Examens in het hbo; kwaliteitseisen en kwaliteitsborging

Dr. C. Sluijter
Directeur Hoger Onderwijs Cito
Voorzitter Nederlandse Vereniging voor Examens

Dr. G.J.J.M Straetmans
Toetsdeskundige Cito
Lector Saxion Hogescholen

Inhoudsopgave

Inleiding

De Inspectie van het onderwijs heeft in haar rapport 'Alternatieve afstudeertrajecten en de bewaking van het eindniveau in het hoger onderwijs' (2011) kritiek geuit op de borging van de kwaliteit van de examens en diploma's in het hoger onderwijs. De Staatsecretaris van onderwijs heeft onder meer op basis hiervan in zijn strategische agenda 'Kwaliteit in verscheidenheid' aangegeven dat het wenselijk is zowel de externe legitimering van examens beter te borgen als de toetsingspraktijk te versterken. Hij spreekt daarbij zijn voorkeur uit voor 'landelijke toetsing van …kernvakken' en pleit voor het gebruik maken van externe examinatoren indien landelijke toetsing niet mogelijk is.

De HBO-raad heeft begrip voor het streven van de staatssecretaris, maar is beducht voor het ontstaan van kostbare, starre en bewerkelijke instrumenten. De HBO-raad stelt –terecht- dat er een breder spectrum aan vormen van externe validering mogelijk is. Aan het ene uiteinde van dat spectrum bevinden zich zaken als het betrekken van externe deskundigen en gecertificeerde examinatoren. Het andere uiterste ligt bij gestandaardiseerde landelijke examens. In het Hoofdlijnenakkoord is in dit kader met de Staatssecretaris afgesproken dat de HBO-raad een commissie van deskundigen instelt die zal adviseren over mogelijke vormen van externe validering.

Uitgangspunt voor de commissie is dat concrete invulling van externe validering niet voor elke hbo-opleiding dezelfde kan zijn. Welk proces er ook wordt ontwikkeld en geïmplementeerd, het zal rekening moeten houden met verschillen tussen opleidingen en sectoren. De commissie wordt gevraagd met inachtneming van het voorafgaande een advies te geven dat aangeeft:

- wat zowel in technisch instrumentele zin als in een meer algemene context de voor- en nadelen zijn van de verschillende mogelijke vormen van externe validering (inclusief de budgettaire gevolgen);
- welke vorm of vormen van externe validering het meest geschikt is of zijn voor de verschillende sectoren van het hbo;
- binnen welk tijdpad de meest geschikte vorm of vormen van externe validering kan of kunnen worden ingevoerd.

Daartoe dient de commissie in kaart te brengen welke mogelijkheden tot externe validering van examens en diploma's denkbaar zijn. Daarbij zouden de ervaringen die zijn opgedaan in andere Nederlandse onderwijssectoren betrokken moeten worden, zo mogelijk in vergelijking met ontwikkelingen in het buitenland. Verder zou de voor externe validering relevante startsituatie in de diverse sectoren van het hbo geïnventariseerd moeten worden. In het bijzonder die rond al bestaande gemeenschappelijke 'bodies of knowledge', externe examinatoren en landelijke of gemeenschappelijke toetsen. Bij dit laatste zou ook de vraag betrokken moeten worden of bij bepaalde opleidingen of sectoren sprake is van wettelijke regelingen of afspraken met beroepsverenigingen die van invloed kunnen zijn op de keuze voor de meest wenselijke vorm van externe validering.

De HBO-raad heeft de commissie voorgesteld om te starten met het laten schrijven van een drietal essays over de volgende onderwerpen:

- De mogelijkheden voor examinering en/of assessment die voor het hbo relevant zouden kunnen zijn, inclusief de theoretisch-conceptuele afwegingen die aan deze verschillende opties ten grondslag liggen alsmede de praktische consequenties daarvan;
- De mogelijkheden voor externe validering van examenkwaliteit, waarbij tevens wordt betrokken welke mogelijkheid waar in de praktijk (onderwijstype, -niveau etc.) wordt gebruikt;
- De huidige situatie rond validering van examenkwaliteit in het hbo, echter niet in de zin van een uitgebreide 'tour d'horizon', maar casuïstisch, aan de hand van enkele relevante voorbeelden.

De voorliggende notitie heeft tot doel de voor het hbo relevante mogelijkheden voor examinering te beschrijven, maar zal tevens kort ingaan op mogelijkheden voor externe validering.

Examenkwaliteit; begrippenkader en eisen

Een examen is een door een daartoe bevoegde instantie ingesteld onderzoek naar kennis, inzicht, houding en vaardigheden ('knowledge, skills & abilities') van een kandidaat. De kandidaat –leerling, student of cursist- moet over een samenhangend geheel van leergebieden aan de hand van verstrekte opdrachten[1] een prestatie leveren. Op grond van die prestatie kan vervolgens met inachtneming van vastgestelde concrete prestatie-eisen en beslisregels een bewijs –een diploma of certificaat- worden uitgereikt waaraan specifieke rechten of bevoegdheden kunnen worden ontleend door de kandidaat. Het examen moet zodanig van opzet zijn dat het aantal misclassificaties –ten onrechte geslaagden of gezakten- tot een minimum beperkt blijft.

Examens worden ingezet bij het nemen van beslissingen over personen met zwaarwegende consequenties. Resultaten op examens zijn immers in belangrijke mate bepalend voor de verdere onderwijs- en beroepscarrière van personen. Bovendien vormen examens hét middel bij uitstek om de kwaliteit van opleidingen te borgen. Leerlingen, studenten, vervolgopleidingen en werkgevers moeten vertrouwen kunnen stellen in de kwaliteit van examens. De waarde van het aan een examen verbonden diploma staat of valt hiermee (Onderwijsraad, 2010). Het is daarom dat de kwaliteit van examens volledig buiten kijf moet staan. Hoe transparanter en objectiever de borgingsprocedures, des te hoger het vertrouwen dat in het bijbehorende diploma gesteld kan worden. En hoe terechter dat het betreffende diploma een civiel effect heeft.

Om een uitspraak te kunnen doen over de kwaliteit van een examen moet allereerst onderzocht worden in hoeverre het examen voldoet aan twee fundamentele eisen. De eerste eis is die van meetnauwkeurigheid, of betrouwbaarheid (Haertel, 2006) . Een examen voldoet beter aan deze eis, naarmate de scores die kandidaten erop behalen consistenter, en beter reproduceerbaar zijn, kortom vrijer zijn van meetfouten. Betrouwbaarheid is een noodzakelijke, maar onvoldoende voorwaarde voor de kwaliteit van een examen (zie bijvoorbeeld Wools, 2011; pp.70-71). De tweede eis is die van validiteit (Kane, 2006). Validiteit is een overkoepelend begrip dat betrekking heeft op de betekenis, bruikbaarheid en geldigheid van de conclusies die getrokken kunnen worden uit de scores op het examen.

Twee voor de validiteit en dus voor de kwaliteit van examens belangrijke activiteiten zijn normeren en normhandhaving. Normering is een essentieel onderdeel in het proces van examenconstructie en leidt tot een referentiekader dat het mogelijk maakt de scores op het betreffende examen te interpreteren en te waarderen. Bij examens is dit referentiekader altijd absoluut[2]. De scores van kandidaten worden vergeleken met een vastgestelde prestatie- of beheersingsstandaard.

Normhandhaving moet er hierbij voor zorgen dat de prestatie om te slagen voor overeenkomstige examens gelijk blijft. Kandidaten die op verschillende momenten dezelfde opleiding afronden, moeten de beoordeling krijgen die zij verdienen, ongeacht het specifieke examen dat zij voorgelegd krijgen. Er bestaan drie methoden van normhandhaving (De Groot & Van Naerssen, 1969) die we na de nu volgende alinea's over het begrip cesuur en cesuurbepaling zullen behandelen.

Een belangrijk onderdeel bij absolute normering is het vaststellen van de cesuur of standaard op het examen: de scheiding tussen resultaten die als voldoende en onvoldoende worden beschouwd. Voor het bepalen van de cesuur bestaan uiteenlopende procedures. Voor een overzicht zie bijvoorbeeld Hambleton en Pitoniak (2008), of Zieky & Livingston (2008). Er bestaan twee principieel verschillende benaderingswijzen op dit vlak; opdracht- en kandidaatgerichte methodes (Kane, 2001). Bij opdrachtgerichte methodes baseert men zich op oordelen over de moeilijkheid van de opdrachten om de cesuur te bepalen. Het klassieke voorbeeld hiervan is de methode van Angoff (1971), die

[1] Onder opdrachten verstaan we zowel de vragen in kennistoetsen als de taken die een kandidaat in reële of gesimuleerde werksituaties moet uitvoeren als (onderdeel van het) bewijs voor zijn of haar bekwaamheid.

[2] Normering kan ook relatief zijn. Hierbij worden de prestatie van personen afgezet tegen die van andere personen in een referentiegroep. Deze vorm van normering vinden we bijvoorbeeld bij IQ-tests en bij de Eindtoets basisonderwijs van Cito.

overigens verschillende varianten kent. Leerlinggerichte methodes gaan uit van een oordeel over het prestatieniveau van de kandidaten. Een voorbeeld hiervan is de methode van contrasterende groepen (Livingston & Zieky, 1982). Sanders en Verstralen (2011) geven een korte en heldere beschrijving van beide methoden.

Bij de methode van Angoff worden de opdrachten in het examen(onderdeel) beoordeeld door een panel van personen die over voldoende deskundigheid op het betreffende leergebied beschikken en bekend zijn met het prestatieniveau dat van de kandidaten verwacht mag worden. De beoordelaars wordt gevraagd om van elke opdracht in het examen(onderdeel) aan te geven hoe kandidaten erop zullen presteren die naar hun mening over net voldoende kennis, vaardigheden of bekwaamheid beschikken om een voldoende te behalen. Dergelijke kandidaten worden vaak met de term grenskandidaten of zesjeskandidaten aangeduid. De concrete vraag aan de beoordelaars is: 'Wat denkt u dat de gemiddelde score op deze opdracht zal zijn als honderd grenskandidaten deze opdracht uitvoeren?' Hierna worden per beoordelaar de scores voor alle opdrachten opgeteld. Dit is de voorgestelde cesuur van de betreffende beoordelaar. Bestaan er aanzienlijke verschillen tussen de voorgestelde cesuren van de beoordelaars, dan volgt een bespreekronde en hierna worden de opdrachten opnieuw beoordeeld. Blijven er dan nog behoorlijke verschillen bestaan, dan volgt een nieuwe bespreekronde. Zijn de verschillen tussen de verschillende voorgestelde cesuren uiteindelijk acceptabel, dan wordt het gemiddelde ervan als cesuur voorgesteld aan degenen die eindverantwoordelijk zijn voor het vaststellen van de cesuur. Indien de verschillen tussen de voorgestelde cesuren van de verschillende beoordelaars redelijk groot blijven, kan men bij voldoende beoordelaars het gemiddelde van de voorgestelde cesuren berekenen na weglating van de voorgestelde cesuren van de meest extreme beoordelaars.

De methode van contrasterende groepen gaat ervan uit dat docenten in staat zijn kandidaten die een examen(onderdeel) gaan maken over twee of meer contrasterende groepen te verdelen op basis van beoordelingen van hun kennis, vaardigheden of bekwaamheid. Bijvoorbeeld in groepen kandidaten waarvan zij denken dat die voor het examen(onderdeel) zullen slagen en kandidaten die zullen zakken. Vervolgens worden de oordelen van de docenten over welke kandidaten zullen slagen en welke zullen zakken vergeleken met de scores van de kandidaten op het examen(onderdeel). Als cesuur voor het examen(onderdeel) wordt dan gekozen voor de score die de minste misclassificaties oplevert.

Ook bespreken Sanders en Verstralen (2011) kort de 'intuïtieve methode', waarbij docenten de cesuur bepalen op basis van intuïtie en ervaring. Men kiest dan eenvoudigweg voor 55% of 60% van het totaal aantal te behalen scorepunten als cesuur, omdat dat percentage overeen zou komen met het beheersen van iets meer dan de helft van de leerstof. Het grote nadeel van deze methode is, aldus Sanders en Verstralen, dat deze methode erg subjectief is, omdat uit onderzoek bekend is dat docenten de moeilijkheidsgraad van opdrachten moeilijk kunnen inschatten. Er kunnen bij gebruik van deze methode grote verschillen in moeilijkheid ontstaan tussen examens die gelijkwaardig zouden moeten zijn.

De Groot en van Naerssen (1969) maken bij normhandhaving onderscheid tussen de relatieve cesuurmethode, de absolute normering, compromismethoden, expertoordelen en normhandhaving via deelsets van opgaven die van opeenvolgende toetsen deel uitmaken. Bij de relatieve cesuurmethode is de aanname dat opeenvolgende cohorten van kandidaten gelijkwaardig zijn. De cesuur wordt dan steeds zo gekozen dat het percentage geslaagden in opeenvolgende cohorten gelijk is. In feite komt dit neer op de zojuist beschreven intuïtieve methode. Bij absolute normering is de aanname dat de toetsen gelijkwaardig zijn en wordt dus bij opeenvolgende cohorten kandidaten de cesuur bij dezelfde toetsscore gelegd. Compromismethoden en expertoordelen zijn volgens de Groot en van Naerssen minder exacte procedures dan normhandhaving via deelsets van opgaven. Bij deze laatste techniek maken deze deelsets het mogelijk om statistische methoden toe te passen waarmee scores –en dus de cesuur- op de ene toets vertaald kunnen worden in scores op de andere toets.

Twee andere definities die nodig zijn om verwarring bij het lezen van de rest van deze notitie te voorkomen, zijn die van 'toets' en 'assessment'. Een toets is een instrument voor het meten van kennis[3] die door middel van studie, praktijkervaring en/of onderwijs op een bepaald vakgebied verworven is. Dat gebeurt door het voorleggen van een reeks vragen. De term assessment wordt in deze notitie gebruikt als containerbegrip om andere, relatief nieuwe, toetsvormen aan te duiden. Dit kunnen bijvoorbeeld (computer)simulaties zijn, proeven van bekwaamheid of situatiebeoordelings-testen, maar ook observaties op de werkplek. De term is in zwang gekomen, sinds in verschillende onderwijsvormen in binnen- en buitenland leeromgevingen zijn gecreëerd waarin competent handelen centraal staat (Mulder, 2003). Overigens heeft de term assessment in de literatuur een veel ruimere betekenis: "Assessment is any systematic procedure for collecting information that can be used to make inferences about the characteristics of people or objects (AERA, 1999). In deze notitie wordt hiervan afgeweken omwille van helderheid van het betoog.

In de praktijk is sprake van inconsistent gebruik van de termen 'examen' en 'examens'. Zo is het in het voortgezet onderwijs gebruikelijk om te spreken van het havo-, of vwo-examen voor een specifiek vak, terwijl formeel sprake is van een examenonderdeel. Het havo- en vwo-examen bestaan uit een reeks toetsen voor verschillende vakken, waar een leerling cijfers voor behaalt. Dat cijfer wordt bepaald op basis van de score op de toets en de positie van de cesuur. Of de leerling slaagt of zakt voor het examen, wordt bepaald via een beslisregel die uitgaat van de cijfers op alle examenonderdelen – toetsen- tezamen.

Ook in het hbo is dit vrijwel zonder uitzondering het geval. Studenten ontvangen hun diploma op basis van een 'dossier' waarin hun prestaties op een reeks examenonderdelen zijn opgeslagen. De student is geslaagd voor het examen, zodra zijn of haar prestaties op de verschillende onderdelen tezamen zodanig zijn dat aan het examen gekoppelde uitslagregel een positieve uitslag geeft. Om verwarring op dit vlak te voorkomen, zal in deze notitie daar waar nodig de term examenonderdeel gebruikt worden.

Het op een transparante en voor alle stakeholders acceptabele wijze bepalen van de cesuur op een examen of examenonderdeel is van groot belang. Omdat een examen, hoe goed ook, altijd behept is met meetfouten, is het niet mogelijk misclassificaties te voorkomen. Het percentage misclassificaties wordt niet alleen bepaald door de meetnauwkeurigheid van het examen, maar ook door de keuze van de cesuur. Ligt de cesuur relatief laag, dan zullen er maar weinig kandidaten zijn die ten onrechte zakken, maar relatief veel kandidaten die ten onrechte slagen. Ligt de cesuur relatief hoog, dan zullen er maar weinig kandidaten zijn die ten onrechte slagen, maar relatief veel kandidaten die ten onrechte zakken. Hoe streng of mild de beslissende instantie zich opstelt, is afhankelijk van de consequenties van de beide vormen van misclassificaties; onterecht zakken en onterecht slagen. Hoewel voor individuele kandidaten niet valt te bepalen of sprake is van een misclassificatie, is het wel mogelijk op basis van de betrouwbaarheid van het examen of examenonderdelen het percentage misclassificaties te schatten (Sanders, 2011). Voor de centrale examens havo en vwo is dat bijvoorbeeld recentelijk concreet uitgerekend (Van Rijn, 2009; Van Rijn, Beguin & Verstralen, 2009)

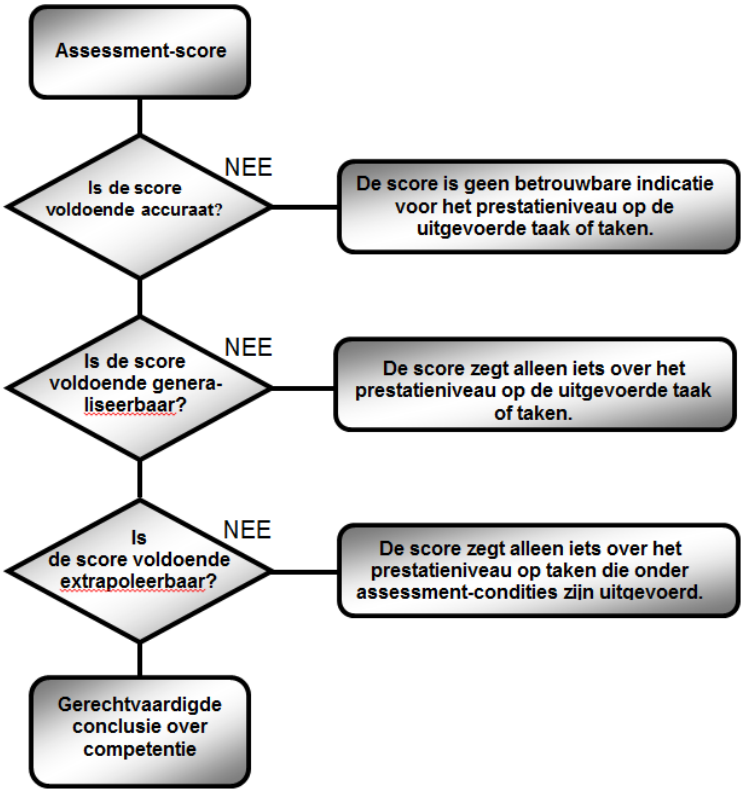
De scope van deze notitie blijft beperkt tot examens, c.q. examenonderdelen die zich richten op het beantwoorden van de vraag of een kandidaat gezien kan worden als een reflecterende professional; een beroepsbeoefenaar die tot meer in staat is dan het mechanisch uitvoeren van ingeoefende procedures en die ook in onbekende probleemsituaties bruikbare oplossingen vindt. De term beroepsgerichte examens wordt hiervoor gereserveerd. In de loop van deze notitie hopen we duidelijk te maken dat beroepsgerichte examens het best uit een combinatie van toetsen en assessments kunnen bestaan.

[3] Kennis wordt hier gebruikt zoals bedoeld door Bloom (Bloom, 1975, Anderson & Krathwohl, 2001). Hierbij is sprake van vier kennisdimensies; feitenkennis, conceptuele kennis, procedurele kennis en metacognitieve kennis en zes handelingsdimensies; oproepen, begrijpen, toepassen, analyseren, evalueren en creëren.

Kwaliteitseisen voor assessments

Voor het beschrijven van de eisen waar assessments aan moeten voldoen, gebruiken we het gedachtengoed van Straetmans (2004a) Hij stelt dat de informatie die assessments opleveren voldoende accuraat, generaliseerbaar en extrapoleerbaar moet zijn. De termen generaliseerbaarheid en extrapoleerbaarheid gebruikt hij om de validiteitseisen voor assessments in generieke zin te preciseren. Accuraatheid is de term die Straetmans hanteert voor meetnauwkeurigheid in plaats van de meer gebruikelijke term betrouwbaarheid. Dat is gepast, omdat bij assessments in de regel een dichotome beslissing - iemand is bekwaam of niet- genomen wordt en er dus ook sprake van misclassificaties. De term accuraatheid wordt op verschillende vakgebieden gebruikt om te refereren aan het percentage misclassificaties bij dichotome beslissingen. Bijvoorbeeld in de medische wetenschap bij het stellen van diagnoses.

Figuur 1 laat zien dat het alleen mogelijk is om een onderbouwd antwoord te geven op de vraag of een kandidaat geslaagd is, indien de informatie die een assessment oplevert voldoet aan de drie eerder genoemde eisen. Informatie is accuraat als deze een voldoende nauwkeurige indicatie geeft van het prestatieniveau op de uitgevoerde opdrachten. Een gebrek aan accuraatheid kan bijvoorbeeld tot uiting komen indien verschillende beoordelaars een andere waardering geven aan dezelfde scriptie. Kennelijk spelen dan andere factoren een rol dan uitsluitend de prestatie van de kandidaat of keken de beoordelaars anders aan tegen bepaalde aspecten van de prestatie. De accuraatheid zal toenemen, indien gebruik gemaakt wordt van eenduidig geformuleerde prestatiecriteria die zich richten op concreet waarneembaar gedrag. De accuraatheid zal ook toenemen, indien men gebruik maakt van meer dan één beoordelaar en uitgaat van het gemiddelde oordeel. Ook het trainen van beoordelaars in het werken met de prestatiecriteria kan een positief effect hebben op de accuraatheid (Sanders, 2011).



Figuur 1; Kwaliteit en interpretatie van assessmentscores (Straetmans, 1998)

Informatie is generaliseerbaar als deze het mogelijk maakt uitspraken te doen over alle mogelijke andere opdrachten waar het assessment ook betrekking op had kunnen hebben. Om de generaliseerbaarheid van de informatie te waarborgen is het nodig om zoveel mogelijk verschillende opdrachten aan te bieden in een assessment. Informatie is tot slot extrapoleerbaar als op basis ervan ook iets gezegd kan worden over prestaties die in de werkelijke beroepscontext verwacht mogen worden. Zo zal de informatie die een assessment oplevert beter extrapoleerbaar zijn, naarmate de opdrachten en situaties waarin kandidaten ze moeten uitvoeren meer gelijkenis vertonen met die uit de toekomstige reële werksituatie.

Verschillende auteurs, zoals Frederiksen & Collins (1989), Haertel (1991), Linn, Baker & Dunbar (1991) en Baartman, Bastiaens, Kirschner & van der Vleuten (2007), pleiten voor de introductie van nieuwe kwaliteitscriteria die meer rekening houden met de specifieke eigenschappen van assessments dan de klassieke eisen van betrouwbaarheid en validiteit. In het kader van deze notitie is het niet nodig expliciet stil te staan bij de verruimde of aangepaste criteria. Ze hebben namelijk geen invloed op de conclusie die in dit essay ten aanzien van de vormgeving van beroepsgerichte examens getrokken zal worden. Het is de kunst om vanuit onderwijsmeetkundig oogpunt optimale informatie te verzamelen bij examens, daarbij rekening houdend met randvoorwaarden als de beschikbare tijd voor ontwikkeling en afname en middelen in de vorm van geld en geschikte cases.

Een praktische indeling van assessments

In de literatuur is sprake van een groot scala aan termen die gebruikt worden om assessmentvormen te omschrijven. Voor een overzicht zie bijvoorbeeld Van Berkel en Bax (2008), Dijkstra (2011) of . Binnen het bestek van deze notitie is het niet doenlijk om een uitputtende lijst op te stellen en van iedere mogelijke assessmentvorm aan te geven in hoeverre deze aan de drie genoemde kwaliteitseisen, of daarvan afgeleide eisen, voldoet. Het is dan ook zaak om tot een meer globale indeling te komen.

Roelofs en Straetmans (2006) en Straetmans (2004a) categoriseerden assessments in een globale driedeling van ‘hands-on’ instrumenten, simulaties en ‘hands-off’ instrumenten. Hands-on instrumenten worden gebruikt voor het beoordelen van prestaties in reële werksituaties. De opdrachten die kandidaten krijgen, dienen zij uit te voeren in complexe werksituaties die niet of nauwelijks van de werkelijkheid verschillen. Bij deze examenvorm zijn de omstandigheden niet volledig beheersbaar. De moeilijkheidsgraad van het examen is derhalve ook niet volledig onder controle. Bij een simulatie moet een kandidaat in een tot op zekere hoogte artificiële situatie demonstreren dat hij of zij vaardig dan wel bekwaam is. In een dergelijke examenvorm worden aspecten uit de beroepspraktijk gemanipuleerd en is dus sprake van een versimpeling van de werkelijkheid. Bij een hands-off examen krijgen deelnemers opdrachten uit te voeren waaruit moet blijken of zij de cognitieve component van een competentie beheersen. Dit kan op papier, maar gebeurt tegenwoordig steeds vaker met hulp van de computer, waardoor opdrachtsituaties levensechter worden. Zie bijvoorbeeld http://www.vandermaesenkoch.nl/ecartoon/nl; http://www.lumc.nl/con/5000/90318051931221/903311018125226, of http://www.ergometrics.org/covt.cfm.

Een veelomvattender indeling komt tot stand door uit te gaan van de aard van de ‘stimulus’ die gegeven wordt bij een assessment en de ‘respons’ van de kandidaat. De stimulus bestaat uit de opdrachten en de verdere relevante informatie, waaronder de instructies die een kandidaat krijgt bij een assessment. Een stimulus bij een assessment kan expliciet of impliciet zijn. Een stimulus is explicieter, naarmate de opdrachten verder tot in detail zijn uitgewerkt, waardoor het gewenste gedrag (en de daaruit voortkomende al dan niet tastbare resultaten) beter en efficiënter uit te lokken is. De respons is uitvoering van de opdrachten door de kandidaat in de vorm van het beantwoorden van de vragen, of het uitvoeren van de opgedragen opdrachten op basis van de verstrekte informatie). Is de kandidaat zich bewust van de aanwezigheid van de assessor(en) bij het assessment, dan zal hij of zij

zich normaal gesproken maximaal inspannen om tot een zo goed mogelijke prestatie te komen. Er is dan sprake van ‘maximal performance’. Is de kandidaat zich niet bewust van de aanwezigheid van de assessor(en), dan wordt ‘typical performance’ waargenomen. De kandidaat vertoont dan het gedrag dat representatief is voor zijn of haar reguliere prestatieniveau.

Door de aard van de prestatie (maximal versus typical performance) af te zetten tegen de aard van de stimulus (impliciet versus expliciet), ontstaat de indeling in tabel 1.

		Stimulus	
		Impliciet	Expliciet
Respons	Maximal performance	(1) Proeve van bekwaamheid	(2) Prestatiebeoordeling: - Hands-off - Simulatie - Hands-on
	Typical performance	(3) Taxatie	(4) ‘Mystery guest’

Tabel 1; Indeling van assessmentvormen op basis van gedrag van de kandidaat en aard van de stimulus

In cel 1 –maximal performance bij een impliciete stimulus- vallen assessments waarbij de kandidaat weet dat hij of zij tijdens het reageren op de stimulus beoordeeld zal worden. De opdracht kan echter niet tot in detail worden uitgewerkt, vanwege de complexe en niet beheersbare context waarin deze moet plaatsvinden. Deze vorm van assessment valt te omschrijven als ‘proeve van bekwaamheid”. Denk bijvoorbeeld aan de LIO-stage. De assessments die in deze categorie vallen zijn authentiek, maar ook relatief inefficiënt als methode, omdat de opdracht voor de kandidaat per definitie niet tot in detail kan worden uitgewerkt. Een ander concreet voorbeeld vinden we in een relatief nieuw onderdeel van het rijexamen waarbij de examinator de kandidaat vraagt om van punt A naar punt B te rijden, zonder daar verder instructies bij te geven.

In cel 2 –maximal performance bij een expliciete stimulus- valt de prestatiebeoordeling (performance assessment), het betreft hier de assessmentvormen die Roelofs en Straetmans (2006) categoriseerden. De kandidaat weet dat hij tijdens het reageren op de stimulus beoordeeld zal worden. In tegenstelling tot de proeve van bekwaamheid kan bij deze categorie door middel van gerichte opdrachten het gewenste gedrag (en de daaruit voorkomende, al dan niet tastbare resultaten) uitgelokt worden. De meest sturing valt daarbij te geven in de hands-off variant; de minste bij de hands-on variant. Van de eerste variant is bijvoorbeeld sprake bij de ‘situational judgment test, waarbij kandidaten aan moeten geven wat de correcte handeling is in een beschreven situatie. Van de laatste variant was bijvoorbeeld sprake bij het traditionele rijexamen van enkele jaren geleden. Hierin was sprake van een examinator die de kandidaat exact vertelde welke opdrachten deze allemaal precies moest doen.

De assessments in cel 3 –typical performance bij een impliciete stimulus- zijn typisch voor in-service beoordelingen van beroepsbeoefenaren, al dan niet in opleiding. Bij deze assessmentvorm –de taxatie- is de kandidaat zich er niet van bewust dat en/of wanneer hij beoordeeld wordt en ook is niet duidelijk wat exact het gewenste gedrag is. Denk hierbij bijvoorbeeld aan systemen voor 360° feedback. Dit type assessment heeft een hoog authentiek gehalte, maar is eveneens slecht beheersbaar. Er wordt veel gevraagd van het geheugen van de beoordelaar(s) en procedures zijn slecht te standaardiseren. Een dergelijke vorm van assessment zou bijvoorbeeld goed ingezet kunnen worden tijdens de LIO-stage.

Bij de assessments in cel 4 -typical performance bij een expliciete stimulus- is wederom sprake van welomschreven opdrachten, maar niet aan de kandidaat, maar aan een persoon, de mystery guest, die tot taak heeft het gewenste gedrag uit te lokken. Ook deze vorm van assessment leent zich goed voor toepassing tijdens stages.

De assessmentvorm in cel 1 is te verkiezen boven de altijd wat kunstmatige methodes in cel 2. De methodes in cel 3 en 4 zijn in onderwijssituaties vooral van belang om zicht te krijgen op de attituditionele aspecten van bekwaamheid. Ze zijn dus bij uitstek geschikt om een beeld te krijgen van de werkelijke beroepshouding van beroepsbeoefenaren (in opleiding).

Sterke en zwakke punten van de verschillende assessmentvormen

Op basis van de voorafgaande informatie is het nu mogelijk om kort en helder weer te geven in hoeverre verschillende assessments naar verwachting aan de voornoemde algemene kwaliteitseisen zullen voldoen. In navolging van Straetmans (2004a) die dit al deed voor prestatiebeoordelingen, kunnen we de informatie uit de verschillende assessmentvormen beoordelen in termen van de accuraatheid, generaliseerbaarheid en extrapoleerbaarheid. De resultaten van deze exercitie zijn te vinden in tabel 3.

	Kwaliteitseis		
Examenvorm	Accuraatheid	Generaliseerbaarheid	Extrapoleerbaarheid
Proeve van bekwaamheid	**--**	**--**	**++**
Hands-on methode	**-**	**-**	**+**
Simulatie	□	□	□
Hands-off methode	**++**	**++**	**--**
Taxatie	**-**	□	**+**
Mystery guest	□	**-**	**+**

--*: relatief zeer zwak;* **-**: relatief zwak; □*: gemiddeld;* **+***: relatief sterk;* **++***: relatief zeer sterk*

Tabel 3; Evaluatie van assessmentvormen voor het vaststellen van competentie in termen van accuraatheid, generaliseerbaarheid en extrapoleerbaarheid

De tabel laat zien dat de proeve van bekwaamheid informatie oplevert die relatief weinig accuraat en generaliseerbaar is, maar dat de extrapoleerbaarheid ervan uitstekend is. De hands-on methode levert informatie die wat accurater en generaliseerbaarder is, maar ook weer minder extrapoleerbaar. De simulatie levert op zijn beurt informatie op die weer wat accurater en generaliseerbaarder is, maar dit gaat wederom te koste van de extrapoleerbaarheid ervan. De hands-off methode levert informatie op die relatief erg accuraat en generaliseerbaar is, maar de extrapoleerbaarheid van die informatie laat zeer te wensen over. De informatie uit een taxatie is relatief weer redelijk extrapoleerbaar, maar niet werkelijk generaliseerbaar en weinig accuraat. De mystery guest, ten slotte, verschaft informatie die eveneens redelijk extrapoleerbaar is, maar de generaliseerbaarheid is beperkt en de accuraatheid laat wel wat te wensen over. Dit leidt tot de onontkoombare conclusie dat er geen assessmentvorm bestaat die in voldoende mate aan alle drie de kwaliteitseisen tegemoetkomt.

Tot diezelfde conclusie kwam Kane (1992) twintig jaar geleden ook al: “Valid assessment of professional competence has proven to be an elusive goal. Objective tests, direct observation of performance, overall ratings of competence and simulations have been tried and found wanting in one way or another. Objective tests are criticized as being unrealistic and therefore invalid. Direct observation tends to be very unreliable and therefore invalid. Simulations and overall ratings of competence share both of these flaws to some extent. Basically, you can’t win!”.

Gelukkig is het mogelijk om dit dilemma op te lossen. En dat is door het inzetten van verschillende assessmentvormen. De kunst is om vanuit onderwijsmeetkundig oogpunt de beste informatie te verzamelen, rekening houdend met randvoorwaarden als de beschikbare tijd voor ontwikkeling en

afname en middelen in de vorm van geld en geschikte cases. Zo'n methodenmix (Straetmans & Sanders, 2001) of competentie-assessment programma (CAP) (Baartman, 2008) bevat een reeks van assessments die in ieder geval al voor een deel gedurende de opleiding worden afgenomen. Welke vorm die assessments aannemen, is mede afhankelijk van de aard van de opleiding.

Besluitvorming bij assessments

Een verantwoorde en inzichtelijke manier om resultaten op verschillende assessments te combineren tot een eindoordeel is ontwikkeld door Straetmans (2004a). Het door hem ontwikkelde Procotol Portfolio Scoring (PPS) laat zien hoe prestaties op uiteenlopende assessments op verantwoorde wijze te integreren zijn tot een eindbeslissing. Het onderliggende principe hierbij is dat een zinvolle integratie van prestaties op verschillende assessmentvormen uitsluitend mogelijk is indien die prestaties via dezelfde systematiek beoordeeld zijn.

Om volgens het PPS-principe te kunnen werken, dient het beoordelingsproces op twee niveaus vormgegeven te worden. Op het hoogste niveau is sprake van beoordelingsaspecten; op het niveau daaronder van indicatoren. Wat het precies behelst om bekwaam in een bepaald vakgebied te zijn, dient om te beginnen uitgewerkt te worden in de formulering van een reeks (globale) beoordelingsaspecten. Deze aspecten operationaliseren generieke categorieën van gewenst gedrag of gewenste kenmerken van ontwikkelde producten die passen binnen het betreffende vakgebied. Niet alle beoordelingsaspecten hoeven noodzakelijkerwijs aan bod te komen binnen ieder assessment dat ontwikkeld wordt. Welke beoordelingsaspecten relevant zijn, is afhankelijk van de opdrachten binnen een assessment en de vorm van dat assessment.

De beoordelingsaspecten worden op hun beurt geoperationaliseerd in een reeks opdrachtspecifieke indicatoren. Dat wil zeggen dat voor elke examenopdracht een specifieke set van prestatiecriteria wordt ontwikkeld op basis van de beschikbare beoordelingsaspecten. Zie voor een concrete uitwerking van dit principe figuur 2.

Figuur 2; Uitwerking van de competentie 'spoedeisende hulp verlenen' in een set beoordelingsaspecten, inclusief de operationalisatie van één beoordelingsaspect in een reeks indicatoren

De PPS-methode houdt in dat er een dossier van assessmentresultaten –een prestatiedossier- wordt aangelegd op basis van de volgende principes:

- Multi shot: bekwaamheden zijn veelomvattende constructen die niet met één prestatie aantoonbaar te maken zijn;
- Multi-method: geen enkele assessmentvorm levert informatie die zowel voldoende accuraat, generaliseerbaar als extrapoleerbaar is. De inzet van meerdere uiteenlopende assesments ligt daarom voor de hand, al is het verstandig om een proeve van bekwaamheid of hands-on performance assessmant als meest gewenste keuze te beschouwen;
- Multi-rater: Prestatiecriteria zijn, zonder dat ze triviaal worden, nooit zodanig te formuleren dat ze objectief scoorbaar zijn. De meest ideale oplossing daarvoor is te werken met meerdere assessoren en het middelen van hun scores. Maar als dit om financiële en organisatorische redenen niet haalbaar mocht zijn, dan is het aan te bevelen om niet alle prestaties van een kandidaat door dezelfde assessor te laten beoordelen;
- Gestandaardiseerde set prestatiecriteria: Elke prestatie die als bewijs voor de aanwezigheid van een bepaalde (beroeps)bekwaamheid deel uitmaakt van het dossier wordt beoordeeld op grond van dezelfde prestatiecriteria. Dat wil zeggen op basis van dezelfde set van beoordelingsaspecten, want de indicatoren die gebruikt worden om de score te bepalen op een bepaald beoordelingsaspect zijn opdrachtspecifiek;
- Gestandaardiseerde beslissingsprocedure: De besluitvorming over het al dan niet aanwezig zijn van de bekwaamheid is volledig vastgelegd in regels en kan in principe automatisch verlopen. Zodra de scores op de beoordelingsaspecten aan een aantal randvoorwaarden voldoen, worden scores horizontaal –per prestatie over de beoordelingsaspecten heen- én verticaal –per beoordelingsaspect over de prestaties heen geëvalueerd door ze te vergelijken met op een onderbouwde[4] manier vastgestelde cesuurscores. Deze procedure is volledig transparant en objectief.

Beoordelingsportfolio SpoedEisende Hulp verlenen
17-01-2012

School	Universiteit Harderwijk
Docent	F. Alsvanouds
Deelnemer	Hansen, E.

Randvoorwaarden:

Minimum aantal bewijsstukken:	4
Minimum aantal bewijsstukken gebaseerd op 'hands-on'-methode:	2
Maximale ouderdom van de bewijsstukken in maanden:	12

bewijsstuk	datum	assessment-methodiek	assessor	beoordelingsaspecten A	B	C	D	E	F	G	H	I	J	K	L	totaal-score	horizontale standaard	resultaat
1	13-12-10	hands-off		4			5		3							12	14,4	-
2	13-01-10	hands-off		5			5		6							16	14,4	+
3	20-04-11	simulatie	J.Pav.	5	5											10	9,8	+
4	28-06-11	hands-on	H. Biem.	5	5	4	4	4								22	23,5	-
5	14-08-11	hands-on	J.Pav.	5		5	6	5	6							27	23,3	+
6	17-11-11	simulatie	A.v.d.Bo.	5	4			5	6							20	19,2	+
7																0	0,0	
8																0	0,0	
9																0	0,0	
10																0	0,0	
11																0	0,0	
12																0	0,0	
	gemiddelde score per aspect			5,0	4,7	4,5	5,0	4,7	6,0									
	grenswaarde			4,8	5,0	4,3	4,8	4,6	4,8									
	resultaat			+	-	+	+	+	+									

Scoren Opslaan

Tabel 3; Een fictief voorbeeld van een prestatiedossier met horizontale en verticale standaarden

Ter nadere toelichting is in tabel 3 een fictief voorbeeld van een prestatiedossier afgebeeld. Voor het onderdeel 'spoedeisende hulp verlenen' bevat dit dossier de resultaten op zes verschillende

[4] Het is aan de betrokken opleiding(en) om te bepalen welk evaluatieresultaat minimaal behaald moet worden voor een positieve uitslag.

assessments. Er zijn voor dit onderdeel zes verschillende beoordelingsaspecten, gelabeld A tot en met F. In het eerste assessment is via een hands-off methode informatie verzameld over prestaties opdrie van de zes beoordelingsaspecten; A, D en F. De score voor deze kandidaat –gebaseerd op het al dan niet voldoen aan een reeks gedragsindicatoren, is voor de drie beoordelingsaspecten respectievelijk 4, 5 en 3. Dit leidt tot een totaalscore van 12 voor dit assessment dat echter een cesuur[5] heeft van 14,4. Daarmee is het resultaat op dit assessment onvoldoende. Verder laat de tabel zien dat ieder beoordelingsaspect minimaal twee maal aan bod gekomen is in de totale reeks assessments. Merk op dat beoordelingsaspect A in alle zes de assessments is meegenomen en beoordelingsaspect C slechts in twee. Hierdoor heeft het beoordelingsaspect A een relatief zwaarder gewicht bij de besluitvorming dan beoordelingsaspect C. Niet alleen voor ieder assessment, maar ook voor ieder beoordelingsaspect is een cesuur vastgelegd. Die cesuur is voor aspect D bijvoorbeeld 4,8. De kandidaat heeft op dit beoordelingsaspect bij de assessments 1, 2, 4 en 5 respectievelijk de scores 5, 5, 4 en 6 behaald. Dat leidt tot een gemiddelde score op dit beoordelingsaspect van 5,0. Deze gemiddelde score ligt boven de cesuur en daarmee scoort de kandidaat op dit beoordelingsaspect voldoende. Verder laat de tabel zien dat er aan een aantal randvoorwaarden voldaan moet zijn. Het minimum aantal bewijsstukken moet 4 zijn; er moet minimaal tweemaal een hands-on methode gebruikt zijn en de bewijsstukken mogen niet ouder zijn dan twaalf maanden. Merk tot slot op dat er bij de simulaties en hands-on methoden gebruik gemaakt is van verschillende assessoren om beoordelaarseffecten enigszins uit te kunnen sluiten. Omdat de kandidaat nog niet op alle beoordelingsaspecten boven de cesuur heeft gescoord en er ook een aantal assessments met onvoldoende resultaat zijn afgesloten, kan nog niet geconcludeerd worden dat de betreffende kandidaat al in staat is op een minimaal gewenst niveau spoedeisende hulp te verlenen.

Een uitgebreide toelichting op het PPS-principe en de bijbehorende werkwijze is te vinden in Straetmans (2004b). Overigens is ook aannemelijk gemaakt dat het PPS-principe naadloos kan aansluiten bij- en geïntegreerd kan worden binnen leeromgevingen waar competent handelen centraal staat (Sluijsmans, Straetmans & Van Merriënboer, 2008).

Landelijke toetsen als onderdeel van beroepsgerichte examens

Omdat de staatssecretaris in zijn strategische agenda de voorkeur uitspreekt voor “landelijke toetsing van één of meer kernvakken”, ligt het voor de hand in deze notitie in concreto in te gaan op de voor- en nadelen van een dergelijke examenvorm. Deze examenvorm beschrijven we hier met de term ‘gestandaardiseerde summatieve toets’. De term summatief geeft aan dat de toets gebruikt wordt aan het einde van een onderwijsleerproces of een onderwijsperiode om vast te stellen of deelnemers aan het onderwijs hier voldoende van hebben opgestoken en/of ter evaluatie van het genoten onderwijs. We gebruiken de term gestandaardiseerd om duidelijk te maken dat het afnameproces, de scoring en beoordeling voor iedere kandidaat bij een specifieke afname identiek is. Dat hoort in het kader van een gelijke uitgangssituatie voor kandidaten zo te zijn. Bij de ‘klassieke’ gestandaardiseerde toets zijn bovendien ook de opgaven voor iedere kandidaat identiek, zodat een zuivere vergelijking tussen kandidaten mogelijk is. Dit is bijvoorbeeld het geval bij de afnames van de centrale examens havo en vwo, waar binnen ieder tijdvak krijgen alle kandidaten dezelfde toets. Tegenwoordig maken technologische en onderwijsmeetkundige ontwikkelingen (zie bijvoorbeeld Béguin, 2001; Eggen, 2004 en Verschoor, 2007) het echter ook mogelijk om de gelijkwaardigheid van toetsen te garanderen die niet identiek van samenstelling zijn voor alle kandidaten.

De algemene kwaliteitseisen waar een gestandaardiseerde summatieve toets aan moet voldoen zijn al eerder in deze notitie in de paragraaf ‘Examenkwaliteit begrippenkader en eisen’ (p.2) besproken. Vertaald naar het begrippenkader van Straetmans (2004a) geldt voor dit instrument dat de accuraatheid van de resulterende informatie goed in orde is. De generaliseerbaarheid ervan –in de zin dat op basis van de voorgelegde opgaven conclusies getrokken kunnen worden over beheersing van

[5] Bijvoorbeeld vastgesteld via de methode van Angoff (1971)

het gehele domein aan mogelijke opgaven- is eveneens in orde. De extrapoleerbaarheid is echter zeer beperkt. Net als voor iedere andere examenvorm geldt dan ook voor de gestandaardiseerde summatieve toets dat deze niet in te zetten valt als enige examenvorm om te bepalen of iemand gekwalificeerd is als beginnend beroepsbeoefenaar op hbo-niveau. Als onderdeel van een beroepsgericht examen kan dit type instrument echter uitstekend geschikt zijn.

Het eerste belangrijke voordeel van een dergelijke toets is dat deze bij onderdelen van curricula waar sprake is van een algemeen –tenminste over verschillende overeenkomstige opleidingen heen- geaccepteerde kennisbasis ('body of knowledge') het middel bij uitstek is om op een verantwoorde en transparante manier expliciet vast te stellen of de beoogde kennis in voldoende mate verworven is en of het onderwijs het beoogde doel wat kennisverwerving betreft heeft gerealiseerd. Hierover valt via de toets ook concreet verantwoording af te leggen. Het expliciet vaststellen of de benodigde kennis op een bepaald minimumniveau beheerst wordt is niet of in veel minder efficiënte mate mogelijk via een (reeks) assessment(s). De door de betrokken opleidingen geaccepteerde kennisbasis vormt een degelijke en transparante basis voor het ontwikkelen en onderhouden van het opgavendomein waar uit geput kan worden om de gemeenschappelijk gestandaardiseerde summatieve toetsen samen te stellen.

De resultaten op dergelijke gemeenschappelijk ontwikkelde gestandaardiseerde summatieve toetsen zijn op uiteenlopende wijzen te aggregeren. Zo kan een overzicht gegeven worden van het gemiddelde kennisniveau van studenten in onderscheiden opleidingen, waardoor de kwaliteit van de opleidingen -waar het het primaire onderwijsproces betreft- transparant gemaakt worden. Aggregatie leidt in algemene zin tot stuurinformatie die instellingen kunnen gebruiken om gericht maatregelen te nemen om ongewenste ontwikkelingen tegen te gaan, maar ook om te laten zien dat aan een acceptabel minimumniveau voldaan kan worden. Indien een goede methode van normhandhaving wordt toegepast, is het bovendien ook mogelijk het kennisniveau van verschillende cohorten studenten te vergelijken om positieve of negatieve trends te signaleren.

Een tweede belangrijk voordeel is dat het ontwikkelen van gemeenschappelijke toetsen wat kosten van opgavenontwikkeling en cesuurbepaling betreft aanzienlijk goedkoper is dan het decentraal ontwikkelen van opgaven. Naarmate meer opleidingen betrokken zijn, wordt de benodigde inspanning om voldoende opgaven te ontwikkelen per opleiding lager. Tevens kan de efficiëntie toenemen, omdat minder mensen betrokken hoeven te zijn bij de normering.

Een derde voordeel is dat naarmate meer studenten deelnemen aan de toets, het beter mogelijk wordt om psychometrische kwaliteitscontrole plaats te laten vinden. Op basis hiervan valt dan beter aan te tonen dat het instrument betrouwbaar en valide is en dat de gegevens die het oplevert voor de beoogde doelstellingen gebruikt kunnen worden. De maatregelen die men zich moet getroosten om deze psychometrische kwaliteitscontrole plaats te laten vinden brengen echter de nodige kosten met zich mee. Datzelfde geldt voor het verantwoordingsproces dat duidelijk moet maken dat het instrument aan de vigerende kwaliteitseisen (zie bijvoorbeeld RCEC, 2011) voldoet . Deze kosten hebben dan echter weer als voordeel dat de kwaliteit van de toetsen concreet aan te tonen valt, waardoor de waarde van het diploma wat betreft het kennisaspect geborgd wordt en blijft.

Een vierde voordeel is dat de correctie van de toetsen geautomatiseerd kan gebeuren, indien er gebruik gemaakt wordt van gesloten opgaven. Bij toetsing op papier kunnen kandidaten hun antwoorden weergeven op optisch leesbare antwoordbladen. En bij toetsing met behulp van de computer kan de correctie zelfs direct plaatsvinden, zodat het resultaat direct na de toets kan worden gerapporteerd. Terzijde zij hier opgemerkt dat de afkeer van gesloten opgaven bij veel betrokkenen niet gestaafd kan worden door wetenschappelijk onderzoek (zie bijvoorbeeld Rodriguez, 2003). Het grote voordeel van de inzet van gesloten opgaven is dat zij in meettechnisch opzicht leiden tot een grotere inhoudsvaliditeit en een hogere meetnauwkeurigheid dan open opgaven. De simpele oorzaak daarvan is dat zij sneller te beantwoorden zijn en dat er daarom meer gesloten dan open opgaven kunnen worden afgenomen in een concrete tijdsspanne. Een economisch voordeel is dat zij

goedkoper te produceren zijn en er bij de correctie geen tussenkomst van menselijke beoordelaars vereist is. Voor automatische scoring van zelfs simpele open vragen is veel inspanning vereist. Zowel voor digitale als papieren examens geldt dat de ontwikkeling van open vragen duurder is en meestal niet noodzakelijk.

Tevens is bij de inzet van gemeenschappelijk ontwikkelde summatieve toetsen per definitie sprake van externe validering van de kwaliteit ervan, omdat er in ieder geval meer dan één opleiding bij betrokken is. De mate van robuustheid van die externe validering is afhankelijk van de opzet van het ontwikkelingstraject. De minst robuuste vorm is die waarbij verschillende instellingen gebruik maken van een in gezamenlijkheid ontwikkeld instrument met een in gezamenlijkheid vastgestelde cesuur. Een zeer robuuste vorm van externe validering van de kwaliteit van dergelijke toetsen in de praktijk in Nederland is die bij de centrale examens voor havo en vwo waarin sprake is van een sterke (maar niet strikte) formele scheiding tussen opleidingen en toetsing[6]. De toetsen die onderdeel uitmaken van de centrale examens worden door een externe onafhankelijke instantie –het Cito- ontwikkeld onder strikte geheimhouding in opdracht van een daartoe speciaal door de overheid opgericht orgaan; het College voor Examens (CvE). De kerntaken van het CvE zijn de beschrijving van examenstof, het vaststellen van examenopgaven, en het normeren van de examens. Er worden uitgebreide maatregelen genomen om te garanderen dat prestaties van kandidaten over de jaren heen gelijkwaardig beoordeeld worden. Uiteraard wordt er ook uitgebreid gerapporteerd over verschillende aspecten van het examenproces. Zie bijvoorbeeld het “Verslag van de examencampagne 2011” (Albers & Erens, 2011)

Een ander voorbeeld van een zeer robuuste vorm van externe validering vinden we bij de entreetoetsen voor de pabo die door Cito ontwikkeld zijn. Over de ontwikkeling van de toets voor Rekenen is in expliciete vorm verantwoording afgelegd. De Commissie Testaangelegenheden (COTAN), een orgaan van het Nederlands Instituut van psychologen heeft het instrument met behulp van hun beoordelingssysteem (COTAN, 2010) op basis van de verantwoording door twee onafhankelijke testexperts laten beoordelen. En op basis daarvan heeft de opdrachtgever (HBO-raad) geconstateerd dat het instrument aan zijn doelstellingen voldeed. De entreetoetsen voor de pabo laten overigens ook zien dat er geen inhoudelijke verschil hoeven te bestaan tussen instrumenten die van bovenaf zijn opgelegd, of instrumenten die in gezamenlijkheid door instellingen worden ontwikkeld. De toetsen voor Wereldoriëntatie zijn in opdracht van een vereniging van eigenaren van Pabo’s ontwikkeld, maar volgen dezelfde ontwikkelings- en onderhoudsprocedures als de toetsen voor Rekenen en Nederlands die in opdracht van de HBO-raad zijn ontwikkeld. Overigens heeft het bestuur van de HBO-raad en recentelijk voor gekozen om ook de toetsen voor Wereldoriëntatie landelijk uit te gaan rollen.

Het beoordelingssysteem van de COTAN dat eerder in deze tekst genoemd werd, heeft als nadeel dat het ontwikkeld is voor het beoordelen van psychologische tests en niet voor examens en studietoetsen. Daarom heeft het Research Center voor Examinering en Certificering recent het initiatief genomen voor de ontwikkeling van een beoordelingssysteem dat wel gebaseerd is op het COTAN-systeem, maar specifiek bedoeld is voor examens en studietoetsen. Inmiddels bestaat er een eerste concept (RCEC, 2011), waarvan het de bedoeling is dat het met verschillende stakeholders verder ontwikkeld zal worden. Tabel 4 bevat een globaal overzicht van de zes kwaliteitseisen die in dit beoordelingssysteem aan de orde komen.

[6] Merk overigens op dat de externe validering nog sterker zou zijn, indien de eerste correctie niet zou plaatsvinden door de docenten van de school van herkomst van de kandidaten, maar door docenten van andere scholen.

Doel en gebruik Bij deze eis komt de vraag aan de orde of de ontwikkelaar het doel, het gebruiksdoel en de doelgroep(en) voldoende duidelijk heeft omschreven.
Kwaliteit toets- of examenmateriaal Bij deze eis is sprake van drie basisvragen. In de eerste plaats in hoeverre de opdrachten gestandaardiseerd zijn. In de tweede plaats in hoeverre de scoring van de opdrachten objectief verloopt. En in de derde plaats of de opdrachten vrij zijn van racistische, etnocentrische, seksistische en voor bepaalde bevolkingsgroepen kwetsende inhoud.
Representativiteit Er worden twee basisvragen gesteld bij deze eis. De eerste heeft betrekking op de mate waarin inhoud, duur, omvang en het soort opdrachten in overeenstemming zijn met de specificaties van de toetsmatrijs, het examenmodel of het examenplan. De tweede heeft betrekking op de mate waarin de moeilijkheidsgraad van de opdrachten is afgestemd op de beoogde doelgroep.
Betrouwbaarheid De drie vragen bij deze eis luiden: - Zijn of worden betrouwbaarheidsgegevens verstrekt? - Zijn of worden de betrouwbaarheidsgegevens correct berekend? - Zijn de betrouwbaarheidsgegevens voldoende gezien de beslissingen die met de toets of het examen genomen worden?
Standaardbepaling en normhandhaving Bij deze eis komt de vraag aan de orde hoe de normen van toetsen of examens bepaald zijn en hoe de normen van vergelijkbare of parallelle toetsen of examens gehandhaafd worden.
Afname en beveiliging De twee basisvragen waarvan hier sprake is, hebben respectievelijk betrekking op de mate waarin informatie voor de toets- of examenleider beschikbaar is over de wijze waarop de toets of het examen dient te verlopen en op de mate waarin het voor onbevoegden mogelijk is om toegang te krijgen tot het toets- of examenmateriaal, de toets- of examenresultaten en de toets of het examen zelf.

Tabel 3; Overzicht van de zes criteria uit het RCEC (2011) beoordelingssysteem voor de kwaliteit van toetsen en examens

De gemeenschappelijk ontwikkelde gestandaardiseerde summatieve toets is zonder meer geschikt als onderdeel van een beroepsgerichte examenmix. Namelijk voor die delen van het curriculum waar sprake is van een concrete 'body of knowledge' van substantiële omvang. Het probleem hierbij is dat de kosten van het ontwikkelen en onderhouden van een dergelijke toets snel stijgen naarmate de inzet ervan flexibeler moet zijn. Is bijvoorbeeld de wens om hoogfrequent of zelfs doorlopend te kunnen toetsen, dan is de enige mogelijkheid het werken met een uitgebreide opgavenbank waaruit zonder menselijke tussenkomst (adaptieve) toetsen kunnen worden samengesteld. De kosten hiervan liggen echter aanzienlijk hoger dan wanneer sprake is van enkele afnamemomenten per jaar.

Hoewel gemeenschappelijke gestandaardiseerde summatieve toetsen prima in te zetten zijn als onderdeel van beroepsgerichte examens, valt een ongebreidelde inzet ervan te ontraden. In de eerste plaats is het ontwikkelen en onderhouden van dergelijke instrumenten een kostbare zaak. Niet zozeer vanwege de inzet van menskracht, maar vanwege de kosten die gemoeid zijn met het realiseren van voldoende kwaliteit en het onderhoud ervan. Echter, ook indien de kosten geen beperkende factor zouden zijn is een te grote nadruk op deze vorm van toetsen onverstandig. Het is namelijk een bekend gegeven dat de inhoud van examens het leergedrag en de motivatie van kandidaten sterk stuurt. Dit fenomeen staat bekend onder de naam "backwash" of "washback' effect en er wordt in de onderwijsliteratuur uitgebreid naar verwezen. Recent heeft Joosten-ten Brinke (2011) dat bijvoorbeeld nog gedaan voor de toetsing in beroepsgerichte opleidingen. Anders gezegd, studenten leren alleen maar die onderdelen van het curriculum waarvan duidelijk is dat ze ook werkelijk geëxamineerd worden. Summatieve toetsing van kennis richt zich per definitie op een beperkt aantal aspecten van beroepsbekwaamheid. Een te eenzijdige benadrukking van kennis en vaardigheden in de examinering zou kunnen leiden tot de uitstroom van onvoldoende gekwalificeerde studenten uit het hbo; studenten met voldoende kennis, maar onvoldoende beroepshouding en bekwaamheid. Het is dus bij beroepsgerichte examinering zowel vanuit strategisch als kostentechnisch oogpunt van belang een goede balans te vinden tussen de hoeveelheid assessments en de hoeveelheid toetsen in het examenprogramma.

Een nadeel van het werken met summatieve toetsen is verder dat over afnameperiodes heen niet dezelfde toets kan worden ingezet, omdat de inhoud ervan na de afname bekend kan raken. Het is dus nodig om de inhoud van de toetsen over afnames heen in ieder geval deels te vernieuwen, of maatregelen te nemen om te voorkomen dat eenmaal gebruikte opgaven eenvoudig bij nieuwe kandidaten bekend kunnen raken. Bovendien moeten alle toetsen dezelfde moeilijkheidsgraad en meeteigenschappen hebben, of moeten er maatregelen genomen worden om te voorkomen dat verschillen in meeteigenschappen en moeilijkheidsgraad leiden tot ongelijke behandeling van kandidaten met dezelfde vaardigheid over afnames heen. Verschillende ontwikkelingen op technologisch en onderwijsmeetkundig gebied maken het echter steeds beter mogelijk om dit op een kostenefficiënte wijze te realiseren. Voor een uitgebreid en gedetailleerd overzicht daarvan wordt verwezen naar de vierde editie van Educational Measurement (American Council on Education, 2006).

Het algemene principe bij voornoemde ontwikkelingen is dat er een voldoende grote verzameling opgaven wordt ontwikkeld waarvan de meeteigenschappen op voorhand worden onderzocht. De opgaven die de gewenste meeteigenschappen bezitten, gaan deel uitmaken van een zogenoemde itembank waaruit via automatische procedures toetsen kunnen worden samengesteld met overeenkomstige meeteigenschappen. Een mooi voorbeeld hiervan in het hbo vinden we in de toetsen voor Nederlands, rekenen en wereldoriëntatie die ontwikkeld zijn voor het geven van bindende studie-adviezen aan pabo-studenten tijdens de propedeuse (Straetmans & Eggen, 2011).

Het feit dat het noodzakelijk is om opgaven te hergebruiken in het kader van normhandhaving brengt het gevaar van fraude en handel in oude opgaven met zich mee. Ook voor dit probleem zijn op basis van ontwikkelingen in de psychometrie en technologie een reeks van “checks and balances” in te zetten die ervoor zorgen dat dit probleem onder beheersing is en blijft. Zo zijn er methoden die ervoor zorgen dat de opgaven in een bank gelijkelijk in toetsen worden opgenomen; zie bijvoorbeeld Stocking & Lewis (1995). Ook is het mogelijk door het toepassen van statistische analysetechnieken te controleren of de moeilijkheid van opgaven met de tijd wijzigt.. En verder is het ook mogelijk om te checken of kandidaten antwoordenreeksen produceren op opgaven die duiden op fraude. Professionele bedrijven en instituten die gestandaardiseerde toetsen ontwikkelen zetten dergelijke procedures in binnen hun toetsprojecten. Het bedrijf Caveon (www.caveon.com) is zelfs volledig gespecialiseerd in het voorkomen en ontdekken van fraude bij toetsing.

Combineren van resultaten op assessments en toetsen

In het begin van deze notitie is de term examen gedefinieerd als een onderzoek waarbinnen een kandidaat over een samenhangend geheel van leergebieden een prestatie moet leveren op grond waarvan met inachtneming van concrete prestatie-eisen en beslisregels een diploma of certificaat kan worden uitgereikt. De kwaliteit van het examen is dus niet uitsluitend afhankelijk van de kwaliteit van de verschillende onderdelen, maar ook van de gehanteerde beslisregel. Sanders (2011) onderscheidt drie ‘uitslagregels’:

- Conjunctieve uitslagregels – een kandidaat moet op alle onderdelen voldoende presteren;
- Complementaire uitslagregels - een kandidaat hoeft niet op alle onderdelen voldoende te presteren;
- Compensatorische uitslagregels - onvoldoende prestaties op sommige onderdelen kunnen gecompenseerd worden met voldoende prestaties op andere onderdelen.

In de paragrafen die betrekking hadden op assessments is duidelijk geworden dat het mogelijk is de resultaten van kandidaten op assessments op een verantwoorde, volledig transparante en voldoende objectieve manier tot een eindoordeel te combineren. En daarna is duidelijk gemaakt dat landelijke gestandaardiseerde summatieve toetsen goed deel uit kunnen maken van beroepsgerichte examens.

Door tot slot één van de boven beschreven uitslagregels te hanteren is het mogelijk de resultaten op de assessments en de resultaten op de toetsen met elkaar te integreren tot een volkomen transparant en inzichtelijk eindoordeel. Dat eindoordeel kan luiden dat een kandidaat geslaagd is, maar ook dat er op een onderdeel of onderdelen herkansing nodig is.

Kwaliteitsborging van beroepsgerichte examens

Tot nu toe heeft deze notitie zich geconcentreerd op de borging van de kwaliteit van examens en examenonderdelen op basis van eisen die gelden voor het examen(onderdeel) zélf. Er zijn echter andere, indirecte manieren om de kwaliteit van examens of examenonderdelen te borgen. Dat hoeft niet uitsluitend te geschieden door de intrinsieke kwaliteit van het examen(onderdeel) te controleren. Borging van de kwaliteit van een product is ook mogelijk door te controleren op:

- Ambachtelijke vaardigheid: de deskundigheid van de personen die het product ontwikkelen;
- Proceskwaliteit; de kwaliteit van de processen die leiden tot de totstandkoming van het product;
- Klanttevredenheid: de tevredenheid van de gebruikers van het product.

Dergelijke methoden kunnen gebruikt worden ter aanvulling of verdere versterking van de kwaliteitsborging, maar ook als controle van de intrinsieke kwaliteit van een product te complex en/of te kostbaar is.

Ook moet hier opgemerkt worden dat het examen als product zelf ook weer deel uitmaakt van een veelomvattender proces: de examenketen. Een proces dat start met de registratie van kandidaten en eindigt met het uitreiken van diploma's of certificaten aan geslaagde kandidaten en het herregistreren van gezakte kandidaten. De voornoemde eisen kunnen ook gesteld worden aan dit bredere proces. Vanuit de perspectieven van ambachtelijke vaardigheid, proceskwaliteit en klanttevredenheid zijn binnen dit kader daarom de volgende mogelijkheden voor kwaliteitsborging van examens te onderscheiden:

- Controle van de deskundigheid van de ontwikkelaars van toetsen en assessments; maar ook
- Controle van de deskundigheid van andere personen die betrokken zijn bij het examenproces, zogeheten examenfunctionarissen. Naast de ontwikkelaars van toetsen en assessments, zijn dat bijvoorbeeld leden van examencommissies en (praktijk)assessoren, maar ook medewerkers van examenbureaus en zelfs surveillanten;
- Controle van de kwaliteit van het ontwikkelproces van toetsen en assessments; maar ook
- Controle van de kwaliteit van de gehele examenketen, waaronder de examenlogistiek, kandidaatadministratie en –registratie, examenomstandigheden, klachtenregistratie, resultaatopslag en rapportage;
- Controle van de tevredenheid van kandidaten over het examen; maar ook
- Controle van de tevredenheid van andere stakeholders, zoals het beroepenveld, opleidingen, ouders, en de overheid.

In de examenpraktijk zijn de voornoemde activiteiten impliciet dan wel expliciet aan te treffen in zeer uiteenlopende combinaties. De meest volwassen en robuuste vormen van kwaliteitsborging zijn die waarbij sprake is van concrete controle op al deze aspecten op basis van harde criteria. In zulke gevallen zijn alle controle-activiteiten expliciet beschreven en is borging en onderhoud ervan opgenomen in een PDCA of Deming-cyclus. Dit laatste wil zeggen dat systematisch gewerkt wordt aan verbetering van de kwaliteit door allereerst een plan (Plan) op te stellen voor verbetering; dit te laten volgen door het uitvoeren van de geplande verbeteringen (Do); te controleren of de geplande verbetering het beoogde effect heeft (Check); en tot slot eventueel zaken bij te stellen op basis van die controle (Act).

Borging intrinsieke kwaliteit
Voor wat betreft de intrinsieke kwaliteit is sprake van een concreet en uitgewerkt examenplan dat aangeeft welke instrumenten ingezet zullen worden binnen het examen. De inhoud van dit examenplan komt voort uit de onderwijskundige visie van een instelling of opleiding. Verder houdt het examenplan rekening met de eisen die er ten aanzien van de examinering vanuit een toezichthoudende instantie of het bevoegd gezag gesteld worden. Ook maakt het duidelijk hoe het voortkomt uit het competentieprofiel of de eindtermen waar de opleiding zich op richt. Uiteraard zorgt het er ook voor dat ieder examenonderdeel aan de vigerende eisen voldoet. In combinatie met de examenspecificaties leidt het examenplan tot een concreet examenprogramma: het geheel van geplande assessments en toetsen dat het bewijs levert dat nodig is om een onderbouwde beslissing te nemen over de vraag of kandidaten voldoende hebben opgestoken van een opleiding.

In het examenplan wordt ten minste besproken hoe tot een goede mix van examenonderdelen gekomen is, hoe deze ontwikkeld worden of waar deze aangeschaft zijn, aan welke toetstechnische eisen ieder examenonderdeel moet voldoen, onder welke condities de examenonderdelen moeten worden afgenomen en op welke wijze de examenonderdelen leiden tot diplomering.

De documenten die in dit kader ontwikkeld worden kunnen ter beoordeling aan een visiterende instantie worden voorgelegd. De borging zou robuuster worden, indien een instelling de kwaliteit van het gehele examenprogramma door een onafhankelijke instantie op basis van een daartoe geschikt beoordelingssysteem. Naast het al genoemde systeem van het RCEC bestaat een reeks andere generieke systemen. Voor een overzicht daarvan wordt verwezen naar Roorda (2008). Een nog hoger niveau van borging zou tot slot bereikt kunnen worden wanneer ook op dit vlak de eis gesteld wordt dat deze onafhankelijke organisatie geaccrediteerd is door de RvA. In de Verenigde Staten wordt zelfs bij sommige examens zelfs de eis gesteld dat de instantie die de kwaliteit ervan beoordeeld geaccrediteerd is (Wild & Knap 2008).

Borging deskundigheid examenfunctionarissen
Ten aanzien van de deskundigheid van examenfunctionarissen kunnen eveneens concrete eisen gesteld worden. Voor voorbeelden hiervan zie de Nederlandse Vereniging voor Examens (NVE, 2011a; 2011b; 2011c; 2011d; 2011e). De kwaliteitsborging van dit aspect is beter naarmate er sterkere maatregelen getroffen zijn om te borgen dat functionarissen aantoonbaar aan de eisen voldoen. Dat kan variëren van het laten volgen van relevante cursussen of opleidingen tot het formeel laten certificeren (conform het ISO 17024 label voor persoonscertificatie) van betrokken personen. In de meest extreme vorm kan er hier bovendien nog sprake zijn van borging van kwaliteit van de certificerende instantie doordat deze geaccrediteerd is als certificerende instantie door de RvA.

Borging proceskwaliteit
Kwaliteitsborging van de gehele examenketen is beter mogelijk naarmate de keten explicieter beschreven is: de inrichting van de werkorganisatie voor de examens, alle examenonderdelen en de wijze waarop een einduitslag bepaald wordt, inclusief de naleving van de wettelijke voorschriften ten aanzien van het examen. Een voorbeeld van een voorzet voor een concrete en transparante inrichting van onderdelen van een examenproces vinden we in een recente notitie van de Hbo-raad (2011) waarin concreet gemaakt wordt hoe de rol van examencommissies op een goede wijze kan worden ingevuld na inwerkingtreding van de gewijzigde Wet op het Hoger onderwijs.

Analoog aan de borging van intrinsieke kwaliteit zou de borging van de proceskwaliteit aan robuustheid winnen indien deze door een onafhankelijke organisatie beoordeeld zou kunnen worden op basis van een daartoe geschikt beoordelingssysteem. Een bekende vorm van procesborging is die volgens het ISO 9001-systeem. Minder bekend is waarschijnlijk dat er inmiddels ook een ISO-label ontwikkeld is specifiek voor examenprocessen.

Borging betrokkenheid stakeholders
Voor het borgen van de tevredenheid van de stakeholders is een reeks van maatregelen mogelijk. Kandidaten zelf kan gevraagd worden na afloop van een examen(onderdeel) hoe zij het

examen(onderdeel) ervaren hebben. Daarbij moet er zorg voor worden gedragen dat de tevredenheid van kandidaten niet beïnvloed wordt door het feit dat zij geslaagd of gezakt zijn. Een geheel andere vorm van het borgen van de tevredenheid van stakeholders is te vinden in recente ontwikkelingen in het mbo. Daar is het ontvangende bedrijfsleven sinds enige tijd formeler betrokken bij de totstandkoming van het beroepsgerichte deel van de examens dan voorheen. In de zogeheten examenprofielen die voor alle sectoren zijn ontwikkeld en vastgesteld is de betrokkenheid van het beroepenveld concreet geborgd, zowel op het niveau van de sectoren als in de regio's. De verwachting is dat hiermee de tevredenheid van het beroepenveld over de kwaliteit van het examen zal toenemen.

Net als bij beroepsgerichte examens zelf is het ook op het vlak van kwaliteitsborging de vraag welke mix van methoden optimaal is om de kwaliteit van de examinering bij een specifieke opleiding te borgen. En net als bij die examens zelf is het antwoord daarop dat dit afhangt van een reeks van factoren, waaronder de aard van de specifieke opleiding, de context waarbinnen examinering plaatsvindt, de consequenties van misclassificaties en vooral ook van de kosten die de verschillende activiteiten met zich meebrengen.

Zo ligt bij havo- en vwo-opleidingen, vanwege de aard van de opleidingen de nadruk van oudsher – ten minste voor wat het gedeelte betreft dat niet aan de scholen zelf is voorbehouden- op de intrinsieke kwaliteit van de examens. Dat is ook logisch, omdat het centrale gedeelte van het examen bestaat uit een reeks gestandaardiseerde summatieve kennistoetsen[7]. En van dergelijke instrumenten is de intrinsieke kwaliteit goed aan te tonen. Dat neemt niet weg dat ervoor gezorgd wordt dat ook de ontwikkelaars aantoonbaar deskundig zijn. Verder is natuurlijk het ontwikkelproces gestandaardiseerd en voorzien van vele controlepunten. En tot slot wordt er ook periodiek onderzocht hoe het gesteld is met de tevredenheid van verschillende stakeholders. Binnen een beroepsgerichte context is daarentegen de nadruk op de kenniscomponent van het curriculum minder en zal deze bovendien aan het begin van de opleiding liggen. Naarmate de opleiding vordert, zal de nadruk steeds meer op beroepsbekwaamheid komen te liggen. Bovendien zal de opleiding een steeds individueler karakter krijgen, naarmate deze verder vordert. Een volledige focus op de intrinsieke kwaliteit van de assessments in het examenprogramma ligt dan minder voor de hand, vanwege het grote aantal assessments dat dan beoordeeld zou moeten worden. Het ligt dan meer voor de hand om te checken of de ontwikkelaars aantoonbaar deskundig zijn en of assessoren een vorm van formele training hebben gehad. Daarbij zou natuurlijk altijd sprake moeten zijn van een steekproefsgewijze controle van de kwaliteit van de assessments.

Tot slot wordt hier nog opgemerkt dat kwaliteitsborging van de examinering binnen een instelling via een integratie van al de voornoemde facetten zou kunnen culmineren in de formele accreditatie van de instelling als exameninstelling door de RvA.

Relevante ontwikkelingen in andere onderwijssectoren in Nederland

Een beschouwing van de situatie in andere onderwijssectoren in Nederland maakt duidelijk dat de verschillen wat betreft borging van de kwaliteit van het examen aan het afnemen zijn. Door de overheid is een ontwikkeling ingezet die tot doel heeft meer grip te krijgen op de kwaliteit van het onderwijs op relevant geachte leergebieden. Conceptueel gezien groeien de overige onderwijssectoren toe naar het model waarvan sprake is in het voorgezet onderwijs. Het diploma wordt daar uitgereikt op basis van de resultaten van kandidaten op de centrale examens en schoolexamens tezamen.

[7] In de betekenis van Bloom (1975)

Hoewel in het primair onderwijs formeel geen sprake is van examinering, zal de situatie daar naar verwachting op korte termijn desalniettemin vergelijkbaar zijn. De overheid heeft immers het plan om deelname aan de Eindtoets Basisonderwijs van het Cito verplicht te maken voor alle scholen. Daarmee worden –net als in het voorgezet onderwijs twee vormen van informatie relevant voor het bepalen van de verdere leerloopbaan van leerlingen: resultaten op een landelijke summatieve toets en het oordeel van de school zélf. Hiermee zet de overheid ook duidelijk in op de twee functies van de summatieve gestandaardiseerde toets: niet alleen bepalen wat leerlingen hebben opgestoken van het onderwijs op concrete leergebieden, maar ook evaluatie van dat onderwijs zélf.

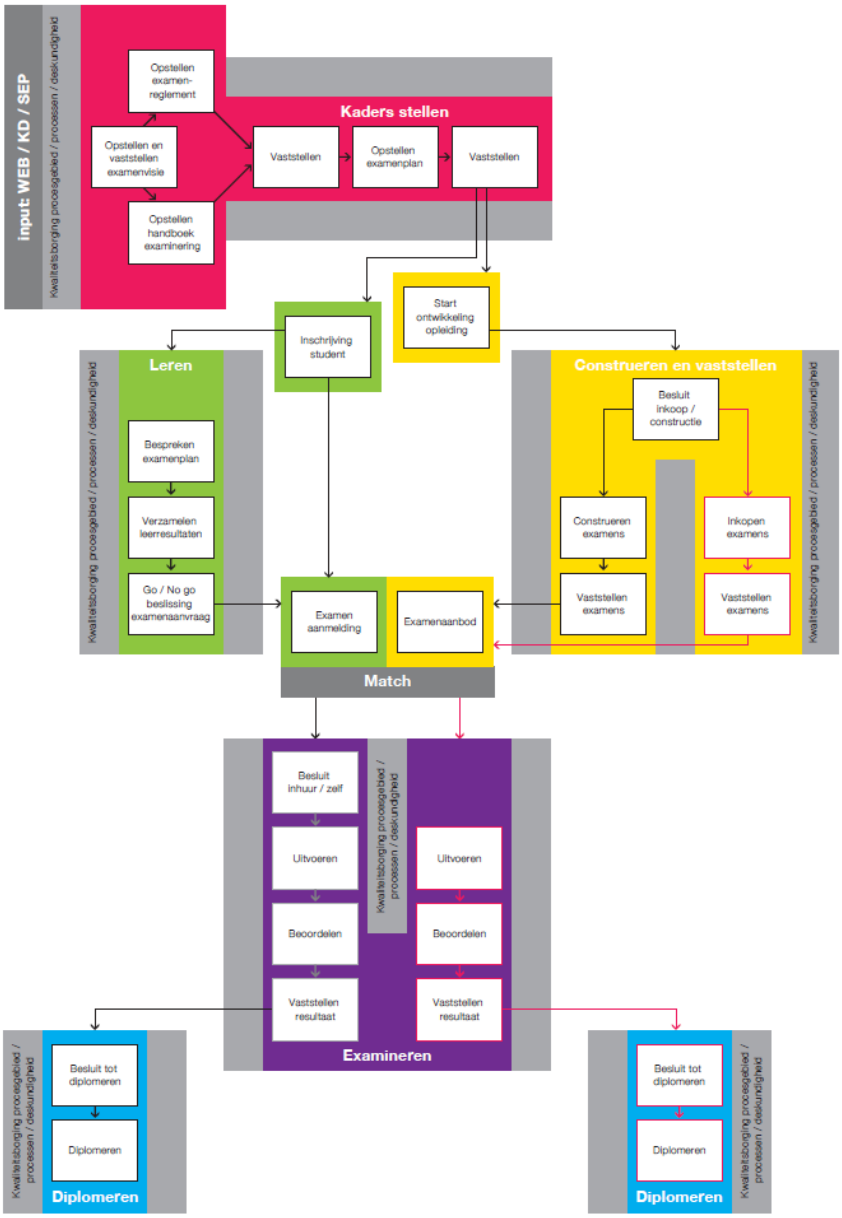
De situatie in het VO is al jaren redelijk stabiel. Een belangrijke ontwikkeling die plaatsvindt is de overgang van klassieke papieren examens naar digitale examens In het mbo daarentegen is sprake van grote veranderingen. Er is inmiddels sprake van een volledig competentiegerichte kwalificatiestructuur. De opleidingen in het mbo zijn of worden dan ook opnieuw ontworpen. De examinering zal zich daarbij met ingang van het schooljaar 2012-2013 dienen te baseren op basis van de herziene standaarden uit het toezichtskader van de Onderwijsinspectie (Inspectie van het Onderwijs, 2011). Het waarderingskader dat hieraan voorafging (Inspectie van het Onderwijs, 20xx) had de vier elementen van kwaliteitsborging in zich die eerder in deze tekst aan de orde kwamen. In het nieuwe waarderingskader zitten deze elementen meer verborgen, maar ze komen nog steeds aan de orde.

De situatie in het mbo is relevant om in beschouwing te nemen, omdat zij voor een deel met dezelfde problematiek te kampen hadden. Om de kwaliteit van de examinering in het mbo op een hoger plan te brengen zijn door de sector zelf en door OCenW verschillende maatregelen genomen. Een majeure operatie op het moment is de introductie van de Centraal Ontwikkelde examens voor taal en rekenen (COE's) in het mbo. Het einddoel van deze operatie is de implementatie van een gecomputeriseerd systeem dat het mogelijk maakt aan kandidaten om zo flexibel mogelijk deze examens te maken. Het streven hierbij is om gebruik te gaan maken van de technieken die eerder in deze notitie beschreven zijn om een zo flexibel mogelijke afname te realiseren.

Verder hebben de MBO raad, COLO (nu opgegaan in de stichting Samenwerking Beroepsonderwijs Bedrijfsleven), AOC raad, NRTO (voorheen PAEPON), VNO-NCW en MKB-Nederland enige jaren geleden de handen ineen geslagen in het Project Examenprofielen. Doel van dit project was om het vertrouwen van het bedrijfsleven te vergroten, om bij te dragen aan landelijke standaardisatie binnen de onderscheiden sectoren en de kwaliteit van de examinering te doen toenemen.

Het examenprofiel is een document in standaard format waarin op drie niveaus afspraken over examinering zijn beschreven: landelijk, sectoraal en regionaal. Het onderwijs, bedrijfsleven en de kenniscentra hebben samengewerkt bij de ontwikkeling van de profielen. Bedrijven zijn betrokken bij het vaststellen van de randvoorwaarden voor goede examinering. Er zijn 24 landelijke sectorale examenprofielen beschikbaar met afspraken rondom examinering voor alle kwalificatiedossiers. De overige producten die binnen het project Examenprofiel zijn ontwikkeld, hebben tot doel de mbo-instellingen te helpen om de kwaliteit van hun examinering te verbeteren.

De procesarchitectuur examinering die schematisch is afgebeeld in figuur 3 is één van die producten. De procesarchitectuur beschrijft alle processen die van belang zijn bij examinering in een mbo-instelling. Alle stappen die een instelling neemt om te zorgen voor goede examinering komen aan de orde: van het examenplan en de inkoop van examens tot het diplomeren en borgen van kwaliteit. De processen zijn te verdelen in zes onderwerpen: kaders stellen, leren, construeren en vaststellen, examineren, diplomeren en kwaliteitsborging. In het examenprofiel staat welke stappen de instelling samen met het bedrijfsleven neemt.



Figuur 3; Schema van de procesarchitectuur examinering zoals die ontwikkeld is voor het mbo (Bron: Landelijke Regiegroep Examinering (2010).

Een ander belangrijk product is het kostenmodel Examinering. Dit model is ontwikkeld door Price Waterhouse Coopers, het expertisecentrum beroepsonderwijs (Ecbo), het project Examenprofiel en diverse mbo-instellingen Het model maakt het mogelijk de feitelijke uitgaven en kosten voor examinering te benoemen en de kostprijs te berekenen. Het houdt rekening met de grote diversiteit in het mbo in uitvoeringsvormen van examinering. Zie http://www.examenprofiel.nl/?page_id=1169.

Tot Slot

Deze notitie heeft tot doel de voor het hbo relevante mogelijkheden voor examinering te beschrijven. De auteurs hopen van harte dat het advies van de commissie zal leiden tot een betere toets- en examencultuur in het hbo. Dat zal leiden tot betere beslissingen over studenten; betere beslissingen over de inrichting van onderwijsleerprocessen; verhoging van het rendement van de sector, gerichtere discussies met toezichthouders en betere kwaliteit van het onderwijs.

Literatuur

American Council on Education (2006). *Educational Measurement* (4th ed.). Westport, CT: American Council on Education and Praeger Publishers.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Anderson, L.W., & Krathwohl, D. (Eds.) (2001). *A taxonomy for learning, teaching and assessing: revision of Bloom's taxonomy of educational objectives*. New York: Longman.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Baartman, L.K.J. (2008). Assessing the assessment; Development and use of quality criteria for Competence Assessment Programs. Utrecht: Utrecht University.

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, 114-129.

Béguin, A. A. (2000) *Robustness of Equating High-Stakes Tests*. Enschede: Universiteit Twente.

Berkel, H. van, en A. Bax (2006). *Toetsen in het hoger onderwijs*. Houten: Bohn Stafleu van Loghum.

Bloom, B.S. (1975). *Taxonomy of educational objectives: Handbook 1: Cognitive domain*.
New York: David McKay Co.

COTAN. (2010). *Beoordelingssysteem voor de Kwaliteit van Tests*. Amsterdam: NIP

Cronbach, L.J. (1989) Construct validation after thirty years. In R.L. Linn (Eds.), *Intelligence: Measurement, theory and public policy. (147-171)*.

Dijkstra, A (2011). T*oetsing en nieuwe beoordelingsvormen; Gids voor borging en optimalisering van toetsbeleid*. Breda: Avans Hogeschool.

Eggen, Th, J.H.M. (2004). *Contributions to the theory and practice of computerized adaptive testing*. Enschede: Universiteit Twente.

Frederiksen, J.R., & Collins, A. (1989) A system approach to educational testing. *Educational researcher, 18, (9), 27-32*.

Groot, A.D. de, & Naerssen, R.F. van. (1969). Studietoetsen, construeren, afnemen, analyseren. Den Haag: Mouton.

Hambleton, R. K., & Pitoniak, M. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433-470). Westport, CT: American Council on Education and Praeger Publishers.

Haertel, E.H. (1991) New forms of teacher assessment*. Review of research in education, 17, 3-29*.

Haertel, E.H. (2006). Reliability. In R. Brennan (Ed.), *Educational Measurement (*4th ed., pp. 65-110*).* Westport, CT: American Council on Education and Praeger Publishers.

HBO-raad (2011). *Geslaagd! Handreiking examencommissies*. Den Haag: Hbo-raad.

Inspectie van het onderwijs (2011) *Alternatieve afstudeertrajecten en de bewaking van het eindniveau in het* hoger *onderwijs.*

Joosten-Ten Brinke, D. (2011). *Eigentijds toetsen en beoordelen.*Tilburg: Fontys Hogescholen.

Kane, M. (1992a) An argument-based approach to validity. *Psychological Bulletin, 112, 527-535*.

Kane, M.T. (1992b). *The validity of assessments of professional competence*. ERIC document reproduction Service No. ED 343958

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.) Setting performance standards: Concepts, methods, and perspectives (pp. 53–88). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Kane, M.T. (2006). Validation. In R. Brennan (Ed.), *Educational Measurement (*4th ed., pp. 17-64*).* Westport, CT: American Council on Education and Praeger Publishers.

Landelijke Regiegroep Examinering (2010). *Procesarchitectuur Examinering.* Xxx: Landelijke Regiegroep Examinering.

Linn, R.L., Baker, E., & Dunbar, S. (1991) Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 16,1-21*.

Messick, S. (1994) The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23, (2), 13-22*.

Mulder, M. (2003). Ontwikkelingen in het competentiedenken en competentiegericht beroepsonderwijs. In: M. Mulder, R. Wesselink, H. Biemans, L. Nieuwenhuis, & R. Poell (Eds.), *Competentiegericht Beroepsonderwijs. Gediplomeerd, maar ook bekwaam?* (p. 15-32). Houten: Wolters Noordhof.

Nederlandse Vereniging voor Examens (2011a). *Assessor; Functieprofielen Examenorganisatie*. Enschede: Nederlandse Vereniging voor Examens
Nederlandse Vereniging voor Examens (2011b). *Toetsconstructeur; Functieprofielen Examenorganisatie*. Enschede: Nederlandse Vereniging voor Examens
Nederlandse Vereniging voor Examens (2011c). *Toetsvaststeller; Functieprofielen Examenorganisatie*. Enschede: Nederlandse Vereniging voor Examens
Nederlandse Vereniging voor Examens (2011d). *Medewerker Examenbureau; Functieprofielen Examenorganisatie*. Enschede: Nederlandse Vereniging voor Examens
Nederlandse Vereniging voor Examens (2011e). Examencommissielid*; Functieprofielen Examenorganisatie*. Enschede: Nederlandse Vereniging voor Examens
Onderwijsraad (2010). *Een diploma van waarde.* Den Haag: Onderwijsraad.
RCEC (2011). *Beoordelingssysteem voor de kwaliteit van toetsen en examens*.
Rijn, P. W. van (2009). Mogelijke effecten van verschillende uitslagregels. *Examens, 6, 3, 5-8.*
Rijn, P.W. van, Béguin, A.A., & Verstralen, H.H.F.M. (2009). Zakken of slagen? De nauwkeurigheid van examenuitslagen in het voortgezet onderwijs. *Pedagogische Studiën, 86, 185-195.*
Rodriguez, M.C. (2003). Construct equivalence of multiple choice and constructed-response items: a random effects synthesis of correlations. *Journal of Educational Measurement, 40 (2)*, 163-184.
Roelofs, E. (2006). Een procesmodel voor de beoordeling van competent handelen. *Tijdschrijft voor Hoger Onderwijs, 24,* 152-167.
Roelofs, E. & Straetmans, G. (Eds.) (2006). *Assessment in Actie. Competentiebeoordeling in opleiding en beroep*. Arnhem: Cito.
Roorda (2008) Quality Systems for Testing. In: Wild, C. L. and Ramaswamy, R. (Eds.): *Improving Testing: Applying Process Tools and Techniques to Assure Quality*, (pp 145-176) New York: Lawrence Erlbaum Associates.
Sanders, P.F. (2011) De betrouwbaarheid van toetsscores. In: P.F. Sanders (Red.) *Toetsen op School*. Arnhem: Cito
Sanders, P.F. & Verstralen, H.F.M (2011) Het beoordelen van toetsscores. In: P.F. Sanders (Red.) *Toetsen op School*. Arnhem: Cito
Sluijsmans, D. M. A., Straetmans, G., & Van Merriënboer, J. (2008). Integrating authentic assessment with competency based learning: the Protocol Portfolio Scoring. *Journal of Vocational Education and Training, 60(2), 157-172.*
Stocking, M.L. & Lewis, C. (1995). *Controlling Item Exposure Conditional on Ability in Computerized Adaptive Testing*. (Research Report 95-24). Princeton, NJ: Educational Testing Service.
Straetmans, G.J.J.M. (1998). Toetsing van competenties. In Schramade, P.W.J. (Red.). *Handboek Effectief Opleiden, 9, 1, 3.01-3.36*. 's-Gravenhage: Elsevier Bedrijfsinformatie B.V.
Straetmans, G.J.J.M. (2004a). Protocol Portfolio Scoring. *OnderwijsInnovatie, 2, pp. 17-27*.
Straetmans, G.J.J.M. (2004b). Protocol Portfolio Scoring. *Een methode voor het systematisch scoren en vaststellen van competenties*. BVE en HO Brochurereeks Perspectief op assessment, nr 4. Arnhem: Citogroep.
Straetmans, G.J.J.M.& Eggen, T.J.H.M. (2011). WISCAT-pabo: Ontwerp, kwaliteit en resultaten van een geruchtmakende toets. In: Schramade e.a. (Red.) *Handboek Effectief Opleiden, sectie 9-2.2*. 's Gravenhage: Reed Business Information BV.
Straetmans, & P.F. Sanders (2001). *Beoordelen van competenties van docenten*. Den Haag: Programmamanagement EPS/HBO-raad.
Verschoor, A.J. (2007). *Genetic Algorithms for Automated Test Assembly*. Enschede: Universiteit Twente.
Wild, C.L. & Knapp, J.E. Standards in the Testing Industry. In: Wild, C. L. and Ramaswamy, R. (Eds.): *Improving Testing: Applying Process Tools and Techniques to Assure Quality*, (pp 59-81) New York: Lawrence Erlbaum Associates
Wools, S. (2011). De validiteit van toetsscores. In: P.F. Sanders (Red.) *Toetsen op School* . Arnhem: Cito
Zieky, M.J., Perie. M., & Livingston S.A. (2008). *Cutscores: A manual for setting standards of performance on educational and occupational tests.* Princeton: Educational Testing Service.