De betrouwbaarheid van de praktijksimulatie bij de Nederlandse brandweer

Esther Oprins (TNO), Ronald Heus (Nbbe), Chantal Ruijten (Nbbe), Gerard Veldhuis (TNO), Ward Venrooij (TNO)

De betrouwbaarheid van praktijkexamens

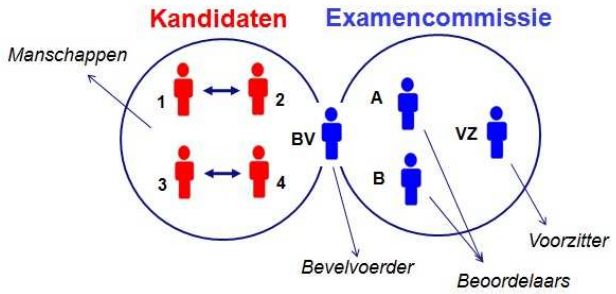
De betrouwbaarheid van praktijkexamens is van groot belang. Betrouwbaarheid bij het beoordelen van prestaties (cf. 'performance assessment') houdt in dat herhaalde afname tot hetzelfde resultaat moet leiden, ook al wisselen de taak, de beoordelaars of de omstandigheden (e.g. Berk, 1986; Murphy & Cleveland, 1995). Het inrichten van een betrouwbaar examen waarbij menselijke beoordelaars op gedrag beoordelen blijkt in de praktijk een uitdaging te zijn. De vergelijkbaarheid van de taak (bijv. scenario's) is soms lastig te organiseren. De mate van overeenstemming tussen beoordelaars ofwel de 'interbeoordelaarsbetrouwbaarheid' is vaak beperkt (Guion, 1998). Beoordelingsfouten zijn in de praktijk nauwelijks te vermijden, bijv. het halo-effect: het generaliseren van een oordeel over een kandidaat over meerdere aspecten in positieve dan wel in negatieve zin (Thorndike, 1920). Indien het examen met meer dan één kandidaat wordt afgenomen, speelt het probleem van onderlinge beïnvloeding. De criteria waarop wordt beoordeeld zijn niet altijd eenduidig vast te stellen omdat het werk in de praktijk vaak op verschillende manieren uitgevoerd kan worden. De beoordelingscriteria zouden afgeleid moeten zijn van competenties en zo goed mogelijk geformuleerd in observeerbaar gedrag (e.g. Berk, 1986; Flin & Martin, 2001; O'Connor et al, 2002; Murphy & Cleveland, 1995). De norm moet juist zijn benoemd en begrensd om het aantal 'valse negatieven' (ten onrechte gezakt) en 'valse positieven' (ten onrechte geslaagd) zo laag mogelijk te houden (Roe, 2005). Voor functiegroepen waarbij een fout in taakuitvoering kan leiden tot ernstige gevolgen op persoonlijk, materieel en immaterieel vlak (bijv. in de luchtvaart, hulpdiensten en medische beroepen), geldt dat de 'false positives' het meest ongewenst zijn.

Voor domeinen waarin de veiligheid centraal staat, is de betrouwbaarheid van een examen cruciaal. Dit kan voorkomen dat er kandidaten ten onrechte slagen en fouten gaan maken in de praktijk. Binnen de luchtvaart (piloten, luchtverkeersleiding) is relatief wat meer onderzoek uitgevoerd naar het ontwerpen van een betrouwbaar beoordelingssysteem en het evalueren hiervan (e.g. Flin & Martin, 2001; O'Connor et al, 2002; Oprins et al, 2006, 2008; Oprins, 2008). Binnen het domein van de brandweer is dit relatief nieuw. Het aantonen van de betrouwbaarheid is van belang om verbeteringen te kunnen aanbrengen in het ontwerp van het examen. In de praktijk loopt men in dergelijk onderzoek echter regelmatig tegen praktische beperkingen aan zoals te lage aantallen kandidaten, niet gedigitaliseerd afnemen van examens, en gebrek aan heldere (kwantitatieve) eisen waaraan de betrouwbaarheid van praktijkexamens zou moeten voldoen.

Het onderzoek beschreven in dit paper levert een bijdrage aan de manier waarop onderzoek naar de betrouwbaarheid van praktijkexamens uitgevoerd kan worden en hoe dergelijke examens het beste ingericht kunnen worden. Het onderzoek is uitgevoerd bij het Nederlands bureau brandweerexamens (Nbbe) dat de wettelijke taak heeft brandweerexamens in Nederland af te nemen. Dit paper beschrijft de gehanteerde methodemix van kwantitatief en kwalitatief onderzoek die toepasbaar is in de meeste praktijkexamens, zeker als het veiligheidsfuncties betreft. De resultaten hebben geleid tot aanbevelingen voor de optimale inrichting van dergelijke praktijkexamens.

De praktijksimulatie bij het Nbbe

Het Nbbe werkt sinds 2008 met een nieuwe vorm van examens, de zogenaamde Proeve van Bekwaamheid (PvB). Het Nbbe doet in samenwerking met TNO onderzoek naar de psychometrische kwaliteit en bruikbaarheid van de PvB van de brandweerfunctie Manschap A ter evaluatie en verbetering; de betrouwbaarheid maakt deel uit van dit onderzoek. Deze PvB is een methodemix waarbij de kandidaat eerst bepaalde verrichtingen voldoende moet hebben ('toetskaart'), vervolgens drie theorietoetsen moet halen en vervolgens drie 'praktijksimulaties' moet doen. In dit praktijkexamen worden de prestaties van kandidaten beoordeeld in een realistische setting. Voorbeelden hiervan zijn: brandbestrijding in een gebouw of het redden van slachtoffers bij auto-ongelukken (= technische hulpverlening). Qua teamsamenstelling is de praktijk exact nagebootst: vier kandidaten vormen een team van vier manschappen die tijdens de inzet worden aangestuurd door de bevelvoerder, zie Figuur 1:



Figuur 1: Teamsamenstelling kandidaten en examencommissie

De bevelvoerder heeft tevens een beoordelende rol en wordt in dit paper daarom 'bevelvoerder-beoordelaar' genoemd. Hij beoordeelt het eerste deel van de inzet namelijk voor, tijdens en vlak na aanrijden naar de lokatie (bijv. brandend gebouw). Daarna observeren twee beoordelaars die in het scenario verder geen rol hebben het gedrag van de vier kandidaten tijdens een fictieve inzet; elke beoordelaar beoordeelt twee kandidaten tegelijk. Elke kandidaat wordt dus door de bevelvoerder-beoordelaar en één (aparte) beoordelaar beoordeeld. De examencommissie staat onder leiding van een voorzitter die de organisatie van de examens op zich neemt. Alle kandidaten doen twee inzetten. Bij de tweede inzet krijgen de kandidaten een andere beoordelaar.

Na afloop vullen alle beoordelaars een hiervoor vervaardigd beoordelingsformulier in dat voor de bevelvoerder-beoordelaar andere criteria vraagt dan voor de aparte beoordelaars. Ze geven allen op de set beoordelingscriteria aan of de kandidaten voldoende of onvoldoende hebben geacteerd. De beoordelingscriteria zijn direct afgeleid van de nationale eisen voor de beroepsgroep. Ze zijn verdeeld over diverse rubrieken. Elke rubriek bevat 1 tot 3 criteria waarop de beoordelaar een rating moet geven. Voorbeelden zijn: 'afwegen veiligheid en geaccepteerd risico tegenover het te dienen belang' behorende bij de rubriek 'veiligheid', en 'terugkoppelen relevante informatie' behorende bij de rubriek 'communicatie'. Sommige criteria wegen zwaarder dan andere. De weging is terug te vinden in drie type onvoldoendes, te weten een O, een O* (telt dubbel) of een KO (kardinale onvoldoende; kandidaat direct gezakt) voor het maken van veiligheidsfouten. De beoordelaar heeft steeds de keus tussen V (voldoende) of één van de drie type O (onvoldoende) die bij het criterium hoort. Feitelijk is er dus sprake van een tweepunts ratingschaal. Als de kandidaten beide inzetten hebben gedaan worden alle scores bij elkaar opgeteld en dit leidt tot een eindoordeel geslaagd of gezakt. Hiervoor is een cesuur vastgesteld waarbij een specifieke expert-consensus

methode is gevolgd gebaseerd op Angoff (1971). Dit is een kwantitatieve methode waarin de experts de criteria op mate van belangrijkheid moeten scoren met als referentie een 'grenskandidaat' die net geslaagd zou zijn.

Onderzoeksmethoden

De centrale vraagstelling in het onderzoek was: 'hoe betrouwbaar is de praktijksimulatie die wordt gebruikt voor het examineren van kandidaten opgeleid voor de functie Manschap A bij het Nbbe'? Om deze vraag te beantwoorden is een mix van kwalitatieve en kwantitatieve onderzoeksmethoden toegepast. De resultaten zijn met elkaar gecombineerd om op basis hiervan aanbevelingen te geven ter verbetering van de examens.

Ten eerste is een kwantitatieve analyse uitgevoerd op de beoordelingsformulieren van één van de drie praktijksimulaties die worden afgenomen voor de functie Manschap A, namelijk 'brandbestrijding' (BB). Er is gekeken naar de frequentieverdelingen per criterium en meer specifiek naar het onderscheid gezakte versus geslaagde kandidaten. De samenhang tussen criteria (interne consistentie) is onderzocht door de Cronbach's alphas te berekenen waarbij in dit geval een minimum waarde van .70 wordt aangehouden (Schmidtt, 1996). Dit is een absoluut minimum; zeker op individueel niveau wordt ook wel .80 of .90 aangehouden (Clark & Watson, 1995). De alpha's zijn ook afhankelijk van het aantal gebruikte items in een schaal en mogelijke intercorrelaties. Er is eigenlijk een tweepuntsschaal gehanteerd (zie vorige paragraaf). Daarom is in plaats van een gewone factoranalyse een 'principal component analysis allowing for categorical data' (CATPCA) uitgevoerd (Theunissen et al., 2003) om inzicht te krijgen in de interne structuur. De onderzoeksresultaten leveren informatie op over hoe de in gebruik zijnde formulieren verder geoptimaliseerd kunnen worden. Bovendien is geëxploreerd met behulp van een variantieanalyse (ANOVA) in hoeverre beoordelaars verschillen in het gemiddelde aantal (on)voldoendes dat ze geven (Oprins, 2008).

Vervolgens is de manier waarop de praktijksimulaties worden afgenomen kwalitatief geëvalueerd. Er zijn interviews gehouden met ontwerpers van de praktijksimulatie en de gekozen aanpak, en daarnaast met kandidaten en beoordelaars. Ook zijn observaties uitgevoerd tijdens de uitvoering van praktijksimulaties en tijdens een training voor beoordelaars. Een bestaande set kwaliteitscriteria, genaamd ´wheel of competency assessment´ (Baartman et al., 2006) is hierbij als basis gebruikt. In de interviews en observaties is gekeken naar het verloop van het examenproces zoals de rol en werkwijze van de beoordelaars, de scenariokeuze, en de manier waarop de formulieren worden ingevuld. Het evaluatieproces is door drie experts op het gebied van beoordelen bediscussieerd om tot consensus te komen.

Onderzoeksresultaten

De dataset van Brandbestrijding (BB) bestond uit 1768 formulieren: de helft is ingevuld door de bevelvoerder-beoordelaars en de andere helft door de (aparte) beoordelaars. Omdat elke kandidaat per examen twee inzetten doet waarbij in totaal vier formulieren ingevuld zijn, betreft de dataset dus de praktijksimulaties van 442 individuele kandidaten.

<u>Frequenties</u>

Tabel 1 presenteert de frequenties van de bevelvoerder-beoordelaar, voldoende (V) of onvoldoende (O), voor alle criteria behorende bij de rubrieken: 1. Voor het aanrijden, 2. Tijdens het aanrijden, 3.

Ter plaatse. Tabel 2 toont de frequenties van de beoordelaar voor de rubrieken: 4. Communicatie, 5. Gevaars- risicoherkenning, 6. Persoonlijke bescherming, 7. Materialen, 8. Veiligheid, 9. Samenwerking, 10. Professionele houding. Indien een O* of een KO hoort bij een criterium in plaats van O staat dit erbij. Uit beide tabellen blijkt dat op alle criteria verreweg het meeste een V wordt gegeven. Er blijkt dat criterium 2.3 het vaakst met V wordt beoordeeld (99,3%) en criterium 8.3 het minst (84,7%).

	1.1	2.1	2.2	2.3	2.4	3.1 (O*)	3.2	3.3
O	13	11	7	6	115	78	79	70
V	871	873	877	877	769	804	804	813

Tabel 1: Frequenties Bevelvoerder-beoordelaar

	4.1	4.2	5.1 (KO)	6.1 (O*)	6.2	7.1	7.2	8.1 (KO)	8.2 (KO)	8.3	9.1	9.2	10.1	10.2
O	40	29	32	124	81	126	72	125	45	135	49	45	100	32
V	844	853	852	760	802	755	810	756	833	748	835	838	784	851

Tabel 2: Frequenties Beoordelaar

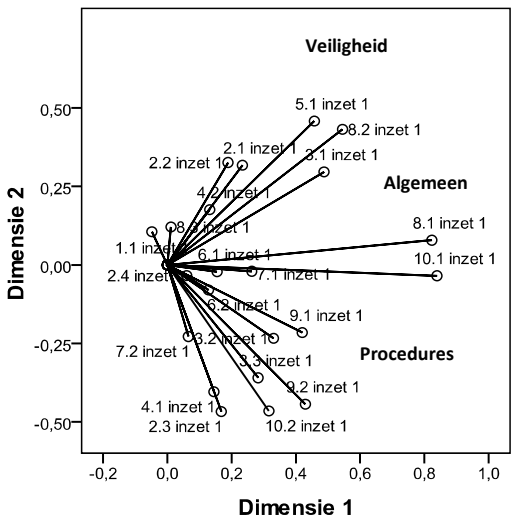
Daarnaast is geanalyseerd hoe de verdeling was voor geslaagde versus gezakte kandidaten. Van de 442 geëxamineerde kandidaten zijn er 313 (71%) geslaagd en 129 (29%) gezakt. 124 van de 129 gezakte kandidaten zakken op een KO. Veruit het meest van deze KO's (97%) worden gegeven op criterium 8.1 ('optreden (totale inzet'), gevolgd door 78% op criterium 10.1 ('operationele inzetbaarheid'). De overige 5 kandidaten zakken op een combinatie van O's en O*'s volgens de cesuur. De verhouding KO's bij inzet 1 versus inzet 2 is 54% versus 46%. Dit zal te maken hebben met een leereffect: het is waarschijnlijk dat de kandidaten relatief beter presteren tijdens inzet 2 omdat ze al een keer een praktijksimulatie hebben gedaan in dezelfde setting en dus hebben geoefend. Het zal niet zozeer aan de beoordelaar liggen omdat de kandidaat bij de tweede inzet een andere beoordelaar heeft.

Interne consistentie

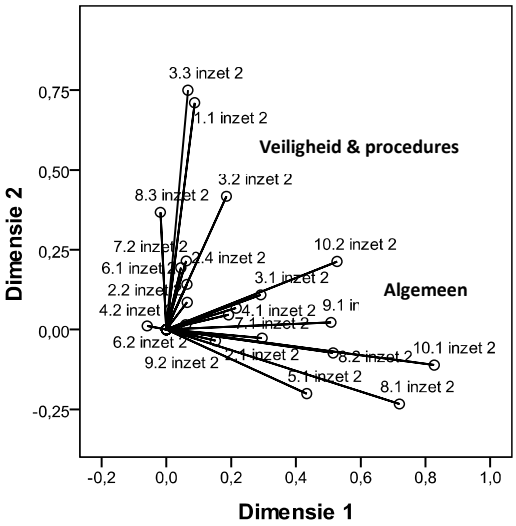
De Cronbach's alpha voor het geheel van de 22 criteria bij 'brandbestrijding' voor inzet 1 is .59 en voor inzet 2 is de alpha .49. Dit is volgens de standaard van .70 (Schmiddt, 1996) behoorlijk laag gezien het aantal criteria en het feit dat de uitslag moet wijzen op 'goed vakmanschap'. Dit resultaat is waarschijnlijk deels veroorzaakt door de gehanteerde tweepuntsschaal en (hierdoor) lage spreiding bij de meeste kandidaten. Normaal gesproken worden de alpha's juist ook per competentie of rubriek berekend maar dat is in dit geval niet mogelijk omdat het aantal criteria per rubriek sterk verschilt en zelfs soms maar uit één criterium bestaat.

Factoranalyse

Figuur 2 presenteert de resultaten van de CATPCA voor inzet 1: dimensie 1 (Cronbach's alpha = ,71; eigenvalue = 3,054) en dimensie 2 (Cronbach's alpha = ,48; eigenvalue = 1,841). Figuur 3 presenteert de resultaten voor inzet 2: dimensie 1 (Cronbach's alpha = ,64; eigenvalue = 2,537) en dimensie 2 (Cronbach's alpha = ,42; eigenvalue = 1,667).



Figuur 2: CATPCA inzet 1



Figuur 3: CATPCA inzet 2

De CATPCA voor inzet 1 leidt tot grofweg drie clusters criteria. De indeling in rubrieken wordt hierin niet direct herkend maar er is wel inhoudelijke samenhang. Het eerste cluster bestaat uit criteria die sterk met veiligheid te maken hebben waaronder 5.1 'herkennen van toepasselijke gevaren' en 8.2 'afwegen veiligheid en geaccepteerd risico tegenover het te dienen belang'. Het tweede cluster heeft slechts uit twee algemene criteria en beslaat het 'optreden van de totale inzet' (8.1) en de 'operationele inzetbaarheid' (10.1). Dit zijn tevens de twee criteria waarop verreweg de meeste KO's zijn gegeven. Het derde cluster wordt gevormd door criteria die met procedures te maken hebben waaronder 2.3 'omhangen persoonlijke beschermingsmiddelen en 4.1 ' Terugkoppelen van relevante informatie'. De uitkomsten van de analyse van de tweede inzet komen ongeveer overeen maar er is hier eerder sprake van twee clusters criteria. Het eerste cluster bestaat uit criteria die met procedures en veiligheid te maken hebben zoals 1.1 'Alarmering en controle kleding bij elkaar' en 8.3 'Omgeving' (8.3). Het tweede cluster bestaat net als bij inzet 1 uit de algemene criteria 8.1 'optreden van de totale inzet' en 10.1 'Operationele inzetbaarheid', maar ook 8.2 'Afwegen van veiligheid tegen belang (8.2). Het valt op dat alle KO's in het tweede cluster vallen.

<u>Vergelijking tussen beoordelaars</u>

De resultaten van de variantieanalyse (ANOVA) voor alleen de (aparte) beoordelaars (N=46) voor 886 formulieren laten zien dat gemiddelde aantal voldoendes op 14 criteria significant verschilt per beoordelaar (*F*=4,448, *p*=,000) net als het gemiddelde aantal onvoldoendes (*F*=6,423, *p*=,000). Hetzelfde is gedaan voor bevelvoerder-beoordelaars (N=45) voor 891 formulieren: het gemiddelde aantal voldoendes verschilt per beoordelaar significant op 8 criteria (*F*=6,716, *p*=,000) alsmede het gemiddeld aantal onvoldoendes (*F*=6,472, *p*=,000). Dit suggereert dat er verschillen tussen beoordelaars bestaat bijv. beoordelaars die strenger zijn dan anderen. Hierbij moet echter rekening gehouden worden met het feit dat zij steeds andere kandidaten in verschillende scenario's hebben beoordeeld. Dit kan de verschillen tussen de beoordelaars hebben vergroot.

<u>Check op evaluatieraamwerk (Baardman et al., 2006).</u>

Tabel 3 geeft een samenvatting van de kwalitatieve resultaten volgens het evaluatieraamwerk (Baardman et al., 2006) verkregen op basis van observaties en interviews. Vooral de eerste twee

(vergelijkbaarheid en herhaalbaarheid) hebben betrekking op de beoordeling van de betrouwbaarheid maar de rest van de kwalitatieve criteria draagt ook hieraan bij.

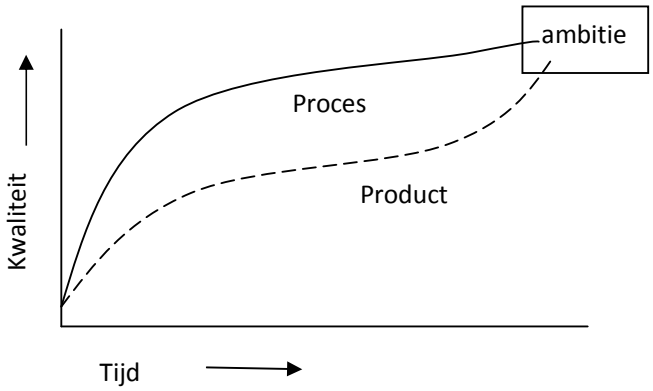
Vergelijkbaarheid: De condities waarin de PvB wordt afgenomen, incl. de criteria, scoringssysteem en beoordelingstaak moeten constant blijven zodat kandidaten met elkaar vergeleken kunnen worden.
• Het beoordelingsproces is zorgvuldig ingericht bijv. het protocol en samenstelling examencommissie. • De criteria zijn gelijk voor alle examens, maar beoordelaars interpreteren ze niet altijd op dezelfde wijze. • Het scenario wordt vastgesteld door de examencommissie, maar is nog niet volledig gestandaardiseerd.
Herhaalbaarheid van beslissingen: Beslissingen gemaakt op basis van de uitkomst van de PvB moeten accuraat en constant zijn over de tijd, over meerdere beoordelingssituaties, en meerdere beoordelaars.
• Er zijn twee inzetten per kandidaat t.b.v de herhaalbaarheid over meerdere beoordelingssituaties. • Een kandidaat wordt altijd door minimaal twee verschillende beoordelaars beoordeeld. • Er is (nog) niet kwantitatief onderzocht hoe hoog de interbeoordelaarsbetrouwbaarheid is.
Acceptatie: Organisatoren, beoordelaars en kandidaten moeten zelf de PvB als praktisch bruikbaar ervaren (bv. formulieren, tools, etc.), en daarmee de PvB graag willen gebruiken
• De beoordelaars vinden de nieuwe PvB een grote vooruitgang ten opzichte van het voormalige examen. • De beoordelaars zijn vooraf uitgebreid getraind; dit bevordert de acceptatie en bruikbaarheid • De kandidaten accepteren de PvB, afgezien van enkele gevallen waarin bezwaar wordt aangetekend.
Transparantie: De PvB (bv. handleidingen, procedures, criteria, scoring) moet helder en begrijpelijk zijn voor organisatoren, beoordelaars en kandidaten. Er mogen geen onduidelijkheden in zitten.
• Het beoordelingsproces verloopt volgens een vast en duidelijk protocol; uitleg vindt plaats in een training. • De formulieren zijn helder van opzet maar de cesuurbepaling is complex en niet bekend bij de kandidaten. • De feedback over de examenuitslagen naar de kandidaten kan in sommige gevallen worden verbeterd.
Rechtvaardigheid: De PvB mag geen onderscheid maken tussen diverse groepen kandidaten of externe criteria (bv. niveau, opleidingslocatie, cultuur), maar mag uitsluitend de vereiste competenties meten.
• Er wordt geen onderscheid gemaakt tussen groepen kandidaten bijv. niveau of opleidingsinstantie. • Ondanks de individuele examinering kunnen buddyparen elkaar beïnvloeden net als in de praktijk.
Authenticiteit: De PvB moet de competenties die op de werkplek zijn vereist meten. Dit geldt voor de beoordelingstaak, de fysieke en sociale context, de beoordelingscriteria en het resultaat.
• De kandidaat wordt maximaal conform de echte praktijk geëxamineerd (scenario, teamsamenstelling). • De bevelvoerder moet er rekening mee houden dat hij een dubbelrol heeft omdat hij ook beoordelaar is.
Cognitieve complexiteit: De PvB, die op competenties is gebaseerd, moet voldoende cognitief complex zijn om bv. probleem oplossen te toetsen, in een praktijkgerichte beoordelingstaak en de juiste criteria
• De taken binnen het scenario zijn voldoende complex waarbij op competenties wordt beoordeeld. • In sommige gevallen is er sprake van diversiteit in werkstijlen afhankelijk van de opleiding en werkplek.
Relevantie: De uitkomsten van de PvB moeten voldoende relevant zijn voor de kandidaten, de opleiders en de praktijk, en voldoende differentiëren tussen de kandidaten.
• De examenresultaten geven informatie over vakbekwaamheid nodig om in de praktijk te gaan werken. • De ratingschaal kan leiden tot weinig differentiatie binnen de groep gezakte dan wel geslaagde kandidaten.

Tabel 3: Samenvatting resultaten geordend volgens evaluatieraamwerk (Baardman et al., 2006)

Deze kwalitatieve uitkomsten worden bij de afronding van het onderzoek (juli 2011) nog aangevuld met waarde oordelen van experts (plus – min).

Discussie en conclusies

De resultaten laten zien dat de betrouwbaarheid van de praktijksimulatie Brandbestrijding bij het Nbbe op een aantal aspecten goed is en op andere aspecten verbeterd kan worden. Hierbij kan onderscheid gemaakt worden tussen verbeteringen voor het *proces* en het *product*, zie Figuur 4.



Figuur 4. Proces versus product

Het *proces* betreft de manier waarop de examens zijn ingericht zoals het beoordelingsproces en de organisatie van de examens. Het *product* betreft de inhoud en materialen zoals de scenario's, de cesuur en de beoordelingsformulieren. Figuur 4 suggereert dat de kwaliteit van het proces sneller hoog is dan de kwaliteit van het product. Dit is meestal het geval voor beoordelingssystemen omdat verbeteringen aan het product pas gedaan kunnen worden nadat er praktijkervaring is opgedaan en kwantitatieve analyses zijn uitgevoerd. De resultaten leveren richtlijnen op voor een optimaal betrouwbaar ontwerp van dergelijke praktijkexamens die principe van toepassing zijn in alle domeinen vooral indien veiligheid een rol speelt, zoals luchtvaart, defensie of politie (Flin & Martin, 2001; O'Connor et al, 2002; Oprins et al, 2006).

Proces

Over het algemeen is het *proces* van examineren bij het Nbbe uitstekend ingericht. Voor de transfer naar de operationele praktijk is het van belang dat de 'authenticiteit' van de beoordeling maximaal is (e.g., Baartman et al., 2006; Berk, 1986). Bij de praktijksimulatie Brandbestrijding zijn de inzetten conform praktijk inclusief teamsamenstelling met vier manschappen en een bevelvoerder. De leden van de examencommissie hebben duidelijke rollen en elke kandidaat wordt door drie beoordelaars beoordeeld op twee scenario's; dit verhoogt de betrouwbaarheid. Het is echter altijd beter als er twee onafhankelijke beoordelaars zijn die een kandidaat op hetzelfde scenario beoordelen (e.g. Guion, 1998). De voorlopige resultaten (variantie-analyse) suggereren dat er verschillen bestaan tussen beoordelaars. Dit kan ook niet anders; in dergelijke praktijkexamens zijn persoonlijke tendenzen en beoordelingsfouten nauwelijks te vermijden (Roe & Daniels, 1994; Oprins, 2008). Een grondige training waarin aandacht wordt besteed aan de interpretatie van de criteria (frame-of-reference training; Bernardin & Buckley, 1981) zoals bij het Nbbe reeds wordt toegepast, draagt bij aan een verhoogde interbeoordelaarsbetrouwbaarheid. Ook kan worden overwogen om te werken met videoregistratie tijdens de examens zodat meerdere beoordelaars achteraf de kandidaat kunnen beoordelen, waarbij kandidaten zelf ook nog zaken kunnen terugzien. Dit bevordert de 'transparantie' (Baartman et al, 2006).

In juni 2011 wordt bij het Nbbe een experiment uitgevoerd om de interbeoordelaars-betrouwbaarheid nader te onderzoeken. In dit experiment gaan de beoordelaars een set videofragmenten beoordelen die praktijksimulaties tonen van verschillende kandidaten waarna zij onafhankelijk van elkaar een beoordelingsformulier invullen. De mate van overeenstemming tussen beoordelaars per criterium en over het geheel wordt berekend (Goldsmith & Johnson, 2002).

Product

De examens bij het Nbbe zijn al een stuk verbeterd dan voorheen: de beoordelingen zijn objectiever geworden met duidelijkere criteria op de formulieren. De producten kunnen op basis van verder onderzoek echter nog verder worden verfijnd en worden doorontwikkeld. Allereerst is het van belang om in dergelijke praktijkexamens de beoordelingstaak (scenario) te standaardiseren ten behoeve van de vergelijkbaarheid tussen kandidaten. Dit betekent overigens niet dat iedere kandidaat altijd hetzelfde scenario moet worden voorgelegd maar dat elk examen een basisset aan onderdelen zou moeten bevatten. Daarmee worden examens meer vergelijkbaar en herhaalbaar. Mogelijke varianten moeten gevalideerd worden om te verifiëren dat alle exameneisen aan bod komen en de mate van complexiteit gelijk blijft (e.g. Berk, 1986; Guion, 1998; Oprins, 2008).

De kwantitatieve onderzoeksresultaten bieden handvatten om de beoordelingsformulieren te verbeteren. Bij het Nbbe zijn de formulieren op een specifieke manier ingericht wat de resultaten verklaart. Uit de frequentietabellen is gebleken dat er een soort tweedeling ontstaat tussen de kandidaten: de gezakten zijn vrijwel allemaal direct gezakt op één criterium terwijl de geslaagden vrijwel uitsluitend voldoendes halen. Kandidaten zakken vooral op de criteria die betrekking hebben op het algehele operationele optreden. Ze zitten op een wat hoger abstractieniveau dan de andere criteria in het formulier. Het 'halo-effect' kan hiervoor een verklaring zijn: het generaliseren van bepaald gedrag over de algehele prestatie van de kandidaat (Thorndike, 1920; Roe & Daniels, 1994; Murphy & Cleveland, 1995), in dit geval de criteria waarop je direct kan zakken omdat deze betrekking hebben op algeheel presteren. Voor wat betreft de geslaagden ontstaat het gevaar op het 'plafond-effect' omdat zij vrijwel uitsluitend voldoendes hebben gescoord; dan wordt er minder onderscheid gemaakt tussen de goede en iets minder goede kandidaten. Meer nuancering kan informatie bieden over mogelijke aandachtspunten in de operationele praktijk. Het werken met kardinale onvoldoendes, doorgaans 'knock-outs' genoemd, is echter wel degelijk zinvol, zeker in veiligheidsfuncties. Ze komen dan ook vaak voor in dergelijke examens (Oprins, 2008). Een oplossing is om de knock-outs maximaal te objectiveren worden door ze te koppelen aan concrete en observeerbare veiligheidsfouten.

Een andere oorzaak voor deze effecten ligt bij de tweepuntsschaal. In de methodologische discussies die worden gevoerd wordt aangegeven dat statistisch gezien deze ratingschaal nadelen oplevert voor de betrouwbaarheid (Preston & Colman, 2000; Lozano, Garcia-Cueto & Muniz, 2008). Het gevaar bestaat dat de spreiding laag is waardoor er niet goed gedifferentieerd kan worden tussen kandidaten. In praktijkexamens zoals bij het Nbbe zou daarom het gebruik van een 4- of een 6-puntsschaal beter zijn. Bij een even schaal kan nog steeds strak onderscheid gemaakt worden tussen onvoldoende en voldoende prestaties op bepaalde competenties waar beoordelaars over het algemeen behoefte aan hebben (Oprins et al., 2006).

Bovendien kunnen de kwantitatieve resultaten gebruikt worden om de indeling in rubrieken en beoordelingscriteria te verbeteren. De Cronbach's alpha waren erg laag en dit betekent dat de samenhang verbeterd moet worden (Schmidtt, 1996). De clusters die uit de factoranalyses naar voren gekomen zijn kunnen gebruikt worden voor een nieuwe indeling. Doorgaans wordt aanbevolen om in praktijkexamens op competenties te beoordelen. Elke competentie bevat een vergelijkbare set van ongeveer 3 tot 6 gedragscriteria die zo concreet en observeerbaar mogelijk geformuleerd moeten worden samen met inhoudsdeskundigen (e.g. Flin & Martin, 2001; O'Connor et al, 2002; Oprins et al, 2006). Dit is een langdurig proces van verbetering op basis van kwantitatieve analyses en praktijkervaring. Hiervoor is het van belang dat de data gedigitaliseerd zijn. Validatieonderzoek is

nodig om te verifiëren of de criteria inhoudelijk voldoende dekkend zijn (Roe, 1995; Oprins, 2008). Zo wordt bij het Nbbe binnenkort een experiment gedaan om de predictieve validiteit van de praktijksimulatie te onderzoeken. Hierbij worden de resultaten gerelateerd aan prestaties in de operationele werkpraktijk. Ook de cesuur kan op deze manier worden verfijnd door aandacht te besteden aan de (on)gewenste 'valse positieven' en 'valse negatieven' (Roe, 2005). Uiteindelijk leidt dit tot een optimaal valide en betrouwbaar praktijkexamen waarmee functiegroepen waarbij fouten cruciale en onherroepbare gevolgen kunnen hebben en 'valse positieven' absoluut vermeden moeten worden, beoordeeld kunnen worden met meer zekerheid.

Referenties

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 in Education.

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

Berk, R.A. (1986). *Performance assessment: methods and applications.* Baltimore: The John Hopkins University.

Bernardin, H.J., & Buckley, M.R. (1981). A consideration of strategies in rater training. *Academy of management review, 6,* 205-212.

Clark, L.A., & Watson, D.(1995). Constructing validity: basic issues in objective scale development. *Psychological assessment, 7*(3), 309-319.

Flin, R. & Martin, L. (2001). Behavioral markers for crew resource management: a review of current practice. *The international journal of aviation psychology, 11,* 95-118.

Goldsmith, T.E., & Johnson, P.J. (2002). Assessing and improving evaluation of aircrew performance. *Human Factors, 12*(3), 223-240.

Guion, R.M. (1998). *Assessment, measurement and prediction for personnel decisions.* Mahwah NJ: Lawrence Erlbaum Associates.

Lozano, L.M., Garcia-Cueto, E., & Muniz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology,* 4(2), 73-79.

Murphy, K.R., & Cleveland, J.N. (1995). *Understanding performance appraisal: social, organizational and goal-based perspectives.* London: Sage Publications.

O'Connor, P., Hormann, H, Flin, R, Lodge, M., & Goeters, K. (2002). Developing a method for evaluating crew resource management skills: a European perspective. *The international journal of aviation psychology, 12*(3), 263-285.

Oprins, E., Burggraaff, E., & Weerdenburg, H. van (2006). Design of a competence-based assessment system for air traffic control training. *The international journal of aviation psychology, 16*(3), 297-320.

Oprins, E., Burggraaff, E., & Weerdenburg, H. van (2008). Reliability of assessors' ratings in competence-based air traffic control training. *Human factors and aerospace safety, 6* (4), 305-322.

Oprins, E. (2008). *Design of a competence-based assessment system for air traffic control training.* Doctoral dissertation, University of Maastricht.

Preston, C.C., & Colman, A.M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta psychologica, 104,* 1-15.

Roe, R.A, & Daniels, M.J.M. (1994). *Personeelbeoordeling: achtergrond en toepassing* (3th ed.). Assen: Van Gorcum.

Roe, R.A. (2005). The design of selection systems: contexts, principles, issues. In A. Evers, O. Smit & N. Anderson (Eds.), *Handbook of personnel selection* (pp. 73-97). Oxford: Blackwell.

Schmidtt, N. (1996). Uses and abuses of coefficient alpha. *Psychological assessment, 8*(4), 350-352.

Theunissen, N.C.M., Meulman, J.J., den-Ouden, A.L., Koopman, H.M., Verrips, G.H., Verloove-Vanhorick, S.P., & Wit, J.M. (2003). Changes can be studied when the measurement instrument is different at different time points. *Health Services and Outcomes Research Methodology, 4* (2), 109-126

Thorndike, E.L. (1920). A constant error in psychological ratings. *Journal of applied psychology, 4,* 25-29.