alpha groter is dan .60. De interbeoordelaarsbetrouwbaarheid is berekend met behulp van jury alpha`s.

Gekeken naar de mate van consensus, support en naar de opmerkingen die werden gegeven bij de criteria (zie hieronder voor de kwalitatieve analyse), zijn de criteria na de eerste en tweede ronde aangepast. Om te testen of de support is toegenomen tussen rondes, zijn variantie analyses voor herhaalde metingen (*repeated measures ANOVA*) uitgevoerd. Deze toets berekent of de gemiddelden per ronde van elkaar verschillen. Als de gemiddelden tussen rondes zijn toegenomen, is dus de mate van support toegenomen Een voorwaarde van de variantie analyse voor herhaalde metingen is dat sfericiteit ontbreekt. Dit is getoetst met Mauchly`s toets.

Om te toetsen of de mate van support is toegenomen zijn Levene`s toetsen voor homogeniteit uitgevoerd. Deze toets vergelijkt de standaardafwijking tussen rondes. Indien de standaardafwijking is afgenomen, is de mate van consensus toegenomen.

Aangezien een bekend risico van een Delphi studie de hoeveelheid uitval is gedurende de rondes, werden de experts die uitvielen tijdens de Delphi studie vergeleken met experts die alle rondes hebben deelgenomen. Deze twee groepen zijn vergeleken met behulp van een t-toets voor twee onafhankelijke groepen. Op basis van deze resultaten kan de generaliseerbaarheid van de selecte onderzoeksgroep besproken worden.

Kwalitatieve analyse

Per criterium hadden panelleden de mogelijkheid om een toelichting te geven bij de gegeven antwoorden en/of een suggestie voor verbetering te doen voor het desbetreffende criterium. Aangezien dit geen verplichte vraag was, hebben ook niet alle deelnemers bij elk criterium een opmerking of suggestie geplaatst. In het totaal zijn 465 opmerkingen geplaatst, over de drie rondes. De opmerkingen en suggesties die zijn gemaakt, zijn geanalyseerd met behulp van een codeboom, zie tabel 2. Elke uiting is in zijn geheel geanalyseerd (niet gefragmenteerd). Een enkele keer werden er meerdere dingen genoemd, de meest constructieve (ter verbetering van het criterium) is in dat geval gecodeerd.

De codeboom is *top-down* ontwikkeld op basis van de drie vragen die gesteld werden per criterium: de relevantie, de haalbaarheid en de correcte formulering van het criterium. Het streven was om met behulp van de gemaakte opmerkingen de criteria te verbeteren. Indien er dus een suggestie werd gedaan voor verbetering van een bepaald criterium, werd deze ingedeeld in één van deze groepen (Relevantie, Haalbaarheid of Formulering). Lastig hierbij was dat regelmatig in een uiting van een deelnemer een suggestie ter ver betering werd gedaan door middel van het aanpassen van de formulering. Echter, indien dit een inhoudelijke wijziging was, bijvoorbeeld een wijziging die er voor zou zorgen dat een criterium beter haalbaar was, werd deze uiting in dit geval gecodeerd als ‘haalbaar’. Als het een tekstuele suggestie ter verbetering was, bijvoorbeeld over de spelling of de duidelijkheid van de formulering, dan werd deze gecodeerd met ‘formulering’.

Indien er geen enkele suggestie ter verbetering te vinden was in de opmerking, kreeg deze één van de drie andere codes. Deze drie codes zijn *bottom-up* toegevoegd. Als de opmerking in ging op de waarde van het criterium, zonder verdere kanttekeningen of suggesties ter verbetering, kreeg deze de

code ‘positief’. Indien de opmerking in ging op een ander belangrijk aspect van het examineren van mondelinge college-examens, kreeg deze de code ‘nieuw criterium’. Ten slotte waren er behoorlijk veel opmerkingen die niet bruikbaar waren voor de verbetering van de criteria en die niet in gingen op de waarde van het criterium of een suggestie voor een nieuw criterium. Denk daarbij aan irritaties die deelnemers uiten over de vragenlijst of enkel ervaringen met de huidige praktijksituatie. Aangezien in de eerste ronde bleek dat er vrij veel opmerkingen geplaatst waren die niet bruikbaar waren, is er in ronde 2 en 3 extra instructie gegeven over het doel van de ruimte om opmerkingen en suggesties te plaatsen.

Na het vaststellen van een eerste versie van de codeboom, werd deze besproken met een tweede beoordelaar. Om objectiviteit van scoring te garanderen, hebben deze tweede beoordelaar en de auteur beide onafhankelijk van elkaar een random steekproef van 5% van de uitingen gecodeerd. Op basis van deze codes werd Cohen`s Kappa berekend. Deze bleek niet voldoende (Cohen`s Kappa = .50) en na aanleiding van de gecodeerde uitingen werd opnieuw de codeboom besproken. De codeboom is daarna, op basis van deze discussie, opnieuw aangepast: er kwam één code ‘positief’ in plaats van de drie codes ‘relevantie positief’, ‘haalbaarheid positief’ en ‘formulering positief’. Het bleek namelijk lastig hier onderscheid in te maken. Daar komt bij dat deze onderverdeling maar een zeer beperkt nut had voor het aanpassen van de lijst met criteria (positieve opmerkingen geven geen aanleiding tot aanpassing van het criterium) en er relatief weinig opmerkingen in de afzonderlijke categorieën behoorden.

Met deze definitieve codeboom werd nogmaals een nieuwe random steekproef van 5% van de uitingen gecodeerd. Na deze ronde was Cohen`s Kappa nog niet van voldoende grootte, waarna besproken werd wanneer een uiting inging op ‘formulering’ en wanneer op ‘relevantie’ en ‘haalbaarheid’ (zie het hierboven besproken onderscheid). Na een derde keer onafhankelijk coderen van een random steekproef van 5% van de uitingen, bleek de Cohen`s Kappa 0,67. Dit is een indicatie voor een goede overeenstemming. Ten slotte werden alle uitingen, met behulp van deze codeboom en de hierboven beschreven richtlijnen, gecodeerd door de auteur. Zoals eerder vermeld werden deze opmerkingen daarna gebruikt voor het verbeteren van de criteria (behalve wanneer gecodeerd als ‘niet bruikbaar’).

Tabel 2

Codeboom - met omschrijvingen en voorbeelden van codes - gehanteerd bij kwalitatieve analyse

Code	Omschrijving	Voorbeeld
Positief	Positieve uiting over de relevantie, haalbaarheid of formulering van het criterium	"Ik vind deze uiterlijke dingen bij een dermate hoge functie als examinator zeer belangrijk." (Criterium 3, ronde 2)
Relevantie	Negatieve uiting/suggestie ter verbetering wat betreft de relevantie van het criterium	"Overbodig. Beide hebben er een mening over en overleggen. Daar komt een cijfer uit voort." (Criterium 17, ronde 1)
Haalbaarheid	Negatieve uiting/suggestie ter verbetering wat betreft de haalbaarheid van het criterium	"Men is niet altijd in de gelegenheid om alles voor te bereiden. Talendocenten krijgen boekenlijsten bijvoorbeeld op de dag zelf door." (Criterium 5, ronde 2)
Formulering	Negatieve uiting/suggestie ter verbetering wat betreft de formulering van het criterium	"Volgens mij kunnen de woorden 'als' en 'is' worden weggelaten." (Criterium 10, ronde 2)
Nieuw	Suggestie voor een nieuw criterium of het toevoegen van een nieuw onderdeel bij een reeds bestaand ander criterium	"Het kennen is niet voldoende. De examinator moet ook de vertaalslag kunnen maken naar vragen kunnen stellen op het juiste niveau (…)." (Criterium 1, ronde 1)
Niet bruikbaar	Uiting gaat niet over het criterium of is enkel een beschrijving van de huidige praktijksituatie of een ervaring van een deelnemer met dit criterium.	"Ik ken een collega die de kandidaten op hun gemak wil stellen, maar ik denk dat dat niet bij alle kandidaten zo overkomt." (Criterium 9, ronde 1)

3. Resultaten

In dit hoofdstuk zullen allereerst de mate van consensus en support wat betreft de criteria per ronde besproken worden. Dit resulteert in de definitieve lijst met criteria. Daarna worden de resultaten van de betrouwbaarheidsanalyses behandeld. Op basis hiervan kunnen de resultaten tussen rondes besproken worden: is er sprake geweest van een significante toename in consensus en support gedurende de Delphi studie? Ten slotte worden de verschillen tussen de onderzoeksgroep (n = 29) en de deelnemers die zijn uitgevallen gedurende de Delphi studie gerapporteerd.

3.1 Resultaten per ronde

Ronde 1

In de eerste ronde is gestart met de lijst met criteria zoals beschreven en weergegeven in paragraaf 2.1. Van de 42 items zijn er 8 items (19%) die onvoldoende scoren op support, waarbij het gemiddelde dus lager is dan 5.5 (zie tabel 3). Dit betreft alle items van criteria 10 en 17 en de items haalbaarheid van criteria 8 en 15. Al deze items scoren ook laag (*low* of *none*) op consensus. Daarnaast scoren nog veel meer items laag op consensus, slechts enkele items vallen binnen de categorie *high* (zie tabel 3). Verder zijn in de eerste ronde veruit de meeste opmerkingen geplaatst bij de criteria (N = 256). Voornamelijk over het aspect haalbaarheid werden kritische kanttekeningen en suggesties ter verbetering geplaatst (zie tabel 4). Dit is in lijn met de lage support op voornamelijk de haalbaarheid van criteria. De meeste opmerkingen werden geplaatst bij de criteria 10 en 17, de criteria die ook het laagste scoorden op het gebied van support (zie tabel 8 in de bijlage).

Met behulp van deze gegevens zijn er na de eerste ronde aanpassingen gedaan aan de lijst met criteria. De formulering van de huidige criteria werd aangepast, waarbij sommige criteria realistischer werden geformuleerd (in het kader van 'haalbaarheid'). Een duidelijk voorbeeld hiervan is criterium 1, waarbij het zinsdeel 'experts in hun vak' veranderd is in de beter haalbare versie 'bekwaam in hun vak'. Tevens werden er, op basis van de opmerkingen van deelnemers, vier criteria toegevoegd aan de lijst. Twee algemene criteria (criteria 3 en 4) en twee criteria op het gebied van afname (criteria 5 en 7; zie figuur 2). Opvallend is dat de deelnemers de lijst zonder deze vier criteria al voldoende volledig vonden (M = 6.17, SD = .71), maar de afwezigheid van overlap tussen de verschillende criteria onvoldoende beoordeelden (M = 5.24, SD = 1.90; waarbij 1 'veel overlap' representeert en 7 'geen overlap'), zie tabel 3. Met behulp van de opmerkingen is getracht het verschil tussen de criteria, die volgens sommige deelnemers sterk op elkaar leken, duidelijker te maken.

Tabel 3

Gemiddelde, standaarddeviatie en overeenstemming(O) per criterium en per ronde (N =29)

	Ronde 1			Ronde 2			Ronde 3		
Criterium	*M*	*SD*	O	*M*	*SD*	O	*M*	*SD*	O
Criterium 1									
Relevantie	6.76	.51	H	6.72	.59	H	6.55	1.15	H
Haalbaarheid	5.97	.87	H	6.14	.74	H	6.17	.71	H
Formulering*	5.79	1.26	L	6.38	.68	H	6.31	1.26	H
Criterium 2									
Relevantie	6.07	1.16	M	6.03	1.38	M	6.21	1.35	H
Haalbaarheid	5.79	.86	L	5.59	1.12	N	5.83	.85	L
Formulering	5.72	1.46	L	6.00	.93	H	5.90	1.68	H
Criterium 3									
Relevantie	-	-	-	6.31	1.00	H	6.66	.55	H
Haalbaarheid	-	--	-	6.21	.94	H	6.28	.80	H
Formulering	-	-	-	5.90	1.52	M	6.41	1.18	H
Criterium 4*									
Relevantie	-	-	-	6.17	1.00	H	6.38	.94	H
Haalbaarheid	-	-	-	5.97	.91	M	6.28	.75	H
Formulering	-	-	-	6.14	.95	M	6.55	1.15	H
Criterium 5									
Relevantie	-	-	-	6.66	.48	H	6.72	.46	H
Haalbaarheid	-	-	-	6.17	.81	H	6.34	.81	H
Formulering	-	-	-	6.41	.87	H	6.34	1.11	H
Criterium 6									
Relevantie	6.31	.97	H	6.45	.78	H	6.45	.87	H
Haalbaarheid	5.86	1.13	M	6.00	.66	H	5.90	.94	M
Formulering	6.00	1.44	M	6.21	.82	H	6.34	.97	H
Criterium 7									
Relevantie	-	-	-	6.28	1.03	H	6.31	1.31	H
Haalbaarheid	-	-	-	6.00	.89	M	5.83	1.23	H
Formulering	-	-	-	5.86	1.30	L	6.38	1.21	H
Criterium 8									
Relevantie	6.24	1.15	M	6.45	.74	H	6.14	1.55	H
Haalbaarheid	5.48	1.18	L	5.90	.90	M	5.62	1.32	M
Formulering	5.90	1.37	L	5.79	1.42	M	6.14	1.60	H
Criterium 9*									
Relevantie	6.31	1.07	H	6.55	.78	H	6.66	.61	H
Haalbaarheid	5.93	1.10	M	6.48	.74	H	6.55	.74	H
Formulering	6.00	1.34	L	6.52	.83	H	6.66	.61	H
Criterium 10									
Relevantie*	5.41	1.38	N	6.45	.63	H	6.45	.91	H
Haalbaarheid*	5.00	1.67	N	6.03	.82	M	6.17	.85	H
Formulering*	15.34	1.72	N	6.28	1.07	H	6.48	1.15	H
Criterium 11									
Relevantie	6.69	.81	H	6.79	.41	H	6.76	.51	H
Haalbaarheid	5.97	.94	M	6.17	.89	M	6.21	.90	H
Formulering	6.45	.95	H	6.28	1.22	H	6.41	1.21	H

Tabel 3. (*vervolg.*)

	Ronde 1			Ronde 2			Ronde 3		
Criterium	M	SD	A	M	SD	A	M	SD	A
Criterium 12									
Relevantie	6.79	.56	H	6.83	.38	H	6.79	.41	H
Haalbaarheid	6.62	.62	H	6.72	.46	H	6.72	.46	H
Formulering	6.55	1.06	H	6.72	.46	H	6.76	.51	H
Criterium 13									
Relevantie	6.62	.68	H	6.76	.51	H	6.62	.73	H
Haalbaarheid	5.93	.92	M	6.28	.88	H	6.31	.66	H
Formulering	6.07	1.13	L	6.28	1.31	H	6.45	1.15	H
Criterium 14									
Relevantie	6.48	.95	H	6.45	.99	H	6.41	.83	H
Haalbaarheid	6.03	.78	M	6.17	.85	H	5.97	.63	H
Formulering	6.21	1.05	H	6.10	1.18	M	6.17	1.31	H
Criterium 15									
Relevantie	6.31	1.00	H	6.55	.74	H	6.55	.69	H
Haalbaarheid*	5.41	1.09	N	5.79	.77	L	6.07	.92	M
Formulering*	5.97	1.27	M	6.28	1.03	H	6.52	.74	H
Criterium 16									
Relevantie	6.10	1.24	M	6.28	1.00	H	6.14	1.30	H
Haalbaarheid	5.97	1.24	M	6.21	.90	M	6.03	1.09	M
Formulering	6.00	1.39	M	6.31	.81	H	6.38	.94	H
Criterium 17									
Relevantie	4.97	1.86	N	4.93	1.71	N	5.00	1.77	N
Haalbaarheid	5.07	1.67	N	5.21	1.45	N	5.55	1.43	L
Formulering	5.24	1.83	N	5.45	1.82	L	6.14	1.22	H
Criterium 18									
Relevantie	6.55	.63	H	6.55	.63	H	6.41	.87	H
Haalbaarheid	6.00	.93	M	5.97	.87	M	6.10	.77	H
Formulering	6.28	.92	M	6.21	1.05	M	6.41	1.35	H
Overige									
Volledig	6.17	.71	H	6.31	.76	H	6.34	.77	H
Overlap	5.24	1.90	L	4.34	2.02	N	5.00	1.87	N

Noot. M = gemiddelde (1 = weinig support, 7 = veel support); *SD*= standaardafwijking; O = overeenstemming/consensus (H = *high*, 70% van de scores in één categorie of 80% van de scores in twee aangrenzende categorieën; M = *medium*, 60% van de scores in één categorie of 70% van de scores in twee aangrenzende categorieën; L = *low*, 50% van de scores in één categorie of 60% in twee aangrenzende categorieën; N = *none*, minder dag 60% van de scores in twee aangrenzende categorieën), waarbij gekeken is naar enkel de categorieën met de scores 6 en 7; Percentages zijn bij de berekening van consensus afgerond op het meest nabije vijftal.

* Er is sprake van een significante toename in support, zie paragraaf 3.3.

Ronde 2

De aangepaste lijst met criteria werd in de tweede ronde beter beoordeeld. Zoals in de tweede kolom van tabel 3 te zien is, werd de mate van support nu voor alle criteria, met uitzondering van criterium 17, voldoende ($M < 5.5$) beoordeeld. Voor dit criterium was in de eerste ronde ook al onvoldoende support en scoort in de tweede ronde ook nog steeds laag op consensus (*low* en *none*). Daarnaast is er weinig consensus over de haalbaarheid van criterium 2 (*none*) en criterium 15 (*low*). Beide criteria scoorden in de eerste ronde ook al laag op haalbaarheid (respectievelijk *low* en *none*). Verder is over

het algemeen de consensus bij elk criterium toegenomen of gelijk gebleven. De nieuw toegevoegde criteria scoren allemaal voldoende op support en enkel bij criterium 7 is *low* consensus op de formulering. De vraag in hoeverre de deelnemers de huidige lijst volledig genoeg vinden, scoort ruim voldoende (*M*= 6.31, *SD* = .76) en ook de opmerkingen geven geen aanleiding tot het aanvullen van de lijst. Er zijn 3 opmerkingen gemaakt binnen de categorie 'nieuw', maar deze suggesties konden onder gebracht worden bij reeds bestaande criteria (anders dan het criterium waarbij de opmerking is gemaakt). Het totaal aantal gemaakte opmerkingen is meer dan gehalveerd ten opzichte van de eerste ronde (*N* = 124). De bruikbare opmerkingen gingen voor het grootste deel over de haalbaarheid en de formulering van de criteria. Op basis van deze opmerkingen werden de criteria opnieuw aangepast, waarbij dus geen nieuwe criteria zijn toegevoegd. Het betrof voornamelijk kleine aanpassingen in de formulering.

Tabel 4

Frequenties gemaakte opmerkingen per ronde en per code.

Ronde	Positief	Relevant	Haalbaar	Formulering	Nieuw	Niet bruikbaar	Totaal
1	37	19	64	38	6	92	256
2	15	13	27	25	3	39	124
3	0	10	27	28	0	20	85
Totaal	52	42	118	91	9	151	465

Noot. Alle opmerkingen die gemaakt zijn, zijn meegenomen in de analyse. Dus zowel van examinatoren als van oud-staatsexamenkandidaten en zowel opmerkingen van experts die alle drie de rondes hebben geparticipeerd, als van de experts die één of twee rondes hebben deelgenomen.

Ronde 3

De aanpassingen na ronde 2 resulteerde in weer meer consensus en support over de criteria in ronde 3. Alleen voor de relevantie van criterium 17 is onvoldoende support, de overige criteria worden allemaal voldoende beoordeeld. Daarbij is er over het algemeen ook een toename in de mate van consensus. De meerderheid van de criteria valt binnen de categorie *high* consensus. Criteria 2 en 17 blijven helaas op respectievelijk haalbaarheid (*low*) en relevantie (*none*) en haalbaarheid (*low*) laag scoren. Al zijn er bij deze criteria ook verbeteringen te zien ten opzichte van ronde 2. Verder valt bij een paar criteria de haalbaarheid nog niet in de categorie *high*, maar in *medium*, te weten criteria 6, 8, 15 en 16. Alle overige criteria vallen in de categorie *high*.

Het aantal gemaakte opmerkingen daalde nog meer ten opzichte van ronde 2 (*N* = 85), waarbij de meeste bruikbare opmerkingen weer gingen over de haalbaarheid en formulering van criteria. Er werden geen suggesties gedaan voor nieuwe criteria en de deelnemers beoordeelden de volledigheid van de lijst weer ruim voldoende (support: *M* = 6.34, *SD* = .77; consensus: *high*). De afwezigheid van overlap tussen de criteria werd echter nog steeds onvoldoende beoordeeld (support: *M* = 5.00, *SD* = 1.87; consensus: *none*).

Algemene criteria

1. Examinatoren zijn bekwaam in hun vak en hebben de vakinformatie van hun vak paraat.
2. Examinatoren hebben affiniteit en werkervaring met de doelgroep die zij toetsen. Zij hebben deze werkervaring maximaal vijf jaar geleden opgedaan en de ontwikkelingen op hun vakgebied goed bijgehouden.
3. Examinatoren gedragen zich professioneel tegenover zowel kandidaten als collega`s, hetgeen zich bijvoorbeeld uit in op tijd aanwezig zijn, het nakomen van afspraken en het dragen van representatieve kleding.
4. Examinatoren werken op een prettige en constructieve wijze met elkaar samen.

Criteria bij het afnemen

5. Examinatoren bereiden het examen en de materialen die ze daarbij nodig hebben (bijvoorbeeld teksten, casussen, illustraties, hulpmiddelen of vragen) zo goed mogelijk voor.
6. Examinatoren zorgen ervoor dat ze de onderwerpen die ze behandelen tijdens het examen zoveel mogelijk spreiden over de verschillende te toetsen domeinen (zoals vastgesteld in de vakinformatie).
7. Examinatoren laten kandidaten zo veel mogelijk aan het woord, hetgeen afhankelijk kan zijn van context en kandidaat. Zij maken hierbij gebruik van gespreksvaardigheden zoals het anticiperen op de antwoorden van de kandidaat, het aanbrengen van variatie in hun vragen en het soepel kunnen overstappen op een ander onderwerp.
8. Examinatoren nemen kennis van eventuele aangegeven (lichamelijke) handicaps, leerproblemen of andere problemen van een kandidaat uit het voortgezet onderwijs. Zij houden tijdens het examen hier rekening mee, zonder daarbij aanpassingen te doen aan het niveau van het examen.
9. Examinatoren stellen, aan het begin van een mondeling examen, kandidaten op hun gemak en geven kort aan hoe het mondeling examen zal verlopen.
10. De examinator die protocolleert doet dit zorgvuldig, objectief en zo uitgebreid mogelijk.
11. Examinatoren streven ernaar tijdens het examen onbevooroordeeld te zijn en elke kandidaat gelijke kansen te geven.
12. Examinatoren gaan vertrouwelijk om met de gegevens van de kandidaat en (het protocol van) zijn examen.
13. Examinatoren formuleren hun vragen duidelijk en vermijden aanstootgevende woorden (bijvoorbeeld vloeken en racistisch taalgebruik). Het taalgebruik wordt afgestemd op het niveau (vmbo/havo/vwo) waarop de kandidaat examen doet.

Criteria bij het beoordelen

14. Examinatoren beoordelen het examen van kandidaten op basis van de eisen die worden gesteld aan het desbetreffende vak, op het desbetreffende niveau, in de vakinformatie en nemen daarbij de moeilijkheid van de gestelde vragen in acht.
15. Examinatoren zijn zich bewust van de mogelijke subjectiviteit van een mondeling examen. Zij streven ernaar zich bij de beoordeling van een examen niet te laten beïnvloeden door niet relevante kenmerken van de kandidaat, zoals het uiterlijk van een kandidaat.
16. Examinatoren beoordelen elk examenonderdeel, dat aan bod is gekomen in het examen en vermeld staat in de vakinformatie, afzonderlijk. De uiteindelijke beoordeling is het totaal van deze, eventueel gewogen, afzonderlijke beoordelingen.
17. De examinator die de vragen stelt en de examinator die protocolleert geven eerst onafhankelijk van elkaar een cijfer en komen daarna door middel van overleg tot een gezamenlijk eindcijfer.
18. Examinatoren verwoorden hun beoordeling dermate duidelijk, dat de examenvoorzitter deze begrijpt.

Figuur 2. *Definitieve lijst met criteria (criteria 3, 4, 5, en 7 zijn extra toegevoegd na ronde 1)*

Aangezien de criteria, op één item na, allemaal voldoende scoren op support en ook de grote meerderheid van criteria *high* scoort op consensus, is besloten om na de derde ronde geen volgende Delphi ronde toe te voegen. Enkele criteria scoren helaas nog niet hoog genoeg, maar de vraag is of deze criteria wel hoger zouden gaan scoren in volgende rondes. Bij het vergelijken van de scores van criterium 17 over de verschillende rondes, valt op dat deze – ondanks aanpassingen – nauwelijks stijgen (behalve het aspect formulering na de tweede ronde). De meningen op het gebied van relevantie en haalbaarheid lijken fundamenteel te verschillen. Indien het criterium aangepast zou worden op basis van de opmerkingen van het panel van experts, zou het criterium inhoudelijk niet meer consistent zijn met de literatuur. Dit geldt bijvoorbeeld heel duidelijk voor criterium 17. Uit de scores van het item 'relevantie' en de gemaakte opmerkingen bij criterium 17 blijkt dat de meerderheid van het panel er geen waarde aan hecht dat twee beoordelaars onafhankelijk van elkaar een cijfer geven. Zij geven aan dat het in de praktijk meestal door overleg gebeurd en dat deze methode ook hun voorkeur heeft. Doordat deze mening fundamenteel verschilt van de literatuur – die stelt dat het belangrijk is om onafhankelijk van elkaar een examen te beoordelen – is de verwachting dat de support voor dit criterium niet zal toenemen door een volgende Delphi ronde toe te voegen (waarbij het criterium inhoudelijk niet gewijzigd zou worden op basis van de literatuur). In figuur 2 wordt de definitieve lijst met criteria weergegeven. Dit is de lijst welke in de derde ronde beoordeeld is door het panel van experts.

3.2 Betrouwbaarheidsanalyses

Om de interne consistentie tussen de drie beoordeelde aspecten (relevantie, haalbaarheid en formulering) per criterium te meten, is Cronbach`s alpha per criterium berekend. Volgens de richtlijnen van COTAN (2010) is een schaal voldoende betrouwbaar indien Cronbach`s alpha groter is dan .60 (test op groepsniveau). Vanaf .70 is de betrouwbaarheid van de schaal te kwalificeren als 'goed'. Zoals te zien is in tabel 5, is de waarde van Cronbach`s alpha bij een groot aantal criteria in één of meerdere rondes niet voldoende. Dit kan mogelijk verklaard worden doordat relatief weinig items ($n = 3$) gezamenlijk een schaal vormen. Daarbij zijn deze drie items inhoudelijk aparte constructen. De relevantie, haalbaarheid en formulering te samen geven aan hoe het criterium over het geheel beoordeeld wordt, maar beogen niet hetzelfde te meten.

Indien Cronbach`s alpha bij een bepaald criterium in alle rondes voldoende tot goed is, zal in het vervolg van de analyses het schaalgemiddelde bij het desbetreffende criterium gebruikt worden (in plaats van de gemiddelden van losse items). Dit is het geval bij criteria 3, 4, 6, 9, 12, en 17. Bij de overige criteria is Cronbach`s alpha in één of meerdere rondes onvoldoende. Daarom zal bij deze criteria per item geanalyseerd en gerapporteerd worden

Tabel 5

Interne consistentie (Cronbach`s alpha) en beoordelaarsbetrouwbaarheid (Jury alpha) per criterium (C), per ronde (n = 29)

	Ronde 1		Ronde 2		Ronde 3	
	Cronbach`s alpha	Jury alpha	Cronbach`s alpha	Jury alpha	Cronbach`s alpha	Jury alpha
C.1	.27	.90	.60	.88	.84	.61
C.2	.57	-.01	.59	.50	.68	.11
C.3	-	-	.65	.35	.63	.54
C.4	-	-	.73	-.33	.68	.04
C.5	-	-	.51	.76	.72	.73
C.6	.71	.49	.72	.79	.74	.82
C.7	-	-	.46	.29	.90	.85
C.8	.62	.76	.45	.75	.90	.78
C.9	.68	.32	.89	-3.80	.82	-.69
C.10	.83	.31	.46	.54	.84	.58
C.11	.85	.93	.13	.76	.42	.70
C.12	.68	.21	.84	.35	.86	-1.04
C.13	.52	.84	.33	.65	.58	.26
C.14	.67	.65	.74	.46	.52	.53
C.15	.82	.92	.58	.88	.60	.80
C.16	.81	-3.67	.57	-6.07	.77	.35
C.17	.74	-1.88	.86	.52	.75	.88
C.18	.45	.75	.64	.81	.78	.47

.

Tevens zijn er per criterium en per ronde de jury alpha`s berekend. Ook hierbij geldt dat de beoordelaarsbetrouwbaarheid per schaal voldoende is indien alpha groter is dan .60. In tabel 5 is te zien dat bij een heel aantal schalen jury alpha onvoldoende is. Dit duidt op een lage correlatie tussen de beoordelingen (op de criteria) van de verschillende experts. Naar mate de waarde van jury alpha lager is, zijn de experts het dus minder met elkaar eens. De experts zijn het, kijkend naar alle drie de rondes, alleen over criteria 1, 5, 6, 8, 11 en 15 voldoende eens.

Als gekeken wordt naar de interne consistentie en beoordelaarsbetrouwbaarheid van de drie verschillende aspecten (relevantie, haalbaarheid en formulering) waarop elk criterium beoordeeld is, valt op dat nu bijna alle waarden wel voldoende zijn. Alleen in de laatste ronde zijn de experts, op het gebied van de formulering van criteria, het onvoldoende met elkaar eens. Zie hiervoor tabel 6.

Tabel 6

Interne consistentie (Cronbach`s alpha) en beoordelaarsbetrouwbaarheid (Jury alpha) per aspect, per ronde

	Ronde 1		Ronde 2		Ronde 3	
	Cronbach`s alpha	Jury alpha	Cronbach`s alpha	Jury alpha	Cronbach`s alpha	Jury alpha
Relevantie	.79	.89	.89	.90	.93	.88
Haalbaarheid	.79	.81	.83	.80	.90	.78
Formulering	.90	.72	.91	.67	.96	.50

3.3 Resultaten tussen de rondes

Toename consensus en support per aspect

Om te toetsen of de mate van support is toegenomen gedurende de Delphi studie, zijn er variantie analyses voor herhaalde metingen uitgevoerd voor elk aspect waar alle criteria op beoordeeld zijn (relevantie, haalbaarheid en formulering, zie tabel 7). Om te controleren voor de assumptie van afwezigheid van sfericiteit zijn er Mauchly`s toetsen uitgevoerd. Alle drie de aspecten voldeden aan de assumptie. Bij relevantie is er sprake van een significant toename in support, $F(2, 56) = 3.12$, $p = .052$. De posthoc toets geeft aan dat er een significante toename is tussen ronde 1 en 2 ($p = .03$), maar niet tussen ronde 2 en 3 ($p = .54$). Bij haalbaarheid bleek er ook een significante toename in support te hebben plaatsgevonden, $F(2, 56) = 16.68$, $p < .001$. De posthoc toets toont aan dat er tussen ronde 1 en 2 sprake is van een significante toename ($p < .001$). Tenslotte is er ook een significante toename op het aspect 'formulering', gemeten over alle drie de rondes, $F(2, 56) = 7.16$, $p = .002$. Echter, de posthoc toets wijst hier uit dat er een geen significante toename was tussen ronde 1 en 2 ($p = .08$) en tussen 2 en 3 ($p = .07$).

Tabel 7

Gemiddelden en standaarddeviaties per aspect, per ronde

	Ronde 1		Ronde 2		Ronde 3	
	M	*SD*	*M*	*SD*	*M*	*SD*
Relevantie	6.26	.55	6.44	.53	6.40	.68
Haalbaarheid	5.79	.58	6.06	.45	6.11	.55
Formulering	5.97	.87	6.17	.69	6.38	.90

Om te toetsen of de consensus binnen het panel van experts is toegenomen gedurende de Delphi rondes, zijn er Levene`s toetsen voor homogeniteit uitgevoerd. Zoals te zien is in tabel 7 nemen de standaarddeviaties af tussen ronde 1 en 2, maar zet deze afname niet door in het vervolg van de Delphi studie. Levene`s toetsen tonen dan ook geen significante verschillen in standaarddeviaties tussen de opeenvolgende rondes. Dit duidt er op dat de consensus op deze aspecten niet is toegenomen gedurende de Delphi studie.

Toename consensus en support per criterium

De besproken analyses, variantie analyses voor herhaalde metingen en Levene`s toetsen voor homogeniteit, zijn ook uitgevoerd per criterium. Hiermee kan in kaart gebracht worden of het panel van experts meer support en consensus heeft bereikt over de afzonderlijke criteria gedurende de Delphi studie. De toename in support bij de criteria die zijn toegevoegd na de eerste ronde zijn gemeten met behulp van een t-toets voor afhankelijke groepen (dit betreft criteria 3, 4, 5 en 7). In het totaal zijn bij vijf criteria significante verschillen gevonden, waarbij het verschil elke keer een toename in support

betrof (dus de gemiddelde score nam toe per volgende ronde). Deze significante verschillen zullen in het vervolg van deze paragraaf besproken worden en zijn ook gemarkeerd in tabel 3.

Allereerst is er een toename in de support van de formulering van criterium 1, $F(2, 56) = 4.54$, $p = .02$, $\eta_p^2 = .14$. De posthoc toets wijst uit dat het verschil tussen ronde 1 en 2 significant is ($p = .01$, $\eta_p^2 = .23$), maar tussen ronde 2 en 3 niet ($p = .70$). Daarbij is er ook een significant verschil gevonden bij criterium 4, $t(28) = -2.97$, $p = .01$, $d = .55$. Bij dit nieuw toegevoegde criterium in ronde 2 is dus sprake van een toename in support in ronde 3. Bij criterium 9 is er ook sprake van een toename in support, maar hier wees Mauchly`s toets uit dat dit criterium niet voldeed aan de assumptie van afwezigheid van sfericiteit ($\chi^2(2) = .45$, $p < .001$) en daarom is gebruik gemaakt van de Greenhouse-Geisser correctie ($\varepsilon = .65$), $F(1.29, 36.19) = 9.71$, $p = .002$, $\eta_p^2 = .26$. De posthoc toets wijst uit dat de toename alleen tussen ronde 1 en 2 significant is ($p = .01$, $\eta_p^2 = .23$). Bij criterium 10 is er bij alle drie de items een toename vast te stellen. Alle drie de items voldoen niet aan de assumptie van afwezigheid van sfericiteit (Relevantie: $\chi^2(2) = .49$, $p < .001$; Haalbaarheid: $\chi^2(2) = .69$, $p = .01$; Formulering: $\chi^2(2) = .75$, $p = .02$), dus wordt ook nu gebruik gemaakt van de Greenhouse-Geisser correctie ($\varepsilon = .66$; .76; .80). De relevantie verschilt significant tussen ronde 1 en 2 ($p < .001$, $\eta_p^2 = .43$), $F(1.33, 37.11) = 17.39$, $p < .001$, $\eta_p^2 = .38$. De haalbaarheid verschilt ook significant tussen ronde 1 en 2 ($p < .001$, $\eta_p^2 = .40$), $F(1.52, 42.67) = 14.44$, $p < .001$, $\eta_p^2 = .34$. En tenslotte verschilt ook de formulering significant tussen ronde 1 en 2 ($p = .004$, $\eta_p^2 = .27$), $F(1.60, 44.78) = 11.33$, $p < .001$, $\eta_p^2 = .29$. Als laatste zijn significante resultaten gevonden bij de haalbaarheid en formulering van criterium 15. Het item ‘haalbaarheid’ voldeed niet aan de assumptie van afwezigheid van sfericiteit ($\chi^2(2) = .73$, $p = .02$), daarom is gebruik gemaakt van de Greenhouse-Geisser correctie ($\varepsilon = .79$), $F(1.58, 44.24) = 4.92$, $p = .02$, $\eta_p^2 = .15$. Uit de posthoc toets blijkt dat de toename in support tussen ronde 1 en 2 heeft plaatsgevonden ($p = .03$, $\eta_p^2 = .17$). Het item ‘formulering’ voldeed wel aan de assumptie van non-shericity, $F(2, 56) = 3.85$, $p = .03$, $\eta_p^2 = .12$. Uit de posthoc toets blijkt echter dat tussen de afzonderlijke rondes er geen significante toename in support is ($p = .19$ tussen ronde 1 en 2; $p = .17$ tussen ronde 2 en 3).

De Levene`s toetsen wijzen uit dat er bij criteria 10 en 15 tevens sprake is van een significante toename in consensus. Bij criterium 10 op alle aspecten (relevantie, haalbaarheid en formulering allen $p < .001$) en bij criterium 15 op het aspect ‘formulering’ ($p = .03$). Daarnaast zijn er significante verschillen gevonden bij de formulering van criterium 2 ($p = .04$), criterium 6 ($p = .03$) en de formulering van criterium 16 ($p = .02$). Bij deze (aspecten van) criteria is de consensus onder het panel van experts dus toegenomen gedurende de Delphi studie.

3.4 Vergelijking onderzoeksgroep met groep uitgevallen deelnemers

Zoals in paragraaf 2.2 (Deelnemers) al reeds is aangegeven, is er bij deze Delphi studie sprake van veel uitval. Van de 60 deelnemers die in één of meerdere rondes hebben geparticipeerd, hebben er helaas maar 29 alle drie de rondes meegedaan. Ook is er al eerder aangegeven dat, aangezien bekend is dat Delphi studies een groot risico wat betreft (veel) uitval hebben, er geanalyseerd wordt in hoeverre de onderzoeksgroep (n = 29) verschilt van de groep met deelnemers die zijn uitgevallen. In de eerste ronde hebben 23 andere deelnemer geparticipeerd, die daarna nog één of geen enkele ronde hebben geparticipeerd. In de tweede ronde zijn dit 12 deelnemers en in de derde ronde 10 deelnemers.

De twee groepen, de onderzoeksgroep en de groep met uitgevallen deelnemers, zijn met elkaar vergeleken met behulp van een t-toets voor twee onafhankelijke groepen. Daaruit bleek dat er geen verschil zit in de gemiddelde leeftijd, het aantal vakken waarin de oud-kandidaten college-examens hebben afgelegd, het aantal jaren dat de examinatoren docent zijn en betrokken zijn bij de staatsexamens.

De beide groepen zijn ook vergeleken per criterium en per ronde. In de eerste ronde zijn er enkel significante verschillen gevonden bij de beoordeling van de haalbaarheid ($t(48.86) = 2.11$, $p = .04$) en formulering ($t(44.02) = 2.36$, $p = .02$) van criterium 1 en de haalbaarheid van criterium 14 ($t(50) = 2.18$, $p = .03$). In beide gevallen beoordelen de deelnemers die uitgevallen zijn de criteria op desbetreffende criteria hoger dan de onderzoeksgroep. In de tweede ronde beoordelen de uitgevallen deelnemers de haalbaarheid van criterium 1 weer hoger dan de onderzoeksgroep ($t(39) = 2.26$, $p = .03$). Tevens beoordeelt deze groep in ronde 2 ook de formulering van criterium 11 ($t(28.00) = 3.19$, $p = .003$) en criterium 17 ($t(39) = 2.51$, $p = .02$) hoger. Opvallend is dat criterium 17 in de derde ronde juist hoger wordt beoordeeld door de onderzoeksgroep ($t(37) = -2.06$, $p = .05$). Verder zijn er geen significante verschillen gevonden tussen de beide groepen.

4. Discussie

4.1 Bespreking resultaten

In dit onderzoek werd getracht consensus en support te vinden voor een lijst met criteria voor het valide examineren van mondelinge staatsexamens. Op basis van een voorstudie, de *Code of Fair Testing Practices in Education (Revised)* (2004) en literatuur over mondeling examineren is een lijst met criteria opgesteld. Een panel van experts heeft in drie Delphi rondes elk criterium beoordeeld op de aspecten relevantie, haalbaarheid en correcte formulering. Na deze drie rondes kan gesteld worden dat er over het algemeen voldoende consensus en support is gevonden voor de defintieve lijst met criteria. De mate van consensus en support nam hierbij toe per ronde, al was deze toename maar bij enkele criteria statistisch significant. Het aantal gemaakte opmerkingen en suggesties (welke kwalitatief zijn geanalyseerd) per criterium nam af gedurende de Delphi rondes, hetgeen ook kan duiden op meer tevredenheid over de criteria naarmate de Delphi studie vorderde.

Bij enkele criteria was na de derde ronde nog steeds niet voldoende consensus bereikt over het aspect haalbaarheid, maar wel over de aspecten relevantie en formulering. Hetgeen betekent dat de desbetreffende criteria wel belangrijk en juist omschreven worden geacht door het panel, maar zij deze criteria nog niet haalbaar vinden gezien de huidige situatie. Aangezien dit buiten het bereik van dit onderzoek ligt en de desbetreffende criteria wel belangrijk en juist geformuleerd werden bevonden, is besloten dat een volgende Delphi ronde niet de oplossing is om tot meer consensus te komen wat betreft het aspect haalbaarheid. Om deze criteria haalbaar te laten zijn in de praktijk, zullen stappen genomen moeten worden bij de organisatie van de college-examens, het College voor Examens. Deze criteria, met bijbehorende opmerkingen zoals ‘hier is geen tijd voor’, geven aanknopingspunten voor het College voor Examens om de huidige procedures aan te passen, zodat op die manier het handelen volgens deze criteria wel haalbaar wordt.

Alleen over criterium 17, welke stelt dat het belangrijk is om op basis van twee onafhankelijke beoordelingen (van de twee examinatoren) tot een eindcijfer te komen, is na de derde ronde nog onvoldoende consensus over twee aspecten, relevantie en haalbaarheid. Daarnaast is er voor de relevantie van dit criterium ook te weinig support. Het panel blijkt het principieel niet eens te zijn met dit criterium, hetgeen ook blijkt uit het feit dat over dit criterium erg veel opmerkingen zijn geplaatst. Zoals eerder aangegeven, is op basis van deze gegevens ook niet besloten een volgende Delphi ronde in te gaan, aangezien dit criterium alle rondes laag scoorde. Wat betreft dit criterium staat het panel niet achter het principe wat in de literatuur als zeer belangrijk wordt geacht: onafhankelijke beoordelingen (Wass, Wakeford, Neighbour & van der Vleuten, 2003). Ook dit punt geeft een duidelijk signaal naar het College voor Examens om hierover intern duidelijkheid te verschaffen en hierin een keuze te maken.

4.2 Methodologische tekortkomingen

Over het algemeen kan dus gesteld worden dat er na de derde ronde voldoende consensus en support is gevonden voor de lijst met criteria. Het is echter belangrijk om te noemen dat als er statistisch consensus en support aan te tonen is, dit nog niet per definitie stelt dat de onderliggende assumpties en voorkeuren van de panelleden ook gelijk zijn (Van der Schaaf & Stokking, 2011). Dat dit mogelijk ook het geval is in dit onderzoek, bleek uit de kwalitatieve analyse waar bij bijvoorbeeld criterium twee het voorkwam dat twee deelnemers hetzelfde scoorden op de vragen, maar waarbij de onderliggende redenatie van de twee deelnemers verschilde. Zo vond de ene deelnemer dat werkervaring die meer dan vijf jaar geleden opgedaan was, te lang geleden, terwijl de andere deelnemer dit een te korte tijdsperiode vond. Ook dit duidt er op dat het belangrijk is voor het College voor Examens om de definitieve lijst met criteria intern grondig te bespreken.

Tevens moet de kwaliteit van de lijst met criteria ter discussie gesteld worden. Deze lijst werd door de deelnemers in alle rondes als ‘volledig’ beoordeeld, terwijl tegelijkertijd de ‘afwezigheid van overlap’ heel laag beoordeeld werd. Kortom, de deelnemers waren van mening dat er teveel overlap zat tussen de onderlinge criteria. Een verklaring voor deze bevinding is het feit dat criteria vijf tot en met dertien in gingen op de afname van examens en criteria veertien tot en met achttien op de beoordeling van examens. Tussen de beide categorieën van criteria kan een zekere mate van overlap ervaren worden. Zo gaan bijvoorbeeld criterium 11 en criterium 15 beide deels in op het kwaliteitscriterium *gelijkheid* (Poldner et al., 2011), echter het eerste criterium speelt zich af tijdens de afname van het examen en het tweede tweede tijdens de beoordeling. Dit is een belangrijk onderscheid wat in de *Code of Fair Testing Practices in Education (revised)* (2004) wordt gemaakt en ook in deze lijst met criteria aangehouden is.

Ten slotte moeten kritische kanttekeningen geplaatst worden bij de generaliseerbaarheid van de definitieve lijst met criteria. Deze definitieve lijst met criteria is beoordeeld binnen een specifieke context: de mondelinge college-examens. Zoals ook in de inleiding genoemd is, kunnen mondelinge toetsen sterk van elkaar verschillen per context (Joughin, 1998). Daar komt bij dat de definitieve lijst is beoordeeld door een selecte groep deelnemers. Voor de medewerkers van de college-examens geldt dat alle kernteamleden, locatievoorzitters en vakvoorzitters waren uitgenodigd en bij de oud-staatsexamenkandidaten was sprake van een aselecte steekproef. Echter, door de lage respons en de grote uitval zijn de criteria uiteindelijk door een selecte groep beoordeeld. Ook dit zal de resultaten hebben beïnvloed. Al bleek de onderzoeksgroep niet heel sterk te verschillen met de deelnemers die niet alle drie de rondes hebben geparticipeerd (de uitval). Daarbij duiden de verschillen in de meeste gevallen op een onderschatting van de mate van support en consensus bij de onderzoeksgroep. Met andere woorden: de uitgevallen deelnemers beoordeelden de criteria in de meeste gevallen beter.

4.3 Conclusie

Er kan geconcludeerd worden dat examinatoren van mondelinge college-examens de achttien criteria, genoemd in figuur 2, in acht dienen te houden om valide te examineren. Hierbij moet vermeld worden dat deze achttien criteria gelden binnen de specifieke context van de mondelinge staatsexamens in Nederland en niet direct generaliseerbaar zijn naar andere mondelinge assessments.

4.4 Implicaties en aanbevelingen

Tot op heden bestond er bij het College voor Examens nog geen lijst met criteria voor examinatoren van de mondelinge college-examens. De definitieve lijst met criteria, die ontwikkeld is door middel van dit onderzoek, kan allereerst als uitgangspunt dienen om nieuwe examinatoren te selecteren en huidige examinatoren te trainen (Wakeford, Southgate & Wass, 1995). Het nauwkeurig selecteren, trainen en monitoren van examinatoren is namelijk, zoals beschreven in de inleiding, een eerste stap om de kwaliteit van mondelinge assessments te verbeteren (Davis & Karunathilake, 2005; Heller, Scheingold & Myford, 1998; Joughin, 1998; Memon, Joughin & Memon, 2010; Muñoz & Álvarez, 2003; Wakeford, Southgate & Wass, 1995; Wass, Wakeford, Neighbour & Van der Vleuten, 2003; Yaphe & Street, 2003).

Daarbij geeft deze lijst met criteria, en bijbehorende resultaten, aanknopingspunten voor interne discussies bij het College voor Examens over de specifieke uitwerking van deze criteria in de praktijk van mondelinge college-examens. Discussies over bijvoorbeeld het slecht scorende criterium 17 en de criteria die laag scoren op het aspect haalbaarheid. Belangrijk is namelijk dat er intern duidelijkheid is over de geldende procedures om de kwaliteit van examens te waarborgen, bijvoorbeeld over de methode om tot een eindcijfer te komen. Tevens moeten er, in het kader van haalbaarheid, voldoende middelen en tijd geboden worden aan de examinatoren om de examens valide te kunnen af nemen en beoordelen. Dit is zeker van belang aangezien het hier gaat om *high stakes* assessments.

Naast het feit dat deze lijst met criteria gebruikt kan worden bij de mondelinge college-examens van het College voor Examens, kan de lijst ook als uitgangspunt dienen voor mondelinge assessments in andere contexten. Natuurlijk kunnen deze mondelinge assessments sterk verschillen van de mondelinge college-examens die centraal stonden in dit onderzoek. De lijst kan echter wel als voorbeeld of uitgangspunt dienen voor het opstellen van een lijst met criteria binnen een andere context. Zoals in de introductie reeds genoemd is, bestaat er namelijk momenteel geen algemeen geldende lijst met criteria voor mondelinge assessments. Dit onderzoek kan daarom een start zijn om een dergelijke algemeen geldende lijst te ontwikkelen.

Ten slotte zou er in vervolgonderzoek aandacht geschonken kunnen worden aan de overwegingen die ten grondslag liggen aan de keuze voor het inzetten van mondelinge assessments in een specifieke onderwijscontext. Dit viel buiten de reikwijdte van dit onderzoek, maar in de inleiding zijn wel mogelijke valkuilen en interventies besproken om mondelinge assessments kwalitatief goed in te zetten in het onderwijs. Vervolgonderzoek zou op dit theoretische kader verder kunnen bouwen.

Literatuur

Ang-Aw, H.T. & Chuen Meng Goh, C. (2011). Understanding discrepancies in rater judgement on national-level oral examination tasks. *RELC journal, 42* (1), 31-51.

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*(2), 153–170.

Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A. & van der Vleuten, C. P. M. (2007). Evaluating assessment quality in competence-based education: A qualitative comparison of two frameworks. *Educational Research Review, 2*(2), 114–129.

Birenbaum, M. (2007). Evaluating the assessment: Sources of evidence for quality assurance. *Studies in Educational Evaluation, 33*(1), 29–49.

Clayton, M.J. (1997). Delphi: a technique to harness expert opinion for critical decision-making tasks in education. *Educational Psychology, 17*(4), 373-386.

Code of Fair Testing Practices in Education (Revised) (2004). Washington, DC: Joint Committee on Testing Practices.

College voor Examens (n.d.). *Staatsexamens voortgezet onderwijs (VO) – Inhoud.* Gevonden op 15 maart 2011, op http://www.cve.nl/item/staatsexamens_voortgezet_onderwijs

College voor Examens (2010a). Staatsexamens voortgezet onderwijs. Informatie voor nieuwe examinatoren en correctoren. Gevonden op 15 maart 2011, op http://www.cve.nl/item/staatsexamens_voortgezet_onderwijs

College voor Examens (2010b). *Informatie over het CvE.* Gevonden op 15 maart 2011, op http://www.cve.nl/item/vacatures

COTAN (2010). *COTAN beoordelingssysteem voor de kwaliteit van tests.* Amsterdam: NIP.

Crisostomo, A.C. (2011). The effect of standardization on the reliability of the Philippine Board of Surgery oral examinations. *Journal of Surgical Education, 68*(2), 138-142.

Davis, M.H. & Karunathilake, I. (2005). The place of the oral examination in today`s assessment systems. *Medical teacher, 27,* 294-297.

Delbecq, A.L., Van de Ven, A.H. & Gustafson, D.H. (1975). *Group techniques for program planning, a guide to nominal group an Delphi processes.* Glenview: Foresman and Company.

Dienst Uitvoering Onderwijs [DUO] (n.d.). Staatsexamens vakinformatie 2011. Gevonden op 22 maart 2011, op http://www.ib-groep.nl/particulieren/examens/Staatsexamens/Vakinformatie2011/Algemeen.asp

Dierick, S. & Dochy, F. (2001). New lines in edumetrics: new forms of assessment lead to new assessment criteria. *Studies in educational evaluation, 27,* 307-329.

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

Gorney, B. & Ness, R.G. (2000). Evaluation dimensions for full-time head coaches at NCAA division II institutions. *Journal of personnel evaluation in education , 14,* 47-65.

Greatorex, J. & Dexter, T. (2000). An accessible analytic approach for investigating what happens between the rounds of a Delphi study. *Journal of Advanced Nursing, 32*(4), 1016-1024.

Heller, J.I., Sheinhold, K. & Myford, C.M. (1998). Reasoning about evidence in portfolios: cognitive foundations for valid en reliable assessments. *Educational Assessment, 5*(1), 5-40.

Jayawickramarajah, P.T. (1985) Oral examinations in medical education. *Medical Education, 19*, 290–293.

Joughin, G. (1998). Dimensions of Oral Assessment. *Assessment & Evaluation in higher education, 23*(4), 367-378.

Keeney, S., Hasson, F. & McKenna, H. (2005). Consulting the oracle: ten lessons from using the Delphi technique in nursing research. *Journal of Advanced Nursing, 53*(2), 205–212

Kehm, B.M. (2001). Oral examinations at German universities. *Assessment in Education, 8*(1), 25-31.

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

Linstone, H.A. & Turoff , M. (Eds.) (1975). *The Delphi method: techniques and applications,* (pp. 37-71). London: Addison-Wesley Publishing Company.

Memon, M.A., Joughin, G.R. & Memon, B. (2010). Oral assessment and postgraduate medical examinations: establishing conditions for validity, reliability and fairness. *Advances in health sciences education, 15,* 277-289.

Muñoz, A.P. & Álvarez, M.E. (2003). Estimating the validity and reliability of an oral assessment instrument. *Universidad Eafit, 39,* 65-75.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749.

Poldner, E., Simons, P.R.J., Wijngaards, G. & van der Schaaf, M.F., Quantitative content analysis procedures to analyse students` reflective essays: a methodological review of psychometrich and edumetric aspects. *Educational Research Review* (2011).

Preble, J.F. (1983). Public sector use of the Delphi technique. *Technological forecasting and social change, 23,* 75-88.

Schuwirth, L.W.T. & van der Vleuten, C.P.M. (1996). Quality Control: Assessment and Examinations. *Zeitschrift für Hochschuldidaktik, 1-2*, 66-76.

Smith, K.S., & Simpson, R.D. (1995). Validating teaching competences for faculty members in higher education: a national study using the Delphi method. *Innovative Higher Education, 19*(3), 223-234.

Stokking, K., Van der Schaaf, M., Jaspers, J. & Erkens, G. (2004). Teachers` assessment of students` research skills. *British educational research journal, 30*(1), 93-116.

Turoff , M. (1975). The Policy Delphi. In H.A. Linstone, & M. Turoff (Eds.), *The Delphi method: techniques and applications,* (pp. 84-100). London: Addison-Wesley Publishing Company.

Yaphe, J. & Street, S. (2003). How do examiners decide? A qualitative study of the process of decision making in the oral examination component of the MRCGP examination. *Medical education, 37,* 764-771.

Van Berkel, H.J.M. & Bax, A.E. (2006). Toetsen met een mondelinge toets. In H.J.M van Berkel & A.E. Bax (Eds.), *Toetsen in het hoger onderwijs,* (pp. 159-174). Houten: Bohn Stafleu van Loghum.

Van der Schaaf, M.F. & Stokking, K.M. (2011). Construct validation of content standards for teaching. *Scandinavian journal of educational research, 55*(3), 273-289.

Wakeford, R., Southgate, L. & Wass, V. (1995). Improving oral examinations: selecting, training, and monitoring examiners for the MRCGP. *British medical journal, 311,* 931-935.

Wass, V., Wakeford, R., Neighbour, R. & Van der Vleuten, C. (2003). Achieving acceptable reliability in oral examinations: an analysis of the Royal College of General Practitioners membership examination`s oral component. *Medical education, 37*, 126-131.

Woehr, D.J. & Arthur, W. (2003). The construct-related validity of assessment center ratings: a review and meta-analysis of the role of methodological factors. *Journal of management, 29,* 231-258.

Bijlage: Tabel 8

Tabel 8

Frequenties gemaakte opmerkingen per criterium en per ronde.

Criterium	Ronde	Positief	Relevant	Haalbaar	Formu-lering	Nieuw	Niet bruikbaar	Totaal
1	1	3	1	4	7	1	5	21
	2	1	0	1	0	1	4	7
	3	0	0	0	2	0	2	4
2	1	4	1	2	5	0	9	21
	2	1	1	2	6	0	0	10
	3	0	3	2	4	0	0	9
3*	2	1	2	0	4	1	2	12
	3	0	0	1	2	0	0	3
4*	2	0	0	1	1	0	3	5
	3	0	1	0	1	0	0	2
5*	2	0	0	4	0	0	2	6
	3	0	0	3	2	0	0	5
6	1	3	0	5	2	0	9	19
	2	3	0	0	0	0	2	5
	3	0	1	3	1	0	1	6
7*	2	1	3	4	0	0	1	9
	3	0	1	1	2	0	1	5
8	1	2	1	5	1	0	9	18
	2	0	1	3	2	0	3	9
	3	0	1	2	3	0	2	8
9	1	2	2	1	5	0	9	19
	2	1	0	1	0	0	3	5
	3	0	0	0	1	0	1	2
10	1	5	2	16	3	0	2	28
	2	0	1	1	2	1	3	8
	3	0	1	1	1	0	2	5
11	1	1	0	10	3	0	5	19
	2	1	0	1	1	0	2	5
	3	0	0	1	1	0	1	3
12	1	3	1	0	1	0	1	6
	2	0	1	0	0	0	1	2
	3	0	0	0	1	0	1	2
13	1	3	2	1	5	1	7	19
	2	3	1	2	1	0	0	7
	3	0	0	1	1	0	3	5
14	1	2	2	3	0	2	2	11
	2	1	1	1	1	0	1	5
	3	0	0	2	3	0	1	6
15	1	2	1	5	2	0	8	18
	2	0	1	1	3	0	2	7
	3	0	0	1	1	0	2	4
16	1	1	2	2	3	2	3	13
	2	1	0	2	0	0	1	4
	3	0	1	4	0	0	1	6
17	1	5	4	6	0	0	14	29
	2	0	1	3	2	0	6	12
	3	0	1	5	0	0	1	7
18	1	1	0	4	1	0	9	15
	2	1	0	0	2	0	3	6
	3	0	0	0	2	0	1	3

Noot. Alle opmerkingen die gemaakt zijn, zijn meegenomen in de analyse. Dus zowel van examinatoren als van oud-staatsexamenkandidaten en zowel opmerkingen van experts die alle drie de rondes hebben geparticipeerd, als van de experts die één of twee rondes hebben deelgenomen. *Criterium nieuw toegevoegd na ronde 1.