

Beoordelen van beroepsbekwaamheid

Inhoudsopgave

1	Kwaliteitseisen voor het beoordelen van beroepsbekwaamheid	4
2	Een praktische indeling van examenonderdelen	6
3	Sterke en zwakke punten van verschillende examenonderdelen	8
4	Besluitvorming bij beoordelen van beroepsbekwaamheid	10
5	Epiloog: Beroepsgerichte examens zijn dossierexamens	13
	Literatuur	14



1 **Kwaliteitseisen voor het beoordelen van beroepsbekwaamheid**

Meer dan 20 jaar geleden merkte de toetsdeskundige Michael Kane (1992) op dat het ondoenlijk is om beroepsbekwaamheid goed te beoordelen. Zijn stelling was dat alle methoden die daarvoor in te zetten zijn, zoals toetsen en tests, observatie van prestaties, generieke beoordeling en simulaties, in bepaalde opzichten iets tekort komen. Toetsen en tests zijn niet realistisch genoeg. Directe observaties zijn te subjectief en dus niet voldoende betrouwbaar. En simulaties en generieke beoordelingen zijn én niet echt realistisch én niet betrouwbaar. Eigenlijk is er geen beginnen aan, aldus Kane. Toch worden er ieder jaar talloze diploma's uitgereikt in het Nederlandse beroepsonderwijs. En ook beroepsbeoefenaren krijgen tijdens hun carrière diploma's en certificaten uitgereikt. Kunnen de gebruikte examens eigenlijk wel de toets der kritiek doorstaan? Valt op diploma's en certificaten wel voldoende staat te maken? Het doel van deze ToetsSpecial¹ is duidelijk te maken aan welke kwaliteitseisen beoordelingen moeten voldoen, willen we houdbare uitspraken kunnen doen over beroepsbekwaamheid.

¹ Voor deze ToetsSpecial is dankbaar gebruik gemaakt van het gedachtegoed van oud-Cito-collega G.J.J.M. Straetmans en eerdere samen met hem gepubliceerde documenten.

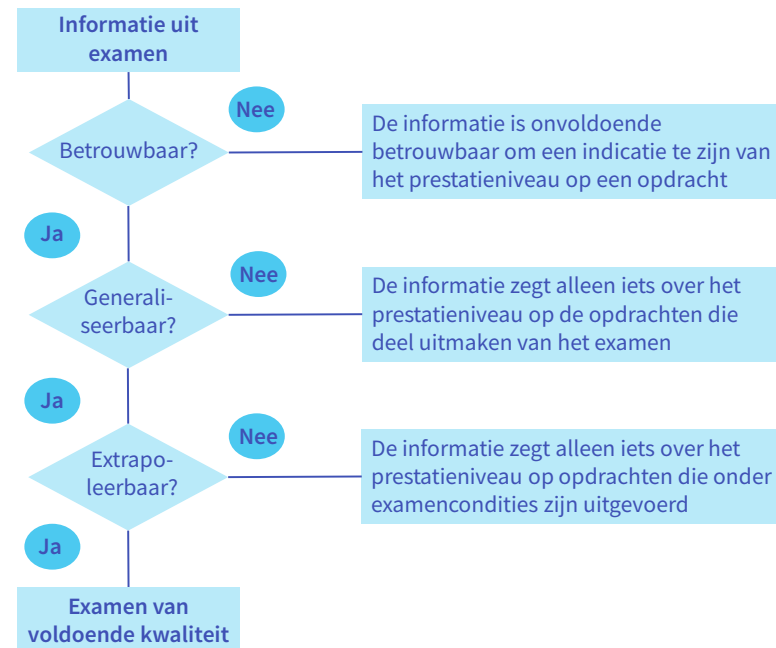
Voor het beschrijven van de eisen waar beroepsgerichte examens aan moeten voldoen, gebruiken we het gedachtegoed van Straetmans (2004a). De informatie die een beroepsgericht examen moet opleveren dient niet alleen voldoende betrouwbaar, maar ook voldoende generaliseerbaar en voldoende extrapoleerbaar te zijn. Wat die eisen betekenen maken we hieronder duidelijk.

Maar eerst een definitie: een *examen* is een onderzoek naar kennis, inzicht, houding en/of vaardigheden van een kandidaat. Die kandidaat dient over een samenhangend geheel van leergebieden, aan de hand van verstrekte opdrachten een aantal prestaties te leveren. Door vervolgens te controleren of de kandidaat voldoet aan concrete prestatie-eisen en transparante beslisregels toe te passen, is te bepalen of een diploma of certificaat verstrekt kan worden.

De informatie die een beroepsgericht examen oplevert, is *betrouwbaar* als de onderdelen waaruit het examen bestaat een voldoende nauwkeurige indicatie geven van het prestatieniveau van de kandidaat. Een gebrek aan betrouwbaarheid kan bijvoorbeeld tot uiting komen als 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 die scriptie. De betrouwbaarheid zal toenemen, als de beoordelaars gebruik kunnen maken van eenduidig geformuleerde prestatiecriteria die zich richten op concreet waar te nemen zaken. Ook het trainen van beoordelaars in het werken met de prestatiecriteria kan een positief effect hebben op de betrouwbaarheid (Sanders, 2013).

De informatie die een beroepsgericht examen oplevert, is *generaliseerbaar* als die het mogelijk maakt uitspraken te doen over andere opdrachten waar het examen óók betrekking op had kunnen hebben. Om de generaliseerbaarheid van de informatie te waarborgen, moet het examen een voldoende aantal uiteenlopende opdrachten omvatten.

De informatie die een beroepsgericht examen oplevert, is – tot slot – **extrapoleerbaar** als op basis ervan ook iets te zeggen valt over prestaties die in de beroepspraktijk te verwachten zijn. Informatie zal beter extrapoleerbaar zijn, naarmate de opdrachten die kandidaten uitvoeren en de situaties waarin ze dat doen meer gelijkenis vertonen met opdrachten en situaties uit de toekomstige werksituatie. Figuur 1 bevat een stroomdiagram dat grafisch laat zien hoe dat zit.



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



2 Een praktische indeling van examenonderdelen

In de literatuur is een groot scala aan termen te vinden voor het beschrijven van onderdelen van beroepsgerichte examens. Zie voor een overzicht bijvoorbeeld Van Berkel en Bax (2008) of Dijkstra (2011). Vanwege de beperkte ruimte is het hier niet mogelijk om een volledige lijst op te stellen en van ieder onderdeel aan te geven in hoeverre dit voldoet aan de drie genoemde kwaliteitseisen of daarvan afgeleide eisen. We gebruiken daarom een meer globale indeling.

Roelofs en Straetmans (2006) en Straetmans (2004a) hanteren een globale driedeling van 'hands-on'-instrumenten, simulaties en 'hands-off'-instrumenten. **Hands-on-instrumenten** zijn geschikt voor het beoordelen van opdrachten in reële werksituaties. Er is sprake van complexe werksituaties die niet of nauwelijks van de werkelijkheid verschillen. Bij dit soort opdrachten zijn de omstandigheden niet volledig beheersbaar. De moeilijkheidsgraad van het examen is daarom ook niet volledig onder controle.

Bij een **simulatie** moet een kandidaat in een tot op zekere hoogte kunstmatige situatie demonstreren voldoende vaardig of bekwaam te zijn. Bij dit type opdracht worden aspecten uit de beroepspraktijk gemanipuleerd. Er is dus sprake van een versimpeling van de werkelijkheid. Bij een **hands-off-instrument** krijgen deelnemers opdrachten uit te voeren waaruit moet blijken of zij over voldoende kennis beschikken om het betreffende beroep uit te oefenen. Dit kan op papier, maar gebeurt tegenwoordig steeds vaker met hulp van de computer, waardoor opdrachtsituaties bijvoorbeeld levensechter worden. Zie bijvoorbeeld: <http://www.vandermaesenkoch.nl/ecartoon/nl>.



Afbeelding 1 | Computersimulatie in de scheepvaart (© Nova College)

Een veelomvattender indeling dan die van Roelofs en Straetmans komt tot stand door uit te gaan van de aard van de assessmenttaak voor de kandidaat en de omstandigheden waaronder de kandidaat de taak uit moet voeren. De assessmenttaak bestaat uit de opdracht met bijbehorende informatie, waaronder de instructies die een kandidaat krijgt om de opdracht te kunnen maken. Zo'n taak kan expliciet of impliciet zijn. Een taak is explicieter, naarmate de opdrachten verder in detail zijn uitgewerkt, waardoor het gewenste gedrag (en de daaruit voortkomende al dan niet tastbare resultaten) beter en efficiënter is uit te lokken. Is de kandidaat zich bewust van het feit dat er een beoordeling plaatsvindt, bijvoorbeeld door de aanwezigheid van een of meer assessoren, dan zal deze zich normaal gesproken maximaal inspannen om tot een zo goed mogelijke prestatie te komen. Er is dan sprake van gecreëerde assessmentomstandigheden en de kandidaat zal zich bij het uitvoeren van de taak normaal gesproken maximaal inspannen om deze tot een goed einde te brengen. Is de kandidaat zich niet bewust van de beoordeling, dan is er sprake van natuurlijke werkomstandigheden en zal de kandidaat gedrag vertonen dat representatief is voor zijn of haar reguliere

prestatieniveau. Denk voor wat betreft maximaal prestatieniveau bijvoorbeeld aan het praktijkexamen rijvaardigheid van het CBR waar geen kandidaat het in zijn hoofd zal halen harder te rijden dan de maximaal toegestane snelheid. Dat het reguliere prestatieniveau van sommige kandidaten na het behalen van het rijbewijs anders is, weet iedereen die wel eens in een auto zit. Door de aard van het gedrag van de kandidaat (maximaal versus regulier prestatieniveau) af te zetten tegen de aard van de assessmenttaak (impliciet versus expliciet), ontstaat de indeling in tabel 1.

Tabel 1 | Indeling van assessmentvormen op basis van gedrag van de kandidaat en aard van de assessmenttaak (naar Sluijter & Straetmans, 2012)

		Assessmenttaak	
		Impliciet	Expliciet
Prestatieniveau	Maximaal	(1) Proeve van bekwaamheid	(2) Prestatiebeoordeling: - Hands-off - Simulatie - Hands-on
	Regulier	(3) Taxatie	(4) 'Mystery guest'

In cel (1) – **maximaal prestatieniveau bij een impliciete assessmenttaak**– vallen examenonderdelen waarvan de kandidaat weet dat deze tijdens het uitvoeren van de taak beoordeeld zullen worden. De opdracht is echter niet in detail uitgewerkt, vanwege de complexe en niet beheersbare situatie waarin deze moet plaatsvinden. Deze vorm van assessment valt te omschrijven als 'proeve van bekwaamheid'. De opdrachten die in deze categorie vallen, zijn levensecht, maar ook relatief inefficiënt als methode, omdat de opdracht voor de kandidaat per definitie niet in detail valt uit te werken. Een ander concreet voorbeeld vinden we in een relatief nieuw onderdeel van het rijexamen waarbij de examinerator de kandidaat vraagt om van punt A naar punt B te rijden, zonder daar verder instructies bij te geven.

In cel (2) – **maximaal prestatieniveau bij een expliciete assessmenttaak**– valt de prestatiebeoordeling (performance assessment); het betreft hier de typen opdracht die Roelofs en Straetmans (2006) onderscheiden. De kandidaat weet dat hij tijdens het reageren op de stimulus beoordeeld zal worden.

In tegenstelling tot de proeve van bekwaamheid wordt bij deze categorie door middel van gerichte opdrachten het gewenste gedrag (en de daaruit voorkomende, al dan niet tastbare resultaten) uitgelokt. De meeste 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 een '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 examinerator die de kandidaat exact vertelde welke opdrachten deze allemaal precies moest uitvoeren.

De examenonderdelen in cel (3) – **regulier prestatieniveau bij een impliciete assessmenttaak**– 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 beoordeling plaatsvindt en ook is niet duidelijk wat exact het gewenste gedrag is. Denk hierbij bijvoorbeeld aan systemen voor 360° feedback. Dit type 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.

Bij de examenonderdelen in cel (4) – **regulier prestatieniveau bij een expliciete assessmenttaak**– 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 bij uitstek geschikt om een beeld te krijgen van de werkelijke beroepshouding van beroepsbeoefenaren (in opleiding).



3 Sterke en zwakke punten van verschillende examenonderdelen

Op basis van de voorafgaande informatie is het mogelijk om in algemene bewoordingen weer te geven in hoeverre verschillende examenonderdelen aan de eerder genoemde algemene kwaliteitseisen voldoen.

In navolging van Straetmans (2004a), die dit al deed voor prestatiebeoordelingen, kunnen we de informatie uit de verschillende assessmentvormen beoordelen in termen van **betrouwbaarheid**, **generaliseerbaarheid** en **extrapoleerbaarheid**. De resultaten van deze exercitie zijn te vinden in tabel 2.

Tabel 2 | Evaluatie van examenonderdelen (Sluijter & Straetmans, 2012)

Examenonderdeel	Kwaliteitseis		
	Betrouwbaarheid	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

De tabel laat zien dat de proeve van bekwaamheid algemeen gesproken informatie oplevert die relatief weinig betrouwbaar en generaliseerbaar is, maar dat de extrapoleerbaarheid ervan uitstekend is. De hands-on-methode levert informatie die wat betrouwbaarder en generaliseerbaarder is, maar ook weer minder extrapoleerbaar. De simulatie levert op haar beurt informatie op die weer wat betrouwbaarder en generaliseerbaarder is, maar dit gaat wederom te koste van de extrapoleerbaarheid ervan. De hands-off-methode levert informatie op die relatief erg betrouwbaar 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 betrouwbaar. De mystery guest, ten slotte, verschaft informatie die eveneens redelijk extrapoleerbaar is, maar de generaliseerbaarheid is beperkt en de betrouwbaarheid ervan laat wel wat te wensen over. Dit leidt tot de onontkoombare conclusie dat er geen examenonderdeel bestaat dat in voldoende mate aan alle drie de kwaliteitseisen tegemoetkomt. En dit gegeven is dan ook de oorzaak van de al eerder aangehaalde verzuchting van Michael Kane.

Toch is de houdbaarheid van de stelling van Kane aan het afnemen. Weliswaar geldt het bovenstaande nog steeds in algemene zin, maar door de voortdurende ontwikkelingen op het vlak van digitalisering en automatisering, komen er steeds meer praktische mogelijkheden om de door Kane aangestipte problematiek tot het verleden te laten behoren. Simulaties worden steeds levensechter en de mogelijkheden om tijdens de simulaties relevante informatie te verzamelen en te analyseren worden ook steeds groter. Een bekend voorbeeld op dit vlak is te vinden in de vliegsimulatoren die bij de opleiding van piloten ingezet worden. Voordat piloten daadwerkelijk een vliegtuig besturen, hebben zij al vele uren doorgebracht in een vliegsimulator en veel ervaring opgedaan in een reeks uiteenlopende levensechte oefeningen.



Afbeelding 2 | KLM vliegsimulator voor de Airbus A330-200E (© KLM)

Een tweede aansprekend voorbeeld is dat van de [Moog Simodont Dental Trainer](#). Dit instrument is ontwikkeld door ACTA, het Academisch Centrum Tandheelkunde Amsterdam, en wordt bij steeds meer opleidingen ingezet bij het trainen van studenten. En bij enkele opleidingen wordt het instrument ook ingezet binnen het examenprogramma van aankomende tandartsen. Het instrument stelt studenten in staat virtuele patiënten te behandelen. Zij krijgen daarbij dezelfde tactiele feedback als bij het boren in een echte tand of kies. De dental trainer maakt onderdeel uit van een complete leeromgeving die in ontwikkeling is: de virtuele kliniek. Studenten kunnen met behulp van dit instrument en een 3D-bril boren in virtuele tanden en kiezen, waarbij zij van de boor dezelfde tactiele feedback krijgen als bij het behandelen van een echte patiënt. Dit [filmpje](#) maakt duidelijk hoe het werken ermee in de praktijk verloopt.



Afbeelding 3 | Dental trainer in gebruik vanuit het standpunt van de gebruiker en totaaloverzicht (© ACTA)

Een derde voorbeeld vinden we in een [voortgangstoets](#) van beeldinterpretatievaardigheden in de specialistenopleiding voor radiologen. In het onderzoek dat aan dit instrument ten grondslag lag, is vastgesteld dat digitaal toetsen met 3D-beelden veilig en betrouwbaar op grote schaal is uit te voeren en dat de kwaliteit ervan beter is dan minder authentieke vormen van toetsen (Ravesloot et al., 2015). De digitale voortgangstoets radiologie maakt onderdeel uit van het

reguliere toetsprogramma van de landelijke medisch-specialistische vervolgopleiding radiologie.

Het ontwikkelen van examenonderdelen als de vluchtsimulator en de dental trainer en hiermee vergelijkbare instrumenten is vooralsnog een uiterst kostbare zaak. Bovendien zijn deze instrumenten niet zo veelomvattend dat zij de complete beroepsbekwaamheid kunnen dekken. Daarom is er een andere insteek nodig om de gesignaleerde problematiek op te lossen. Die andere benadering ligt in het zodanig combineren van verschillende typen assessmenttaken in een examen dat de informatie die het examenprogramma als geheel oplevert te beschouwen is als voldoende betrouwbaar, generaliseerbaar en extrapoleerbaar.

Het is de kunst om vanuit onderwijsmeetkundig oogpunt tot een zo goed mogelijke combinatie van assessmentvormen te komen, rekening houdend met de aard van het beroep of de beroepsopleiding, randvoorwaarden als de beschikbare tijd voor ontwikkeling en afname van de onderdelen 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 opdrachten die in ieder geval al voor een deel gedurende de opleiding worden afgenomen. Welke concrete vorm die verschillende opdrachten het best kunnen aannemen, is zoals aangegeven, grotendeels afhankelijk van de aard van de opleiding.

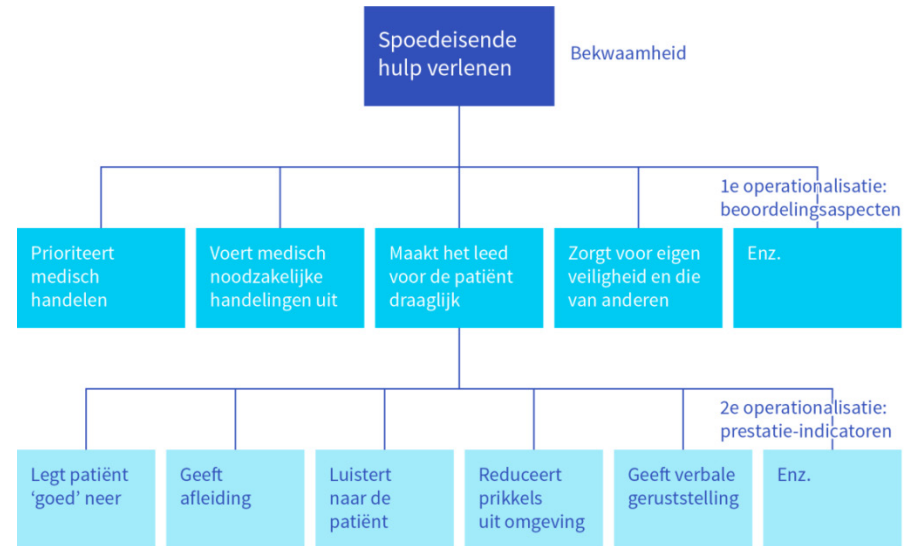


4 Besluitvorming bij beoordelen van beroepsbekwaamheid

Een verantwoorde en inzichtelijke manier om resultaten op verschillende assessments te combineren tot een eindoordeel is ontwikkeld door Straetmans (2004a). Het door hem ontwikkelde Protocol Portfolio Scoring (PPS) laat zien hoe prestaties op uiteenlopende typen opdrachten op verantwoorde en transparante wijze te integreren zijn tot een eindbeslissing. Het onderliggende principe hierbij is dat een zinvolle integratie van prestaties op verschillende assessmentvormen uitsluitend mogelijk is als 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 betekent om bekwaam in een bepaald vakgebied te zijn, dient om te beginnen uitgewerkt te zijn in een reeks concreet omschreven (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 iedere opdracht. Welke beoordelingsaspecten relevant zijn, is afhankelijk van de aard van de opdracht.

De beoordelingsaspecten worden op hun beurt geoperationaliseerd in een reeks opdrachtspecifieke indicatoren. Dat wil zeggen dat bij iedere opdracht een specifieke set van prestatiecriteria hoort die voortkomen uit de beoordelingsaspecten. Zie voor een concreet voorbeeld van een 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 (Sluijter & Straetmans, 2012)

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

- **Multi shot:** beroepsbekwaamheid is een veelomvattend construct dat onmogelijk met één prestatie op één opdracht aan te tonen valt.
- **Multi-method:** geen enkel type opdracht levert op zichzelf informatie die zowel voldoende betrouwbaar, generaliseerbaar als extrapoleerbaar is. De inzet van meerdere uiteenlopende opdrachten is daarom noodzakelijk.
- **Multi-rater:** prestatiecriteria zijn, zonder dat ze triviaal worden, nooit zodanig te formuleren dat ze objectief scorebaar 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 persoon te laten beoordelen.
- **Gestandaardiseerde set prestatiecriteria:** elke prestatie die als bewijs voor de aanwezigheid van een bepaald aspect van (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 opdracht-specifiek.
- **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 manier vastgestelde grenswaarden, ook wel cesuren genoemd. Zie voor meer informatie hierover hoofdstuk 9 van Toetsen op School (Sanders & Verstralen, 2013).

Tabel 3 bevat een fictief en beperkt voorbeeld van een prestatiedossier om een concreet beeld te geven van hoe beslissingen verlopen. Dit dossier bevat de resultaten op zes verschillende opdrachten. Er zijn zes verschillende **beoordelingsaspecten**, gelabeld A tot en met F.

Bij de eerste opdracht is via een hands-off-methode informatie verzameld over prestaties op drie 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 deze opdracht, die echter een **cesuur** (grenswaarde) heeft van 14,4. Daarmee is het resultaat op deze opdracht onvoldoende.

Verder laat de tabel zien dat ieder beoordelingsaspect minimaal twee maal aan bod gekomen is in de totale reeks opdrachten. Merk op dat beoordelingsaspect A in alle zes de opdrachten 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 iedere opdracht, 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 opdrachten 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.

Daarnaast 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. Daarom is opdracht 1 bij het bepalen van alle scores buiten beschouwing gelaten.

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 opdrachten met onvoldoende resultaat zijn afgesloten, kan nog niet geconcludeerd worden dat de betreffende kandidaat al voldoende beroepsbekwaam is.

Tabel 3 | Fictief voorbeeld van een prestatiedossier (naar Sluijter & Straetmans, 2012)

Beoordelingsportefolio				beoordelingsaspecten												totaal-score	horizontale standaard	resultaat
bewijsstuk	datum	assessment-methodiek	assessor	A	B	C	D	E	F	G	H	I	J	K	L			
1	13-12-15	hands-off		4			5		3							12	14,4	-
2		hands-off		3			5		6							14	12,4	+
3		simulatie	J.Pac.	5	5											10	9,8	+
4		hands-on	H.Bol.	5	5	4	4	4								22	23,5	-
5		hands-on	P.Hak.	7		5	6	5	6							29	23,3	+
6		simulatie	J.Pac.	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				+	-	+	+	+	+									

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 beroepsopleidingen (Sluijsmans, Straetmans & Van Merriënboer, 2008). Een bezwaar dat tegen de boven beschreven procedure kan worden ingebracht, is dat uit de literatuur bekend is dat scores binnen sommige taken veel minder kunnen variëren dan andere. En dit komt niet tot uiting in de hierboven beschreven benadering. Een andere mogelijk geschikte toepassing die dit probleem niet kent, is de multivariate generaliseerbaarheidstheorie. Bij toepassing van deze statistische techniek wordt langs empirische weg bepaald of op grond van scores op een verzameling taken betrouwbaar en valide te concluderen valt of een kandidaat voldoende bekwaam is. Via een standaardbepalingsmethode wordt vastgesteld welke gewichten de scores op de gebruikte assessmenttaken krijgen bij het bepalen of een kandidaat voldoende beroepsbekwaam is (Wu, 2015). Een concrete behandeling van deze techniek valt buiten het bestek van deze notitie.



5 Epiloog: Beroepsgerichte examens zijn dossierexamens

De kwaliteit van de examinering bij beroepsopleidingen en de certificering in het bedrijfsleven moet boven alle twijfel verheven zijn. Er is immers sprake van beslissingen over personen met zwaarwegende consequenties voor schoolloopbanen en carrières. Bovendien vormt examinering en certificering – als deze aan de kwaliteitseisen voldoet – het middel bij uitstek om de kwaliteit van beroepsopleidingen en beroepsbeoefenaren en het vertrouwen van de maatschappij daarin te borgen.

Voor havo en vwo vormen de centrale examens in hun huidige vorm het meest kostenefficiënte instrument voor de borging van de kwaliteit van opleidingen en de vergelijkbaarheid van diploma's. Centrale examens bij het voortgezet onderwijs hebben een vorm die de mogelijkheid biedt om te constateren of leerlingen voldoende hebben opgestoken van het voorafgaande onderwijs. En daarnaast geven zij ook de gelegenheid om geldige uitspraken te doen over de kwaliteit van specifieke opleidingen en om de kwaliteit van specifieke opleidingen over instellingen heen met elkaar te vergelijken. Bij beroepsopleidingen is het echter niet mogelijk om het examen uitsluitend uit onderdelen met een vergelijkbare vorm te laten bestaan.

In de eerste plaats niet, omdat instellingen zo niet in staat gesteld worden te bepalen of een opleiding tot voldoende beroepsbekwaamheid heeft geleid. Bij een beroepsopleiding hoort een dossierdiploma: vaststellen of iemand voldoende vakbekwaam is, kan uitsluitend door het inzetten van een evenwichtige mix van uiteenlopende opdrachten die voor een groot deel niet aan het einde maar tijdens een opleiding worden aangeboden. Prestaties op deze opdrachten dienen vastgelegd te worden in een zogeheten

beoordelingsportfolio. Opdrachten met dezelfde vorm als de centrale examens, kunnen overigens wel onderdeel uitmaken van zo'n mix.

De tweede belangrijke reden dat centrale examinering niet de voorkeur verdient bij beroepsopleidingen en het certificeren van beroepsbeoefenaren, vloeit direct voort uit het bovenstaande. Niet alleen is er geen sprake meer van één vast examenmoment, ook de inhoud van het portfolio zal per student verschillen. Er is tot op zekere hoogte sprake van individuele leerwegen binnen beroepsopleidingen. Daaruit volgt dat studenten verschillende onderdelen van hun opleiding op verschillende tijdstippen zullen afronden. Er zal daarom maar een beperkte overlap zijn tussen de portfolio's van verschillende studenten. Datzelfde geldt natuurlijk voor de werkervaring die professionals opdoen tijdens hun carrière. Het uitsluitend werken met centraal vastgestelde examens is in deze situatie praktisch onmogelijk. Er zouden immense hoeveelheden examens ontwikkeld moeten worden om aan het criterium van de geheimhouding te kunnen voldoen.

Natuurlijk kan voorbijgegaan worden aan bovenstaande bezwaren door alleen die onderdelen te examineren die wél centraal te examineren zijn. Dat zou echter grote gevolgen hebben voor de kwaliteit van het onderwijs. Een bekend en niet te veronachtzamen effect van examens is het 'backwash'-effect: de inhoud van examens stuurt het leergedrag en de motivatie van deelnemers sterk. Anders gezegd, studenten leren voornamelijk die onderdelen van het curriculum waarvan duidelijk is dat ze ook werkelijk geëxamineerd worden. Centrale examinering kan zich per definitie niet richten op alle aspecten van de beroepsbekwaamheid en zou bij te eenzijdige benadrukking van bepaalde onderdelen kunnen leiden tot de uitstroom van onvoldoende gekwalificeerde beroepsbeoefenaren. Gediplomeerden beschikken dan wel over voldoende kennis, maar zijn wat beroepshouding en beroepsvaardigheden betreft onvoldoende toegerust om hun vak op een aanvaardbaar niveau uit te voeren. Het uitsluitend werken met een klassiek systeem van examinering bij beroepsopleidingen is dan ook vanuit meettechnisch, praktisch en strategisch standpunt bezien onverstandig.



Literatuur

Baartman, L.K.J. (2008). *Assessing the assessment; Development and use of quality criteria for Competence Assessment Programs*. Proefschrift: Dutch Interuniversity Centre for Educational Research.

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

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

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

Ravesloot, C.J., Van der Schaaf, M.F., Van Schaik, J.P.J, Ten Cate, O.Th J., Van der Gijp, A., Mol, C.P. & Vincken, K.L (2015). Volumetric CT-images improve testing of radiological image interpretation skills. *European Journal of Radiology*, 8 (5), 856-861.

Roelofs, E. & Straetmans, G. (Eds.) (2006). *Assessment in Actie. Competentiebeoordeling in opleiding en beroep*. Arnhem: Cito.

Sanders, P.F. & Verstralen, H.F.M (2013) Het beoordelen van toetsscores. In: P.F. Sanders (Red.) *Toetsen op School. Herziene versie*. Arnhem: Cito

Sluismans, 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.

Sluijter, C & Straetmans, G.J.J.M. (2012). Examens in het hbo; kwaliteitseisen en kwaliteitsborging. Den Haag: HBO-raad.
http://www.vereniginghogescholen.nl/system/knowledge_base/attachments/files/000/000/332/original/Examinering_in_het_hbo_-_Commissie_EVEX.pdf?1442927745

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. & Sanders, P.F. (2001). *Beoordelen van competenties van docenten*. Den Haag: Programmamanagement EPS/HBO-raad.

Wu Y.-F. & Tzou H. (2015). A Multivariate Generalizability Theory Approach to Standard Setting. *Applied Psychological Measurement*, 39, 7, pp. 507-524.

ToetsSpecial

Over deze ToetsSpecial | Bij beroepsonderwijs draait het om het afleveren van competente beginnende professionals. En tijdens hun carrière moeten beroepsbeoefenaren in toenemende mate periodiek aantonen nog steeds berekend te zijn op de snel veranderende vereisten van hun vak. Bij de diplomering van beroepsbeoefenaren en in veel gevallen ook bij hun (her)certificering spelen examens een beslissende rol. Deze special gaat kort in op de vraag hoe (beginnende) beroepsbekwaamheid vast te stellen is. De special is bedoeld voor toetsdeskundigen, examenfunctionarissen, leden van examencommissies en onderwijsgevendenden in het beroepsonderwijs. Maar ook beroepsbeoefenaren zélf en studenten kunnen op basis van deze tekst beter zicht krijgen op de kwaliteit van de examinerings- of en certificeringspraktijk in hun sector.

Colofon

Uitgave: ToetsWijzer | Stichting Cito Instituut voor Toetsontwikkeling

Auteur: Cor Sluijter (Cito)

Datum: december 2017

e info@toetswijzer.nl | i www.toetswijzer.nl



ToetsWijzer