B. Meuffels

beoordelen

GLOBAAL VERSUS ANALYTISCH BEOORDELEN

In Moer *1984/4 werd geschreven over beoordelen met behulp van een studiepuntensysteem. In datzelfde jaar (*Moer *1984/6) kwam een andere oplossing voor beoordelen aan de orde (een evaluatieplan en een puntensysteem).*
In dit nummer reflecteert Bert Meuffels over globaal en analytisch beoordelen van opstellen. Zijn stelling is dat een analytisch schema niet tot een hogere betrouwbaarheid leidt. Daar probeert hij tevens een psychologische verklaring voor te geven.

De betrouwbaarheid van opstelbeoordeling

Een docent in het moedertaalonderwijs die de schrijfvaardigheid van zijn leerlingen beoogt te vergroten, zal die leerlingen regelmatig schrijfopdrachten verstrekken en zal de proeven van kunnen op het gebied van schrijven — opstellen, scripties, nota's, werkstukken en dergelijke — corrigeren en beoordelen. Hij beoordeelt niet omdat hij met alle geweld een oordeel wil vellen over die ene unieke prestatie van een leerling, maar omdat hij door middel van dat oordeel — mits het voldoende specifiek en gedifferentieerd is — de leerling kan informeren over zijn sterke en zwakke punten, fouten en tekortkomingen. Gewapend met die kennis kan de leerling zijn prestaties gericht verbeteren.
Natuurlijk kan de leerling zijn schrijfvaardigheid op grond van de gegeven feedback alleen maar verbeteren wanneer het door de docent geleverde commentaar op het schrijfprodukt voldoende *betrouwbaar* is. Uit empirisch onderzoek is echter bekend dat de betrouwbaarheid van de beoordeling van schrijfvaardigheid vaak te wensen overlaat (Meuffels 1983). Voor de leerling wiens schrijfvaardigheid beoordeeld wordt, is dat een frustrerende zaak. Hij zal zich immers onrechtvaardig behandeld voelen als blijkt dat het cijfer dat zijn docent aan zijn schrijfprodukt heeft toegekend, sterk afwijkt van het cijfer van een andere docent die dat produkt ook heeft beoordeeld. En een leerling zal evenzeer vreemd opkijken wanneer bij een herbeoordeling van zijn opstel door dezelfde docent een heel ander cijfer uit de bus komt. Deze twee situaties zijn karakteristiek voor een onbetrouwbare beoordeling: 1. de cijfers die dezelfde beoordelaar bij herbeoordeling aan dezelfde opstellen toekent, lopen uiteen (deze vorm van betrouwbaarheid staat bekend onder de naam 'beoordelaarsstabiliteit'); 2. de cijfers die twee of meer beoordelaars aan dezelfde opstellen toekennen, lopen uiteen (en deze vorm van betrouwbaarheid staat bekend als 'interbeoordelaarsovereenstemming').

Oordelen over schrijfprodukten moeten in voldoende mate betrouwbaar zijn. Wat blijft er immers over van de pretentie de schrijfvaardigheid van de leerlingen te verbeteren wanneer een leerkracht een opstel een voldoende geeft, terwijl hij datzelfde opstel twee weken daarvóór als onvoldoende heeft gekwalificeerd? Bij een onbetrouwbare beoordeling dient men een groot vraagteken achter elke beoordeling te zetten, en alle praktische beslissingen die men op zo'n beoordeling baseert zijn in hoge mate tentatief en speculatief. De betrouwbaarheid van schrijfvaardigheidsbeoordeling mag dan vaak te wensen over laten, we beschikken volgens sommige onderzoekers wél over kennis en inzicht in methoden die die betrouwbaarheid kunnen verhogen. Zo zal volgens Volovics-Schelvis de betrouwbaarheid (dat wil zeggen én de beoordelaarsstabiliteit én de interbeoordelaarsovereenstemming) toenemen wanneer we de schrijfvaardigheid niet op de gebruikelijke *globale*, maar op *analytische* wijze beoordelen.

Bij een globale beoordelingswijze beoordeelt de beoordelaar een werkstuk als geheel, zonder analyse van het schrijfprodukt in afzonderlijke aspecten. Hij kent in de regel oordelen toe op basis van zijn eigen, niet-geëxpliciteerde normen; hij beoordeelt de opstellen kortom op grond van een relatief ongedifferentieerde totaalindruk, resulterend in één cijfer. Anders dan bij globale beoordeling wordt bij analytische beoordeling de 'overall' kwaliteit van een opstel opgesplitst in een aantal deelaspecten (zoals stijl, spelling, woordkeuze, opbouw), en elk opstel moet op elk van de onderscheiden deelaspecten apart beoordeeld worden. Het analytisch eindoordeel bestaat meestal uit het gemiddelde van alle oordelen op de in het analytisch schema onderscheiden deelaspecten. 'De verwachting is dat het gebruik van een (...) (analytisch) schema de beoordeling meer betrouwbaar (...) zal maken (...). Met meer betrouwbaar bedoelen we hier zowel dat één beoordelaar eenzelfde produkt bij herhaalde beoordeling hetzelfde zal waarderen (...) als dat meerdere beoordelaars hetzelfde produkt met grotere onderlinge overeenstemming zullen waarderen' (Volovics-Schelvis 1979, p. 75).

Het citaat over de positieve invloed van een analytisch schema op de betrouwbaarheid is betrekkelijk willekeurig gekozen uit nationale en internationale publikaties over opstelbeoordeling; heel vaak treft men de in bovenstaand citaat uitgesproken claim in de literatuur aan.

Maar heeft het gebruik van een analytisch schema wel het beoogde effect? Neemt de betrouwbaarheid van opstelbeoordeling inderdaad toe wanneer we opstellen beoordelen aan de hand van een analytisch schema? In dit artikel zal ik mede aan de hand van empirische gegevens uit de onderzoeksliteratuur illustreren dat een analytisch schema niet tot een hogere betrouwbaarheid leidt. Tevens zal ik een psychologische verklaring hiervoor geven.

Een plausibele veronderstelling?

Het ligt alleszins in de rede om te veronderstellen dat het gebruik van een analytisch schema de schrijfvaardigheidsbeoordeling betrouwbaarder maakt. Zo'n schema immers verscherpt de beoordelingstaak door een specificatie van die aspecten van het schrijfprodukt waarop de beoordelaar telkens moet letten.

Bij nadere beschouwing blijkt het gebruik van een analytisch schema toch problematischer te zijn dan in eerste instantie wellicht lijkt. Ook al dwingt een vereenvoudiging c.q. explicitering van de beoordelingstaak in een aantal specifieke kenmerken de beoordelaar tot concentratie op telkens één specifiek aspect, feit blijft dat die afzonderlijke aspecten *beoordeeld* moeten worden en dat de oordelen over die aspecten per opstel op de een of andere wijze tot een eindoordeel worden *gecombineerd*. Het probleem van de beoordeling van een complex geheel ('de' kwaliteit van een schrijfprodukt) wordt bij een analytisch schema in feite verschoven naar het probleem van de beoordeling van afzonderlijke kenmerken (welke, en hoe?) én van het combineren van die afzonderlijke oordelen tot een eindoordeel.

Voor dat laatste problematische aspect — het combineren van verschillende oordelen tot een eindoordeel — hebben de propagandisten van een analytisch schema een wel heel erg simpele oplossing te bieden: tel de oordelen over de aspecten bij elkaar op en middel deze. Deze in de praktijk gangbare gedragswijze druist paradoxaal genoeg in tegen het typische karakter van een analytisch schema. Bij de constructie van zo'n schema gaat men er immers vanuit dat de 'overall' kwaliteit van een schrijfprodukt opgesplitst kan worden in een aantal min of meer *onafhankelijke* deelaspecten — maar zijn die deelaspecten écht onafhankelijk, dan mag men ze natuurlijk niet als appels en

peren bij elkaar optellen en middelen. Zo'n gemiddelde zegt dan niets meer.
Een voorbeeld. Leerling A heeft voor de categorie Spelling een 4, voor de categorie Stijl een 8 behaald, terwijl voor leerling B precies het omgekeerde geldt. Beide leerlingen behalen na middeling hetzelfde eindcijfer. In feite zijn beide eindcijfers incompatibel, want de ene 6 verwijst naar geheel andere schrijfkwaliteiten dan de andere 6. Het verwijt van de propagandisten van een analytisch schema tegen de gebruikers van een globale beoordelingsmethode, dat deze een mysterieuze beoordelingswijze is waarbij het oordeel privéopvattingen weerspiegelt die verborgen blijven in het hoofd van de beoordelaar, kortom het feit dat je niet meer weet waarnaar zo'n oordeel verwijst, verliest veel van zijn kracht als men beseft dat door het optellen en middelen van afzonderlijke analytische oordelen tot een eindoordeel de analytische beoordelingswijze precies hetzelfde verwijt treft.

Een mogelijke verklaring voor het falen van analytische schema's

Resultaten van empirisch onderzoek tonen overduidelijk aan dat een analytische beoordeling niet of nauwelijks betrouwbaarder is dan een globale beoordeling (Morrison & Vernon 1941; Lamb 1953; Nisbet 1955; Volovics-Schelvis 1979; Zondervan 1979; Zijlmans & Blok 1980; Meuffels 1985). Deze uitspraak is echter gebaseerd op de in de literatuur gemaakte vergelijking van de betrouwbaarheid van analytische *eind*oordelen met de betrouwbaarheid van globale oordelen. Het optellen van afzonderlijke analytische oordelen en het middelen daarvan tot een analytisch eindoordeel is, zoals gezegd, in zeer veel gevallen een dubieuze procedure. De betrouwbaarheid van globale oordelen moet vergeleken worden met de betrouwbaarheid van een analytisch oordeel per *afzonderlijke* categorie (zoals spelling, opbouw, oorspronkelijkheid, enzovoort). En uit deze vergelijking blijkt dat een globale beoordeling door de bank genomen zelfs superieur is aan een analytische beoordeling!
Een in de literatuur vaak genoemde verklaring voor het falen van analytische schema's (falen, voor zover het de verhoging van de betrouwbaarheid betreft), is, dat de aanduidingen en eventuele omschrijvingen van de beoordelingscategorieën waaruit analytische schema's zijn opgebouwd, vaak ondoorzichtig en meerduidig zijn. Vaak slecht gedefinieerde categorieën als Stijl, Inhoudelijke Structuur, bieden de beoordelaar relatief weinig steun, zodat deze in zijn cijfergeving grotendeels op privé-normen en -interpretaties is aangewezen (cf. Wesdorp 1981, p. 59-61). Nominaal mogen analytische en globale beoordeling dan wel fundamenteel van elkaar verschillen, in de praktijk komen ze vaak op hetzelfde neer.
In recent empirisch onderzoek blijkt men echter wel degelijk op de hoogte van de gevaren van conceptuele en terminologische vaagheid van de beoordelingscategorieën. Om de interpretatie van die categorieën te verduidelijken en te uniformeren voorziet men de beoordelingscategorieën van een uitgebreide verbale toelichting, en soms wordt de cijfergeving voor een categorie bovendien nog verduidelijkt aan de hand van voorbeeldopstellen. Niettemin vertonen de resultaten van empirisch onderzoek waarin de hier genoemde voorzorgsmaatregelen getroffen zijn, hetzelfde voor het analytisch schema teleurstellende beeld.
Samenvattend kan gesteld worden dat analytische schema's de beoordelingstaak beogen te expliciteren en te vereenvoudigen. Die explicitering en vereenvoudiging, zo wordt algemeen aangenomen, dient een forse betrouwbaarheidswinst op te leveren (zowel wat betreft de beoordelaarsstabiliteit als de interbeoordelaarsovereenstemming). Uit de onderzoeksliteratuur blijkt echter dat analytische schema's niet betrouwbaarder oordelen opleveren dan de globale beoordelingswijze. De ter verklaring hiervan gesignaleerde conceptuele en/of terminologische vaagheid van de beoordelingscategorieën kan wellicht sommige (met name oudere) onderzoeksresultaten verklaren, maar zeker niet alle (en met name niet de meest recente). Een afdoende verklaring voor het falen van analytische schema's moet elders gezocht worden.

Een psychologische verklaring

Een verklaring voor het falen van analytische schema's moet niet zozeer gezocht worden in onvolkomenheden van die schema's zelf, maar veeleer in de manier waarop mensen opstellen beoordelen — of die beoordeling nu op analytische dan wel op globale wijze plaatsvindt.
De kern van de hier voorgestelde psychologische verklaring is dat beoordelaars opstellen niet zozeer *dimensioneel*, als wel *typologisch* beoorde-

len. Typologisch beoordelen is de 'normale', 'natuurlijke' wijze van beoordelen zoals die bij een globale beoordelingswijze plaatsvindt. Een analytisch schema dwingt beoordelaars af te zien van die normale vorm van beoordelen, dwingt hen opstellen dimensioneel te beoordelen, iets waartoe deze niet – althans niet zonder intensieve training – in staat zijn.
De begrippen 'typologisch' en 'dimensioneel' kunnen verduidelijkt worden aan de hand van de wijze waarop mensen elkaar beoordelen, dat wil zeggen elkaar persoonlijkheidseigenschappen toekennen. Stel dat iemand gevraagd wordt zo nauwkeurig en informatief mogelijk drie hem bekende personen (A, B en C) te beschrijven met behulp van de persoonlijkheidseigenschap 'vrolijk'-'somber'. A wordt als een vrolijke jongen gekarakteriseerd, B als een sombere en C als een meisje van ... ja van wat eigenlijk? Van 'gemiddelde' vrolijkheid? Maar wat moet je je daar bij voorstellen?
Door het toekennen van de eigenschap 'vrolijk' (of 'somber') wordt in feite een voorspelling gedaan dat A respectievelijk B zich in de meeste situaties naar alle waarschijnlijkheid vrolijk respectievelijk somber zal gedragen. Maar dat we C met behulp van diezelfde *dimensie* als gemiddeld vrolijk karakteriseren draagt in feite niets bij tot ons vermogen haar gedrag specifiek te voorspellen. Als persoonlijkheidseigenschap verwijst een begrip als 'vrolijkheid/somberheid' niet zozeer naar (de uiteinden van) een *dimensie*, als wel naar een *type*. Als C tot geen van beide typen behoort, is het zaak haar met andere begrippen te beschrijven (cf. Bem & Allen 1974).
Uiteraard is het voorbeeld fictief. In werkelijkheid beoordelen mensen elkaar niet met behulp van slechts één eigenschap, maar met behulp van een aantal persoonlijkheidseigenschappen zoals extravert, intelligent, aardig, standvastig, agressief, enzovoort. Als we personen dimensioneel zouden beoordelen, zou dat betekenen dat we aan alle te beoordelen personen telkens alle persoonlijkheidseigenschappen zouden toekennen waarvan we bij onze persoonsbeschrijvingen gebruik maken, zij het dat de *mate* waarin ze die eigenschappen bezitten, per persoon verschilt. Als we personen echter typologisch zouden beoordelen, dan zouden we juist niet alle eigenschappen – in variërende mate – aan alle te beoordelen personen toekennen, maar verschillende *soorten* eigenschappen aan verschillende personen. Zo zouden we de een bijvoorbeeld aardig noemen, terwijl alle andere persoonlijkheidseigenschappen irrelevant zijn voor een juiste typering van de betreffende persoon, de ander agressief, weer een ander standvastig, enzovoort.
Onderzoek heeft aannemelijk gemaakt dat mensen elkaar niet dimensioneel beoordelen, maar typologisch: verschillende soorten eigenschappen zijn relevant voor het beschrijven van verschillende personen (Barendregt 1978; 1979). Deze gedachtengang kan toegepast worden op opstelbeoordeling. In het ene opstel treft ons de stijl – en alle andere eventueel in de beoordeling te verdisconteren aspecten zijn irrelevant voor een treffende karakterisering van dat opstel – in een ander opstel dringt de oorspronkelijkheid ervan zich op, enzovoort. Het zijn juist deze saillante, per opstel variërende karakteristieken waarop we ons globale oordeel voornamelijk baseren. Wanneer dergelijke saillante karakteristieken in een opstel ontbreken, is de onzekerheid over de juistheid van het oordeel maximaal. Het beoordelen van opstellen is niet zozeer moeilijk doordat binnen elke beschrijvingswijze zich tussen opstellen grote verschillen voordoen, maar doordat verschillende opstellen de beoordelaar uitnodigen tot verschillende beschrijvingswijzen.

Het effect van typologisch beoordelen op de betrouwbaarheid van een analytisch schema

Een analytisch schema dwingt de beoordelaar te abstraheren van zijn 'natuurlijke' manier van beoordelen, het typologisch beoordelen; een analytisch schema dwingt de beoordelaar tot een vorm van dimensioneel beoordelen (i.c. alle opstellen moeten beoordeeld worden op telkens hetzelfde aspect, ongeacht de vraag of dat aspect relevant is voor een treffende karakterisering van een opstel), waartoe deze niet in staat is. Een analytisch beoordelaar blijft typologisch beoordelen, en dat maakt dat analytische beoordeling even betrouwbaar (i.c. stabiel) is als globale beoordeling.
Dat laatste kan als volgt verduidelijkt worden. Stel dat een beoordelaar 20 opstellen nakijkt met behulp van een analytisch schema dat onder meer de categorie 'oorspronkelijkheid' bevat. Als die beoordelaar bij zijn analytische beoordeling typologisch te werk gaat, zoals ik aanneem, dan zal hij een opstel uitsluitend hoog waarderen wanneer dat opstel er op die categorie 'uit-

springt', wanneer met andere woorden dat opstel in vergelijking met de andere opstellen extreem oorspronkelijk is. Alle opstellen die in de ogen van de beoordelaar niet of nauwelijks oorspronkelijk zijn, worden op één neutrale hoop geveegd. Zo'n beoordelaar maakt een heel scherp onderscheid tussen opstellen die extreem oorspronkelijk zijn, en opstellen die dat niet zijn — hij maakt nauwelijks een onderscheid tussen opstellen die weinig oorspronkelijk zijn. Voor de *betrouwbaarheid* van het analytisch oordeel heeft dit de volgende consequentie: de beoordelaar kent betrouwbare analytische oordelen toe, voor zover een opstel extreem oorspronkelijk is, hij kent onbetrouwbare oordelen toe voor zover opstellen weinig of helemaal niet oorspronkelijk zijn. Het analytische oordeel van een typologisch oordelende beoordelaar is derhalve opgebouwd uit én betrouwbare én onbetrouwbare componenten, en niet — zoals bij analytische schema's impliciet wordt aangenomen — uit uitsluitend betrouwbare componenten. Het gevolg hiervan is dat analytische beoordelingen niet betrouwbaarder (i.c. stabieler) zijn dan globale.
De hier voorgestelde theoretische verklaring voor het falen van analytische schema's is op twee manieren empirisch getoetst, enerzijds via een kwalitatieve analyse van protocollen van hardop-denkende beoordelaars tijdens het beoordelen van opstellen, anderzijds via een kwantitatief-statistische analyse van opstelbeoordelingen. De resultaten van dat empirisch onderzoek zijn grotendeels in overeenstemming met de uit de theoretische verklaring afgeleide empirische voorspellingen (Meuffels 1985).

Globaal of analytisch beoordelen?

Om hun schrijfvaardigheid te verbeteren dienen leerlingen vaak in de gelegenheid te worden gesteld de betreffende vaardigheid te oefenen en dienen ze specifieke, gedifferentieerde feedback over hun sterke en zwakke punten te ontvangen. Die feedback moet betrouwbaar zijn. Omdat de betrouwbaarheid van globale beoordeling vaak te wensen over laat, hebben velen hun toevlucht gezocht in een analytische beoordelingswijze, echter zonder veel resultaat. Een globale beoordelingswijze is even betrouwbaar als een analytische, als men de betrouwbaarheid van analytische beoordeling berekent voor analytische eindoordelen. Dit laatste is echter in methodisch opzicht een dubieuze procedure. Vergelijkt men derhalve de betrouwbaarheid van analytische beoordeling per afzonderlijke categorie met die van globale beoordeling, dan blijkt globale beoordeling door de bank genomen zelfs superieur!
Betekent dit nu dat analytische schema's maar afgeschaft moeten worden? Wat pleit er eigenlijk vóór zo'n schema, en wat er tegen?
Het in praktisch opzicht evidente nadeel van analytische schema's is dat deze van een beoordelaar wel erg veel vergen: opstellen moeten telkens afzonderlijk op één aspect beoordeeld worden, zodat niet alleen een groot beroep wordt gedaan op het concentratievermogen, maar tevens op de beschikbare tijd. Via een globale beoordelingsmethode corrigeert men al snel zo'n 20 tot 40 opstellen per uur (Finlayson 1951; Wolowitsj 1976), een onbereikbaar aantal wanneer men de opstellen analytisch zou nakijken. Dit bezwaar klemt des temeer omdat de extra tijdsinvestering bij een analytische beoordelingsmethode in het geheel niet gecompenseerd wordt door een forse betrouwbaarheidswinst.
Het grote voordeel van analytische schema's is de transparantie in de beoordeling. De leerling weet vooraf, op welke aspecten hij zal worden beoordeeld — en zo hoort het ook. Zonder een volledig en uitputtend overzicht van voor- en nadelen van analytische schema's te presenteren is toch zoveel al duidelijk dat wanneer bij een globale beoordelingswijze vooraf de criteria worden gespecificeerd waarop schrijfprodukten beoordeeld zullen worden, en wanneer gedifferentieerde feedback wordt geleverd op basis van in schrijfprodukten aanwezige saillante karakteristieken, er wel heel sterke argumenten moeten worden aangevoerd om een analytische beoordelingswijze te prefereren boven een globale beoordelingswijze. Zonder intensieve training zijn beoordelaars immers niet tot dimensioneel oordelen in staat. Bovendien zijn dimensionele oordelen in het gros van de gevallen voor leerlingen nauwelijks informatief.

Literatuur

Barendregt, J.T. 'Een knelpunt in de psychodiagnostiek' in: *Bulletin Persoonlijkheidsleer* 1978/6, p. 126-128

Barendregt, J.T. 'Een natuurlijke methodologie' in: G.J. Mellenbergh, R.F. van Naersen, H. Wesdorp (red.) *Rede als richtsnoer*; bijdragen over methoden van denken en werken in de gedragswetenschappen aangeboden aan prof. dr. A.D. de Groot bij zijn afscheid van de Universiteit van Amsterdam. 's Gravenhage, Mouton, 1979, p. 13-18

Bem, D.J. & A. Allen 'On predicting some of the people some of the time; the search for cross-situational consistencies in behavior' in: *Psychological Review* 1974, p. 506-520

Finlayson, D.S. 'The reliability of the marking of essays' in: *British Journal of Educational Psychology* 1951/21, p. 134-136

Lamb, H. 'The English essay in secundary selection examinations; a comparision of two methods of marking' in: *British Journal of Educational Psychology* 1953/23, p. 131-133

Meuffels, B. 'Hoe meer, hoe beter?' in; *Tijdschrift voor Taalbeheersing* 1983/5, p. 243-256

Meuffels, B. 'Het beoordelen van opstellen: dimensioneel of typologisch?' in: W.K.B. Koning (red.) *Taalbeheersing in theorie en praktijk* Dordrecht, Foris Publications, 1985, p. 386-392

Morrison, R.L. & P.E. Vernon 'A new method of marking English compositions' in: *British Journal of Educational Psychology* 1941/11, p. 109-119

Nisbet, J.D. 'English composition in secundary school selection' in: *British Journal of Educational Psychology* 1955/25, p. 51-55

Volovics-Schelvis, J.J. 'Het ontwikkelen en beoordelen van een analytisch beoordelingsschema voor schrijfprodukten van mavo-leerlingen' in: *Tijdschrift voor Taalbeheersing* 1979/1, p. 75-94

Wesdorp, H. *Evaluatietechnieken voor het moedertaalonderwijs*; een inventarisatie van beoordelingsmethoden voor de stelvaardigheid, het begrijpend lezen, de spreek-, luister- en discussievaardigheid. 's Gravenhage, SVO-reeks nr. 42, 1981

Wolowitsj, J.J. *Het ontwikkelen van een analytisch beoordelingsschema voor schrijfprodukten van mavo-leerlingen* Amsterdam, RITP, 1976

Zondervan, F. e.a. *Ontwikkeling en toetsing van een analytisch beoordelingsmodel. Een exploratief onderzoek* Int. publ. RU, 1979

Zijlmans, S. & H. Blok *De ontwikkeling van een beoordelingsinstrumentarium voor Gericht Schrijven mavo/lbo* Interimrapport I en II. Amsterdam, RITP, 1980