43

Het beoordelen van opstellen: dimensioneel of typologisch?

Bert Meuffels

1. Inleiding: de betrouwbaarheid van globale en analytische beoordeling

Opstellen kunnen op globale of analytische wijze beoordeeld worden. Een beoordelaar beoordeelt opstellen op *globale* wijze wanneer hij oordelen toekent op basis van zijn eigen intuïtieve, niet-geëxpliciteerde normen; hij beoordeelt de opstellen in dat geval op grond van een ongedifferentieerde totaalindruk, resulterend in één cijfer. Wanneer meerdere beoordelaars dezelfde serie opstellen op globale wijze nakijken, bestaat er een redelijke kans dat hun oordelen uiteenlopen als gevolg van verschillen in opvatting over de beoordelingstaak. Een globale opstelbeoordeling laat immers individuele 'stokpaardjes' de vrije teugel. De één is allergisch voor spelfouten en beoordeelt opstellen primair op het aantal en de aard van die fouten, de ander is van mening dat spelfouten irrelevant zijn voor de 'werkelijke' kwaliteit en acht stilistische fraaiheid van doorslaggevend belang, enzovoort. Resultaten van empirische onderzoek tonen dan ook aan dat de overeenstemming tussen globale beoordelaars door de bank genomen gering is.

Om die overeenstemming te verhogen zoeken veel onderzoekers hun toevlucht in een *analytisch* beoordelingsschema. Anders dan bij globale beoordeling, wordt bij analytische beoordeling de 'overall'-kwaliteit van een opstel opgesplitst in een aantal verschillende, min of meer onafhankelijke deelaspecten (bijvoorbeeld stijl, spelling, inhoudelijke structuur, enzovoort). Elk opstel moet op elk van de onderscheiden aspecten apart beoordeeld worden. Het eindoordeel over een opstel bestaat uit de som of het gemiddelde van de cijfers voor de in het analytisch schema onderscheiden aspecten.[1]

Het ligt voor de hand om te veronderstellen dat de beoordelingstaak door een analytisch schema verduidelijkt en verscherpt wordt, zodat de betrouwbaarheid van analytische beoordeling groter zal zijn dan die van globale beoordeling. De resultaten van empirisch onderzoek waarin de betrouwbaarheid van beide beoordelingsprocedures met elkaar vergeleken werd, spreken deze intuïtief plausibele veronderstelling echter tegen.[2] Analytische beoordeling leidt niet tot een grotere overeenstemming tussen beoordelaars (interbeoordelaarsovereenstemming) en evenmin tot een grotere consistentie in de beoordelingen van één individuele beoordelaar, wanneer deze dezelfde serie opstellen met een tussenperiode van enkele weken opnieuw beoordeelt (beoordelaarsstabiliteit).[3] In dit artikel probeer ik hiervoor een verklaring te geven, althans voorzover het de *beoordelaarsstabiliteit* betreft.

W.K.B. Koning (red.), *Taalbeheersing in theorie en praktijk*. Dordrecht 1985: Foris Publications,

2. Een mogelijke verklaring

Een voor de hand liggende verklaring voor het falen van analytische schema's 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 'inhoudelijke structuur', 'stijl', enzovoort, bieden de beoordelaar relatief weinig steun, zodat deze in zijn cijfergeving grotendeels op privé-normen en interpretaties is aangewezen. 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 in de beoordelingscategorieën (Zijlmans & Blok 1980; Volovics-Schelvis 1979; Zondervan 1979; Meuffels 1983). Resultaten van dit onderzoek, waarin de analytische categorieën verbaal zijn toegelicht of verduidelijkt met behulp van voorbeeldopstellen teneinde de interpretaties van die categorieën te uniformeren, tonen niettemin hetzelfde teleurstellende beeld. Conceptuele en/of terminologische vaagheid kan sommige onderzoeksresultaten misschien verklaren, maar zeker niet alle.

Kern van de hier voorgestelde verklaring voor het falen van analytische schema's is dat beoordelaars opstellen niet zozeer *dimensioneel* als wel *typologisch* beoordelen. Een analytisch schema dwingt de beoordelaar tot een vorm van dimensioneel beoordelen, waartoe deze – gezien de perceptueel-cognitieve beperkingen van het menselijk informatieverwerkend systeem – niet in staat is, althans niet zonder intensieve training.[4] De analytische beoordelaar beoordeelt opstellen typologisch (de 'natuurlijke' wijze van beoordelen), precies zoals de globale beoordelaar dat doet. Het gevolg hiervan is dat analytische beoordeling even betrouwbaar (stabiel) als globale beoordeling is.

3. Dimensioneel versus typologisch beoordelen

Een analytisch schema bestaat uit een aantal beoordelingscategorieën en daarmee corresponderende dimensies, in casu cijferschalen met schaalpunten van bijvoorbeeld 1 tot 10. Er is sprake van een dimensionele beoordeling wanneer de opstellen netjes op een rij op de schaal worden gezet, en wel naar de *mate* waarin ze de door de schaal bestreken eigenschap bezitten.

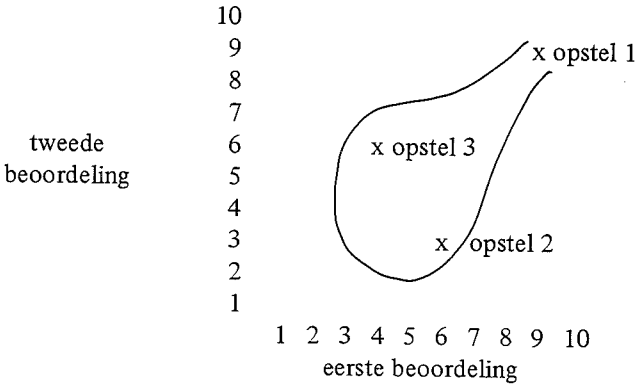
Stel dat iemand gevraagd wordt zo nauwkeurig en informatief mogelijk drie hem bekende personen (A, B en C) te beschrijven in termen van de persoonlijkheidsschaal (dimensie) 'vrolijk-somber'. A wordt als een vrolijke jongen gekarakteriseerd, B als een sombere en C als een jongen van ... ja van wat eigenlijk? Van 'gemiddelde' vrolijkheid? Maar wat moet je je daar eigenlijk 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 zijn 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 hem met andere begrippen te beschrijven (vgl. Bem 1974). Mensen beoordelen casu quo beschrijven elkaar niet in termen van gradaties op telkens dezelfde dimensies, maar karakteriseren verschillende personen in termen van verschillende soorten eigenschappen.

De bovenstaande gedachtengang, ontleend aan de zogenaamde 'personology', kan gegeneraliseerd worden naar opstelbeoordeling. Verschillende opstellen nodigen uit tot een beschrijving in termen van verschillende eigenschappen casu quo kwaliteiten. Het ene opstel treft door een fraaie stijl, het ander door een zwakke spelling en interpunctie, enzovoort. Het beoordelen van opstellen, of dat nu op globale of analytische wijze geschiedt, is niet zozeer moeilijk omdat binnen elke beschrijvingswijze zich tussen opstellen grote verschillen voordoen, maar omdat verschillende opstellen uitnodigen tot verschillende beschrijvingswijzen.

4. Stabiliteit

Als beoordelaars inderdaad opstellen typologisch beoordelen, dan zijn analytische beoordelingen niet stabieler dan globale. Dit kan toegelicht worden met behulp van het volgende voorbeeld. Stel dat één beoordelaar na twee maanden dezelfde serie opstellen met behulp van hetzelfde analytische schema beoordeelt. Stel verder dat het schema de categorie 'oorspronkelijkheid' bevat. Voor die categorie zou het volgende resultaat (grafisch weergegeven) uit de bus kunnen komen:



De beoordelaar geeft opstel 1 bij de eerste beoordeling een 9 voor 'oorspronkelijkheid' en bij de tweede beoordeling eveneens. Opstel 2 wordt door hem bij de eerste keer met een 6 gewaardeerd, bij de tweede keer met een 3. Opstel 3 krijgt bij de eerste beoordeling een 4, bij de tweede beoordeling een 6. De in de figuur weergegeven grafiek vormt in feite een puntenwolk waarbij elk punt in de grafiek één opstel representeert (slechts drie opstellen zijn in de figuur gemarkeerd; in feite is de figuur opgebouwd op basis van de beoordelingen van vele, vele opstellen). De horizontale positie van een opstel in de grafiek wordt bepaald door het cijfer voor dat opstel bij de eerste beoordeling, de verticale positie door het cijfer bij de tweede beoordeling.

De beoordelaar van opstellen volgens bovenstaande figuur meet veel nauwkeuriger in het hoge gebied van de analytische categorie 'oorspronkelijkheid' dan in het lage gebied. (De discrepantie tussen de cijfers bij de eerste en de tweede beoordeling is immers veel kleiner wanneer een opstel als erg oorspronkelijk wordt gekwalificeerd dan wanneer het als matig of nauwelijks oorspronkelijk wordt beoordeeld.) Op grond van de door deze beoordelaar toegekende cijfers heeft het wel zin om een opstel als 'oorspronkelijk', maar weinig zin om een opstel als 'niet' of 'weinig oorspronkelijk' te bestempelen. In psychologisch opzicht vormt oorspronkelijkheid voor deze beoordelaar geen dimensie, maar een typologie.

De beoordelaar van opstellen volgens bovenstaande figuur is een relatief onbetrouwbare beoordelaar, voorzover het het lage gebied van 'oorspronkelijkheid' betreft, een relatief betrouwbare beoordelaar voorzover het het hoge gebied betreft. Gaan we er van uit dat verschillende opstellen uitnodigen tot verschillende beschrijvingswijzen, dan is hiermee een verklaring gegeven voor het feit dat – althans voorzover het de *stabiliteit* van één beoordelaar betreft – analytische beoordelingen niet betrouwbaarder zijn dan globale. De score per analytische categorie en de analytische totaalscore is immers opgebouwd uit èn onbetrouwbare èn betrouwbare componenten en niet – zoals bij analytische schema's impliciet wordt aangenomen – uit uitsluitend betrouwbare componenten.

5. Toetsing

Bovengenoemde verklaring is getoetst via een analyse van protocollen van vier hardop-beoordelende personen. Het ontbreekt hier aan ruimte die analyse – die de verwachtingen overigens grotendeels confirmeerde – te rapporteren en te bespreken.[5] We voegen er daarom – ook al met het oog op de enigszins dubieuze status van protocolanalyse voor theorietoetsing – een statistisch argument aan toe. Als proefpersonen typologisch beoordelen, dan ontstaan (bij herbeoordeling) puntenwolken zoals in de bovenstaande figuur (dat is overigens niet de enige configuratie die indicatief is voor een typologische beoordeling). Als proefpersonen dimensioneel beoordelen, dan ontstaan puntenwolken waarvoor geldt dat de discrepanties tussen cijfers bij de

eerste en tweede beoordeling over de hele cijferschaal min of meer even groot zijn. In de figuur is de puntenwolk gestileerd; empirische gegevens zijn in de regel niet zo gemakkelijk met het 'blote oog' te evalueren. Om uit te maken of beoordelaars typologisch dan wel dimensioneel te werk gaan, gebruiken we een door Van Houwelingen ontwikkelde statistische procedure.[6]

Acht proefpersonen beoordeelden onafhankelijk van elkaar 19 VWO-eindexamenopstellen op analytische en op globale wijze. Twee maanden later vond een herbeoordeling plaats. De resultaten van dit onderzoek staan in onderstaande tabel, waarin per beoordelaar en per analytische categorie (10 in totaal) stabiliteitscoëfficiënten (pmc) staan vermeld. Een beoordelaar beoordeelt stabiel indien zijn stabiliteit groter is dan of gelijk aan .45.

Analytische categorie	Beoordelaar							
	1	2	3	4	5	6	7	8
Inhoud gevraagde?	-.01	.36	.00	.65*	.80*	.78	.00	.62
Structuur	.34	.61	.34	.59	.80	.71*	.35	-.01
Woordgebruik	.67*	.38	.39	.83*	.72*	.36	.14	.70
Aantrekkelijkheid zinnen	.36	.49*	.42	.78*	.67*	.40	.10	.76
Zinsbouw	.38	.63*	.49*	.72*	.50	.42.	.02	.46*
Spelling	.19	.89*	.35	.71*	.49*	.33	.53*	.58*
Leestekens	.27	.09	.10	.47	.45*	.16	-.23	.32
Oorspronkelijkheid	.75	.66*	.45*	.78	.84*	.41	.00	.68*
Argumentatie	.51*	.55*	.28	.73*	.66*	.53	-.11	.73*
Stijl	.60*	.47*	.33	.91	.72*	.31	.26	.80*
anal.totaal	.72	.55	.47	.88	.83	.63	.14	.89
globaal	.49	.68	.50	.68	.89	.61	.45	.89

We concluderen het volgende:

1. Geen enkel analytisch kenmerk wordt door alle beoordelaars stabiel beoordeeld. Dat zelfs relatief eenduidige kenmerken als 'spelling' en 'leestekens' niet stabiel worden beoordeeld, maakt een verklaring van het falen van analytische schema's in termen van conceptuele of terminologische vaagheid van de analytische beoordelingscategorieën implausibel.
2. Slechts twee van de acht beoordelaars (beoordelaars 4 en 5) beoordelen alle analytische kenmerken op stabiele wijze.
3. Anders dan bij de analytische beoordeling blijken alle beoordelaars bij de globale beoordeling stabiel te beoordelen.

4. Het heeft geen zin de interbeoordelaarsovereenstemming bij analytische en globale beoordeling met elkaar te vergelijken, evenmin als het zin heeft de stabiliteit van de totale analytische beoordeling met die van de globale beoordeling te vergelijken. De analytische totalen immers bevatten ruis. Dat analytische beoordelaars het onderling niet meer met elkaar eens zijn dan globale, kan verklaard worden uit de non-stabiliteit in de beoordeling van de analytische kenmerken (per beoordelaar telkens variërend per kenmerk). Men zou hier uit af kunnen leiden dat verschillende opstellen *verschillende* beoordelaars uitnodigen tot verschillende typeringen.
5. Als we voor de 44 stabiel beoordeelde kenmerken statistisch nagaan of deze typologisch dan wel dimensioneel worden beoordeeld, dan blijken er 32 typologisch beoordeeld te worden (in de tabel zijn deze 'hits' met een asterisk gemarkeerd). Het aantal 'misses' zou misschien verklaard kunnen worden uit het geringe onderscheidingsvermogen van de gehanteerde statistische toets (bij n=19). Dit onderzoek zal dan ook gerepliceerd moeten worden, waarbij de beoordelaars meer opstellen zullen moeten nakijken dan in dit onderzoek het geval was.

Op grond van de resultaten kan men voorzichtig concluderen dat verschillende opstellen een beoordelaar uitnodigen tot verschillende typeringen, en – generaliserend – dat verschillende opstellen verschillende beoordelaars uitnodigen tot verschillende typeringen.

Noten

1. De gebruikelijke procedure om de overall-kwaliteit van een opstel op te splitsen in een aantal min of meer onafhankelijke deelaspecten en vervolgens de cijfers voor die aspecten op te tellen tot een totaalscore, is in methodologisch opzicht dubieus. Zijn de onderscheiden aspecten in conceptueel en empirisch opzicht inderdaad onafhankelijk, dan leidt optellen tot een oninterpreteerbare totaalscore.
2. Wesdorp concludeert in zijn overzicht van empirisch onderzoek naar de betrouwbaarheid van analytische en globale beoordeling dat "de samengevatte onderzoeksresultaten teleurstellend" zijn. De stabiliteit is "voor analytische beoordeling iets beter, echter weinig opvallend beter (...). De interbeoordelaars-betrouwbaarheid is ook meestal iets hoger voor analytici" (Wesdorp 1981: 57–58). Gaat men echter (met behulp van de uit de gerapporteerde betrouwbaarheden afgeleide gemiddelde intercorrelatie tussen beoordelaars) na of analytische beoordeling een substantieel, significant hogere betrouwbaarheid heeft dan globale, dan blijkt dat niet het geval te zijn.
3. Zie voor een niet-technische uitleg van (verschillende aspecten van het formele begrip) betrouwbaarheid, Meuffels (1983).
4. Het ontbreekt hier aan ruimte om het cognitieve karakter van de voorgestelde verklaring uiteen te zetten.
5. De beoordelaars lijken opstellen te beoordelen via een proces dat beschreven kan worden in termen van de empirische cyclus van De Groot. Al in een heel vroeg stadium worden relatief specifieke hypothesen geformuleerd, die vervolgens 'getoetst' worden. Bij dat 'toetsen' gaan de beoordelaars strikt confirmistisch te werk.
6. De door Van Houwelingen ontwikkelde procedure is een toets op homoscedasticiteit.