BEOORDELINGSSCHALEN IN COMPETENTIEGERICHTE EXAMINERING

Andere **schalen**, andere **oordelen?**

Hans Kuhlemeier en Tom Erkens

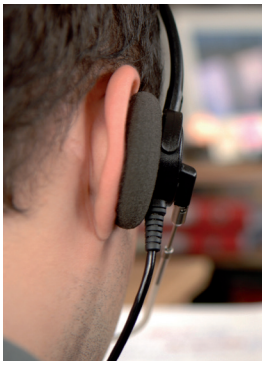
Met de opkomst van competentiegericht onderwijs hebben beoordelingsschalen hun intrede gedaan in de landelijke beroepsgerichte examens. De bruikbaarheid van beoordelingen bij competentiegerichte opdrachten wordt vaak beperkt door beoordelingsfouten zoals de toegeeflijkheidsfout en een gebrekkige discriminatie tussen vaardige en minder vaardige kandidaten. In dit artikel is nagegaan in hoeverre het type beoordelingsschaal van invloed is op de hoogte en spreiding van de beoordelingen.

Inleiding

In het kader van personeelsbeoordeling wordt tegenwoordig veelvuldig gebruikgemaakt van beoordelingsschalen. Aan de beoordelingen die leidinggevenden aan hun ondergeschikten toekennen, zit vaak een luchtje. Het is niet ongebruikelijk dat 80 à 90 procent van de medewerkers als bovengemiddeld bekwaam beoordeeld wordt; de verschillen tussen de beoordelingen die medewerkers ontvangen, zijn meestal veel kleiner dan de werkelijke prestatieverschillen rechtvaardigen (Murphy & Cleveland, 1995). Herhaaldelijk is aangetoond dat beoordelaars ertoe neigen negatieve oordelen te vermijden, weinig spreiding aan te brengen tussen beoordeelden en zelfs kritische fouten onopgemerkt te laten. In veel beoordelingssituaties zijn hiervoor twee bekende beoordelingsfouten verantwoordelijk: de toegeeflijkheidsfout en de fout van *range restriction*. De toegeeflijkheidsfout is de neiging om prestaties van anderen systematisch positiever te beoordelen dan deze in werkelijkheid zijn. *Range restriction* is de neiging om minder spreiding aan te brengen tussen goede en minder goede kandidaten dan de werkelijke prestatieverdeling rechtvaardigt. Niet altijd zijn beoordelingsfouten gebaseerd op onkunde. In een overzicht van onderzoek naar personeelsbeoordelingen maken Murphy en Cleveland (1995) bijvoorbeeld aannemelijk dat leidinggevenden hun ondergeschikten niet accuraat beoordelen omdat zij dat niet kunnen, maar omdat zij dat niet willen.

Overwegend positieve oordelen en weinig spreiding

Tot voor kort kwamen beoordelingsschalen in de centrale examinering in het voortgezet onderwijs nog nauwelijks voor. Met de komst van het competentiegerichte Centraal Schriftelijk en Praktisch Examen (CSPE) voor het vmbo is hierin verandering gekomen. In deze praktijkexamens worden vakkennis, vakvaardigheden, algemene vaardigheden en beroepshoudingen geïntegreerd getoetst. De examinator observeert het handelen van de kandidaten in authentieke, zij het gesimuleerde praktijksituaties. De handelingen en producten van de kandidaten zijn meestal niet objectief en eenduidig scoorbaar, maar moeten beoordeeld worden, waarbij verschillende examinatoren aan eenzelfde prestatie een verschillende score kunnen toekennen.
Het niveau en de complexiteit van de examenopdrachten is afgestemd op hetgeen van een gemiddelde kandidaat na vier jaar vmbo-onderwijs verwacht mag worden. De examenverslagen laten echter zien dat de nieuwe CSPE's voor het vmbo vaak uitzonderlijk goed gemaakt worden. De toets- en itemanalyses maken duidelijk dat de verdelingen van de met schalen beoordeelde examenprestaties over het algemeen sterk negatief scheef zijn (Kuhlemeier, 2005). Dit wil zeggen: het gemiddelde ligt ver boven het (veelal neutrale) midden van de beoordelingsschaal en de spreiding van de prestaties rond dat gemiddelde is gering. Examenopdrachten waarbij vrij-



Beoordelaars neigen ertoe negatieve oordelen te vermijden

wel alle kandidaten vergelijkbaar hoge beoordelingen ontvangen, kunnen moeilijk als objectief, nauwkeurig en rechtvaardig worden beschouwd.

Ongelabelde schaal, checklist en rubriek

Aan geïnflateerde en weinig spreidende beoordelingen kunnen velerlei oorzaken ten grondslag liggen, zoals ambigue beoordelingsmiddelen, ongetrainde of ongemotiveerde beoordelaars of een beoordelingscontext die weinig stimulans biedt voor een objectieve en eerlijke beoordeling. Ter bestrijding van mogelijke beoordelingsfouten staan de examenmaker beperkte middelen ter beschikking (Kuhlemeier & Béguin, 2005). Een mogelijke maatregel is het verbeteren van de beoordelingsmiddelen.
Onderzocht is in hoeverre de hoeveelheid sturing die de beoordelingsschaal biedt, samenhang vertoont met de hoogte en spreiding van de examenprestaties. Hierbij is de verwachting dat de oordelen minder positief zijn en beter spreiden tussen kandidaten naarmate de beoordelingsschaal de vrijheid van de examinator sterker aan banden legt. Onder constanthouding van de examenopdracht en het beoordelingsobject zijn in het onderzoek drie schaaltypen met elkaar vergeleken:

- holistische beoordeling met een ongelabelde schaal;
- analytische beoordeling met een checklist met vijf dichotoom te scoren aandachtspunten;
- holistische beoordeling van dezelfde vijf aandachtspunten met een beschrijvende rubriek.

Figuur 1 geeft van elk schaaltype een voorbeeld.
De drie schaaltypen verschillen in de hoeveelheid steun die de examinator geboden wordt. Met de ongelabelde schaal kan de examinator vrijwel geheel naar eigen inzicht 0, 1, 2, 3, 4 of 5 punten toekennen. De checklist biedt de examinator al wat meer steun. Anders dan de ongelabelde schaal, geeft de checklist de aspecten aan waar de examinator bij het beoordelen op moet letten. De examinator heeft echter veel vrijheid om te bepalen of hij of zij de kandidaat telkens 0 of 1 punt toekent. Cognitieve interviews met praktijkdocenten en analyse van examengegevens

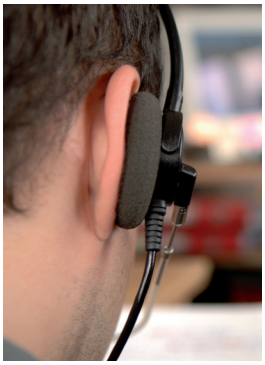
Voorbeeld van een ongelabelde schaal

Voorbeeld van een checklist

Voorbeeld van een rubriek

Figuur 1
Een ongelabelde schaal, een checklist en een rubriek

De hoogte en spreiding van de examenprestaties hangt samen met de sturing die de beoordelingsschaal biedt



doen vermoeden dat examinatoren deze vrijheid regelmatig benutten om kandidaten 'het voordeel van de twijfel' te geven en hoge beoordelingen te geven (Laheij, 2004; Kuhlemeier, 2005). Met een rubriek geeft de examinator één samenvattend kwaliteitsoordeel aan de hand van meerdere criteria tegelijkertijd. In tegenstelling tot zowel de ongelabelde schaal als de checklist zijn de onderscheiden prestatieniveaus van een verhelderende toelichting voorzien. Doordat het beoordelingsobject en de criteria geëxpliciteerd zijn, zouden rubrieken een positieve bijdrage leveren aan het bestrijden van beoordelingsfouten (Moskal & Leydens, 2000).
Verwacht wordt dus dat examinatoren met gebruik van een rubriek lagere beoordelingen toekennen en meer spreiding tussen kandidaten aanbrengen dan met een ongelabelde schaal, waarbij de checklist een middenpositie zal innemen.

Methode van onderzoek

Het onderzoek is uitgevoerd in het najaar van 2006 bij kandidaten die aan het eind van dat schooljaar deelnamen aan het praktijkexamen 'Handel en administratie/verkoop'. Zeventig praktijkdocenten beoordeelden de prestaties van ruim achthonderd examenkandidaten op drie competentiegerichte examenopdrachten: een telefoongesprek met een klant, een reclamebord ontwerpen op de computer en een klant adviseren. Het onderzoek is uitgevoerd in het kader van het schoolexamen of een gesimuleerde eindexamensetting. Voor verdere details over de examenopdrachten, de onderzoeksopzet, de afname- en beoordelingsomstandigheden en de statistische analyse wordt verwezen naar het hoofdrapport (Kuhlemeier & Erkens, 2008).

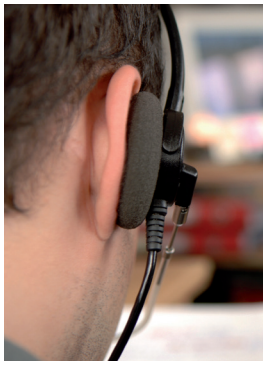
Resultaten

Om de drie schaaltypen te kunnen vergelijken zijn de oordelen van de praktijkdocenten omgezet naar een schaal van 0 tot 100 punten. De checklist leidt tot de meeste positieve beoordelingen, gevolgd door de ongelabelde schaal en tenslotte de rubriek. Met de checklist kennen de examinatoren de kandidaten gemiddeld 73 van de 100 punten toe, met de schaal 66 punten en met de rubriek 61 punten. De tabel in figuur 2 laat zien dat het verschil tussen de drie schaaltypen niet helemaal stabiel is over de drie examenopdrachten. Bij de opdracht 'Ontwerpen reclamebord' zijn de drie gemiddelden statistisch gezien namelijk niet onderscheidbaar. Een aannemelijke verklaring is het procesmatige karakter van de beoordelingsaspecten bij deze opdracht. De examinatoren moesten hier namelijk beoordelen in hoeverre de tekst duidelijk en leesbaar was en het ontwerp goed opviel door de gebruikte letters, de vlakverdeling en de vorm.

Overeenkomstig de verwachting zouden examinatoren met de rubriek een beter onderscheid maken tussen kandidaten dan met een checklist die op zijn beurt weer beter zou spreiden dan een ongelabelde schaal. De varianties blijken te variëren van 401 voor het oordeel over de kwaliteit van het telefoongesprek met een checklist tot 843 voor het oordeel met datzelfde beoordelingsinstrument voor de kwaliteit van het reclamebord (zie figuur 2). Met de rubriek brengen examinatoren gemiddeld de meeste spreiding aan tussen kandidaten, gevolgd door de checklist en tenslotte de ongelabelde schaal. De varianties voor de schaal, checklist en rubriek bedragen gemiddeld respectievelijk 477, 599 en 737. De veel grotere spreiding van de rubriek is des te opmerkelijker als men bedenkt dat het aantal schaalpunten bij de rubriek

Opdracht en beoordelingsaspect	**Gemiddelde beoordeling**	**Variantie**
Voeren telefoongesprek		
Ongelabelde schaal	67,58 (1,58)	506,9 (50,9)
Checklist	77,48 (1,44)	401,4 (41,0)
Rubriek	57,00 (1,89)	765,9 (74,9)
Ontwerpen reclamebord		
Ongelabelde schaal	63,92 (1,35)	455,9 (40,8)
Checklist	67,33 (1,85)	843,3 (76,0)
Rubriek	64,30 (1,57)	634,6 (56.0)
Adviseren klant		
Ongelabelde schaal	65,14 (1,42)	469,2 (43,7)
Checklist	74,87 (1,62)	551,9 (54,3)
Rubriek	62,08 (1,94)	809,6 (78,6)

Figuur 2 Gemiddelde beoordeling en variantie per examenopdracht per beoordelingsmiddel (tussen haakjes: standaardfouten)



Beoordelen met een rubriek is het meest objectief en eerlijk

kleiner is dan bij de andere twee schaaltypen (en de variantie doorgaans toeneemt naarmate het aantal schaalpunten groter is).
De examinatoren is gevraagd de drie schaaltypen op grond van hun ervaring te prioriteren. Volgens de examinatoren bieden de rubriek en de checklist meer steun bij het beoordelen dan de ongelabelde schaal, maar kost de beoordeling wel meer tijd en energie. De rubriek wordt als het meest objectief en eerlijk ervaren.

Discussie

Praktijkdocenten blijken met het ene type beoordelingsschaal duidelijk anders te beoordelen dan met het andere type. Van de drie onderzochte beoordelingsschalen blijkt de rubriek de meest gunstige psychometrische eigenschappen te bezitten: het gemiddelde ligt het dichtst tegen het midden van de schaal en de spreiding is het grootst. Bovendien geeft de rubriek de examinatoren naar eigen zeggen veel steun en is de beoordeling het meest eerlijk en objectief (al kost de beoordeling meer tijd en energie). Op dit moment wordt dit type rubrieken in de praktijkexamens voor het vmbo nog niet toegepast. Het verdient aanbeveling een begin te maken met de introductie van rubrieken en examenmakers te ondersteunen bij de ontwikkeling ervan.
In vergelijking met de rubriek blijken de checklist en de ongelabelde schaal te leiden tot relatief hoge beoordelingen en weinig spreiding tussen kandidaten. Op grond hiervan verdient het aanbeveling deze schaaltypen wat minder vaak te gebruiken voor de beoordeling van beroepsgerichte competenties.
De mate waarin deze onderzoeksresultaten generaliseerbaar zijn is onzeker. Om begrijpelijke redenen is het onderzoek niet uitgevoerd in het 'echte' centrale praktijkexamen, maar in het schoolexamen of een gesimuleerd centraal praktijkexamen (vgl. Kuhlemeier & Erkens, 2008). De beoordelingscondities in het onderzoek zijn mogelijk niet volledig vergelijkbaar met die van een centraal praktijkexamen. Daar staat er voor examinatoren en kandidaten meer op het spel en is de verleiding tot soepel beoordelen waarschijnlijk groter. In dat geval zouden deze resultaten een overschatting zijn van de mogelijkheden om de toegeeflijkheidsfout en een gebrekkige discriminatie tussen kandidaten met alternatieve beoordelingsmiddelen tegen te gaan.

Literatuur

Kuhlemeier, H. (2005). *Inventarisatie van beoordelingsmiddelen in twintig vmbo-praktijkexamens 2005*. Arnhem: Cito.

Kuhlemeier, H. & Béguin, A. (2006). *Over de maakbaarheid en landelijke vergelijkbaarheid van de centrale vmbo-praktijkexamens: Een discussienota voor de Cevo*. Arnhem: Cito.

Kuhlemeier, H. & Erkens, T. (2008). *Effect van holistische, analytische en descriptieve schalen op de oordelen van examinatoren in een competentiegericht praktijkexamen voor het vmbo*. Arnhem: Cito.

Laheij, E. (2004). *Inventarisatie beoordelingspraktijk CSPE '04*. Verslag van een kwalitatief onderzoek bij docenten (doctoraalscriptie onderwijskunde). Arnhem: Cito.

Moskal, B.M. & Leydens, J.A. (2000). Scoring rubric development: Validity and reliability. Practical Assessment, *Research & Evaluation*, 7, 10. Online beschikbaar via http://PAREonline.net/getvn.asp?v=7&n=10.

Murphy, K.R. & Cleveland, J.N. (1995). *Performance appraisal: An organizational perspective*. Boston: Allyn and Bacon.

☐ De heren dr. J.B. Kuhlemeier en drs. T.T.M.G. Erkens zijn respectievelijk als onderzoeker en trainer werkzaam bij Cito.