Het beoordelen van de kwaliteit van rubrics

■ **Johan van Strien**
Johan van Strien is universitair docent bij de Open universiteit. Email: johan.vanstrien@ou.nl

■ **Desirée Joosten-ten Brinke**
Desirée Joosten-ten Brinke is universiteit hoofddocent bij de Open universiteit, lector bij Fontys lerarenopleiding Tilburg.

Van leerlingen en studenten wordt onder meer verwacht dat zij betogen en essays kunnen schrijven, spreekbeurten kunnen geven, onderzoeken kunnen opzetten en rapporteren, kunnen samenwerken en kritisch denken. Het beoordelen van deze complexe vaardigheden is niet eenvoudig. Om deze vaardigheden objectief en betrouwbaar te beoordelen is het goed hanteren van criteria essentieel. Steeds vaker worden voor het verhogen van de kwaliteit van deze beoordelingen rubrics gebruikt. In dit artikel presenteren we de resultaten van een onderzoek naar de kwaliteit van rubrics aan de hand van een 'rubric voor rubrics'.

Rubrics

Een rubric is een beoordelingsinstrument dat inzicht verschaft in de evaluatiecriteria van een vaardigheid en per criterium de niveaus van presteren. De evaluatiecriteria zijn de aspecten waarop wordt beoordeeld en de prestatieniveaus geven de mate weer waarin tegemoet wordt gekomen aan het criterium, bijvoorbeeld van onvoldoende tot zeer goed, of van beginner tot gevorderde. Per criterium en prestatieniveau wordt in een rubric een beschrijving gegeven van waaraan de geleverde prestatie moet voldoen om tot een bepaalde beoordeling te komen, de zogenoemde indicatoren (Reddy & Andrade, 2010; Van den Bos, Burghout, & Joosten-ten Brinke, 2014). Met een rubric kan de docent inzichtelijk maken waarom een score op een bepaald aspect niet 'zeer goed' maar 'goed' is. Indien er alleen met een criteriumlijst gewerkt wordt, geven studenten aan dat zij het moeilijk vinden om de criteria die docenten stellen te interpreteren (Jonsson, 2014). Met een rubric weten studenten beter wat er van hen verwacht wordt (Panadero & Jonsson, 2013). Rubrics kunnen dan ook voor formatieve doeleinden gebruikt worden. Ze zijn bij uitstek geschikt om te laten zien op welke niveau studenten nu zitten en wat zij moeten laten zien om verder te komen. Echter, het werken met rubrics leidt niet automatisch tot betere kwaliteit van beoordelingen (Straetmans, 2015). Door beoordelaars te trainen in het gebruik van rubrics is het mogelijk om met het gebruik van rubrics te komen tot meer betrouwbare en valide beoordelingen (Reddy & Andrade, 2010).

Eisen aan rubrics

Het opstellen van een goede rubric is echter niet eenvoudig. Een rubric moet valide zijn. Dat houdt in dat deze goed moet aansluiten op de doelgroep en de leerdoelen, en inzichtelijk moet maken welke criteria belangrijk en relevant zijn om een vaardigheid aan te tonen (Reddy & Andrade, 2010). Het aantal beoordelingsniveaus moet bovendien, bij formatief

gebruik, betekenisvol zijn en onderscheid kunnen maken in de fase van ontwikkeling van een lerende. Daarnaast moeten indicatoren zo concreet mogelijk zijn en zo min mogelijk op puur kwantitatieve criteria berusten. Ook moeten zij per criterium parallel geformuleerd zijn, dat wil zeggen dat er een logische opbouw is over de prestatieniveaus.
Om de kwaliteit van rubrics te analyseren hebben Arter en Chappuis (2006) een 'rubric for rubrics' ontwikkeld. Deze hebben we vertaald naar het Nederlands en op een aantal punten aangescherpt (zie tabel 1). De rubric omvat drie hoofdcriteria:

- *Validiteit*. Dit betreft de mate waarin de inhoud een goede afspiegeling geeft van wat er nodig is om een goede prestatie te leveren en van wat er van studenten verwacht mag worden gezien hun niveau.
- *Kwaliteit van de criteria.* Dit betreft het aantal, de relevantie van en de samenhang tussen de criteria.
- *Kwaliteit van de indicatoren*. Dit betreft de invulling van de beoordelingsniveaus van de criteria. Zijn deze concreet en parallel geformuleerd?

Case study

Docenten zijn in de afgelopen jaren steeds vaker aan de slag gegaan met het ontwerpen van rubrics, maar geven vaak aan dat het lastig is om te komen tot een rubric waar zij en hun studenten helemaal tevreden over zijn. Via een case study hebben we onderzocht in hoeverre de ontwikkelde rubrics voldeden aan de kwaliteitseisen voor een goede rubric. Op basis van de resultaten is een set van meest voorkomende fouten en aanbevelingen opgesteld.

In deze studie hebben we de rubrics van 20 docenten van de Open Universiteit als uitgangspunt genomen. De docenten hebben in het kader van hun docentprofessionalisering een workshop 'Werken met rubrics' gevolgd.

De workshop duurde drie uur. In het eerste uur werd aandacht besteed aan de theoretische achtergrond: wat zijn rubrics, waarvoor worden ze gebruikt en aan welke eisen moeten ze voldoen? Bij de constructie van rubrics werd nadrukkelijk aandacht besteed aan de criteria van goede rubrics. In de overige twee uur gingen deelnemers aan de slag met het construeren van een rubric voor een eigen cursus. Zij moesten voor minimaal vier criteria indicatoren opstellen. Na de workshop werkten de docenten de rubric verder uit en dienden deze in voor feedback.

Analyse

De rubrics werden beoordeeld door de twee auteurs, waarbij voor elk criterium maximaal 2 punten gegeven konden worden. Om de criteria en indicatoren op dezelfde manier te interpreteren werden eerst negen rubrics gescoord en werden de analyses met elkaar vergeleken. Er zijn meerdere rondes geweest om de indicatoren in de rubric voor rubrics eenduidig te interpreteren. Uiteindelijk werd een overeenstemming van 0,91 gerealiseerd. Op basis van de acht criteria kon de kwaliteit van een rubric met maximaal 16 punten worden beloond.

Resultaten

In tabel 2 staat een overzicht van de gemiddelde scores per criterium. De gemiddelde totaalscore per beoordeelde rubric was 11,6 (SD = 2,66), met een bereik van 5 tot 16 punten. Over het algemeen werd vrij hoog gescoord op de verschillende criteria, met name 'aanwezigheid', 'dekkingsgraad' en 'onderscheidend vermogen' werden hoog beoordeeld. Op het criterium 'weging' werd laag gescoord. Uit de rubrics kon niet worden afgeleid of alle criteria even zwaar telden.

Op het criterium 'transparantie en parallelliteit' werd wisselend gescoord met een vrij grote spreiding.

Tabel 1. De rubric voor het beoordelen van rubrics (bijgestelde versie van Arter & Chappuis, 2006)

	Goed (2 pnt)	Voldoende (1 pnt)	Onvoldoende (0 pnt)
Validiteit			
Dekkingsgraad	De inhoud weerspiegelt correct wat er van studenten realistisch gezien mag worden verwacht gezien hun niveau en de leerdoelen. De inhoud van de rubric geeft de best mogelijke weergave van wat er nodig is om een goede prestatie te leveren op de vaardigheid of het product dat wordt beoordeeld.	De inhoud weerspiegelt wat er van studenten realistisch gezien mag worden verwacht gezien hun niveau en de leerdoelen, al geldt dit niet voor alle criteria. De inhoud van de rubric geeft grotendeels een goede weergave van wat er nodig is om een goede prestatie te leveren, maar er zijn punten waarop verbetering nodig is.	De inhoud weerspiegelt niet wat er van studenten mag worden verwacht gezien hun niveau en de leerdoelen. De inhoud van de rubric geeft geen goede weergave van wat er nodig is om een goede prestatie te leveren, waardoor niet duidelijk is wat de rubric beoogt te meten.
Prestatieniveau	Studenten mogen verondersteld worden in staat te zijn het hoogste niveau te bereiken **op alle criteria**. Om elk criterium op 'voldoende' niveau te beheersen moeten studenten een acceptabele prestatie laten zien.	Studenten mogen verondersteld worden in staat te zijn het hoogste niveau te bereiken op alle criteria. Op **enkele criteria** kan met een zwakke prestatie reeds een voldoende of hoger worden behaald.	De eisen die worden gesteld om het hoogste niveau te bereiken zijn onrealistisch en/of studenten kunnen door een zwakke prestatie **op meerdere criteria** reeds een voldoende of hoger halen.
Kwaliteit van de criteria			
Aantal, samenhang, opbouw	De rubric omvat precies het juiste aantal criteria, zodat de complexiteit van de te leveren prestatie wordt weerspiegeld. De criteria kennen een goede samenhang en logische opbouw.	Het aantal criteria vereist enige aanpassing, zoals het opsplitsen van een afzonderlijk criterium in twee aparte criteria, of het samenvoegen van twee criteria. De criteria kennen verder een goede samenhang, maar de opbouw is niet altijd logisch.	De rubric bestaat uit een lange lijst criteria die niet of nauwelijks met elkaar lijken samen te hangen en waarin geen keuzes lijken te zijn gemaakt. De opbouw is niet logisch.
Weging	Het is duidelijk of de rubric bedoeld is voor formatieve en/of summatieve doeleinden. Indien summatieve doeleinden, dan is in de rubric expliciet aangegeven hoe zwaar elk criterium meetelt, waardoor is af te leiden welke criteria belangrijker zijn dan andere (eventueel in begeleidende tekst). Alle criteria krijgen de nadruk die zij verdienen.	Het is duidelijk of de rubric bedoeld is voor formatieve en/of summatieve doeleinden. Indien summatieve doeleinden, dan is in de rubric weliswaar expliciet (eventueel in begeleidende tekst) aangegeven hoe zwaar elk criterium meetelt en welke criteria belangrijker zijn dan andere, maar sommige criteria krijgen te veel of te weinig nadruk.	Het is niet duidelijk of de rubric bedoeld is voor formatieve en/of summatieve doeleinden. In de rubric wordt niet expliciet (ook niet in begeleidende tekst) aangegeven hoe zwaar elk criterium meetelt en welke criteria belangrijker zijn dan andere. De criteria krijgen hierdoor niet de nadruk die zij verdienen.
Onafhankelijkheid	De criteria zijn onafhankelijk van elkaar, waardoor zij verschillende aspecten meten. Er is geen overlap tussen verschillende indicatoren. Wat hoort bij de ene indicator komt nergens anders terug.	De criteria zijn grotendeels onafhankelijk. Er is enige overlap tussen verschillende indicatoren, maar de rubric is desondanks in staat om verschillende aspecten te meten.	De criteria zijn niet onafhankelijk. Er is veel overlap tussen verschillende indicatoren. Verschillende indicatoren omvatten hetzelfde, waardoor de criteria niet verschillende aspecten meten.
Onderscheidend vermogen	Het aantal beoordelingsniveaus per criterium is logisch en toereikend. Er zijn voldoende niveaus om adequaat onderscheid te kunnen maken tussen goede en minder goede studenten en om voortgang te kunnen meten.	Het aantal beoordelingsniveaus per criterium is niet overal toereikend om adequaat onderscheid te kunnen maken tussen goede en minder goede studenten en om voortgang te kunnen meten, maar het aantal is redelijk eenvoudig aan te passen door één niveau toe te voegen dan wel twee niveaus samen te voegen.	Het aantal beoordelingsniveaus is veel te groot of juist te klein om zinvol en betrouwbaar onderscheid te kunnen maken tussen studenten. Grote aanpassingen zijn nodig.
Kwaliteit van de indicatoren			
Aanwezig	Elk beoordelingsniveau is voorzien van indicatoren.	Niet alle beoordelingsniveaus zijn voorzien van indicatoren.	Geen van de beoordelingsniveaus is voorzien van indicatoren.
Transparantie en parallelliteit	De indicatoren zijn gedetailleerd en concreet genoeg zodat duidelijk is wat de student moet laten zien om een goede prestatie te leveren (ze zijn kwalitatief ipv kwantitatief) en de beoordelaar een juiste beoordeling kan geven. Er is een logische opbouw van indicatoren over de niveaus (parallelliteit)	Er is een poging ondernomen om indicatoren gedetailleerd en concreet te formuleren, maar sommige bevatten nog enkele vage beschrijvingen. Ze zijn wel grotendeels kwalitatief (en niet kwantitatief) Er is veelal sprake van parallelliteit, maar bij een aantal criteria is de opbouw niet logisch.	De indicatoren zijn vaag en weinig concreet beschreven, waardoor niet duidelijk wordt wat er van de student wordt verwacht. Of er wordt alleen onderscheid gemaakt tussen beoordelingsniveaus met woorden als 'zeer', 'erg' en 'enige', of ze zijn volledig kwantitatief. De parallelliteit ontbreekt meestal.

Elk criterium is onderverdeeld in subcriteria en alle criteria tellen even zwaar mee.

Tabel 2. Gemiddelde scores en standaarddeviatie per criterium en voor de rubric in totaal (maximum score per criterium is 2 punten)

Criterium	Gemiddelde score	Standaarddeviatie
Dekkingsgraad	1,8	.41
Prestatieniveau	1,5	.61
Aantal, samenhang en opbouw criteria	1,6	.50
Weging criteria	0,4	.68
Onafhankelijkheid criteria	1,5	.61
Onderscheidend vermogen	1,75	.44
Aanwezigheid indicatoren	1,85	.37
Transparantie en parallelliteit	1,2	.70
Totaal	11,6	2.67

Een rubric is een beoordelingsinstrument dat inzicht verschaft in de evaluatiecriteria van een vaardigheid en per criterium de niveaus van presteren

Conclusie

De meeste rubrics waren van goede kwaliteit en op de meeste criteria werd gemiddeld hoog gescoord. Uitzondering was het criterium 'weging'. Slechts enkele docenten gaven expliciet aan hoe de criteria zich tot elkaar verhielden qua weging. Een mogelijke verklaring is dat de overige docenten impliciet veronderstelden dat alle criteria even zwaar zouden moeten wegen, maar dat was niet af te leiden uit de rubrics. Waar het formuleren van transparante en parallelle indicatoren bekend staat als een uitdaging, ging het docenten over het algemeen aardig af.

Er zat echter wel veel spreiding in de kwaliteit van de indicatoren. Het was (samen met 'weging') het enige criterium waarop de score 1 vaker werd behaald dan de volledige score 2. Het formuleren van goede indicatoren verdient dan ook aandacht.

Uit reacties op de workshops bleek dat veruit de meeste deelnemers de meerwaarde van rubrics inzagen, al erkenden zij ook de complexiteit ervan. Vooral het met collega's discussiëren over wat een goede prestatie is wordt als zeer leerzaam ervaren.

Bij het scoren van de rubrics bleek het inschatten van de dekkingsgraad lastig voor

buitenstaanders, aangezien zij niet per se zicht hebben op het kennisdomein waarop de rubric betrekking heeft en daarom wellicht minder kritisch kunnen kijken naar de accuraatheid van de dekkingsgraad.

Het construeren van rubrics is een doorlopend proces, en training in het gebruik van rubrics door studenten en docenten is noodzakelijk (Brookhart & Chen, 2014; Reddy & Andrade, 2010). ■

Referenties

- Arter, J. A., & Chappuis, J. (2006). Creating & recognizing Quality Rubrics. Boston: Pearson.
- Brookhart, S. M., & Chen, F. (2014). The quality and effectiveness of descriptive rubrics. *Educational Review*, DOI: 10.1080/00131911.2014.929565
- Jonsson, A. (2014). Rubrics as a way of providing transparency in assessment. *Assessment & Evaluation in Higher Education, 39*, 840-852.
- Panadero, E. & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational research review, 9*, 129-144. doi: http://dx.doi.org/10.1016/j.edurev.2013.01.002
- Reddy, Y. M. & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education, 35*, 435-448.
- Straetmans, G. J.J. M. (2015). Gaan rubrics ons helpen om beter te beoordelen? *Examens, tijdschrift voor de toetspraktijk,* 4, 20-25.
- Van den Bos, P., Burghout, C., & Joosten-ten Brinke, D.(2014). Toetsen met rubrics. In H. Van Berkel, A. Bax, & D. Joosten-ten Brinke, *Toetsen in het hoger onderwijs.* (pp. 134-142). Houten: Bohn, Stafleu Van Lochem.

Gesignaleerd

Tien met een griffel voor de school

Een aantal scholen heeft het predikaat ‘excellente school’ gekregen.
“De waardering levert scholen geen extra budget op, zegt Dekker. Maar op termijn wel extra vertrouwen van het ministerie en de onderwijsinspectie. Scholen die langere tijd uitblinken, krijgen de vrijheid om af te wijken van een aantal regels en wetten. Dan mag het bijvoorbeeld wel wat minder met de bureaucratische rompslomp. Of ze geven de taalexamens een andere vorm. Ze vervangen bijvoorbeeld het standaard eindexamen Engels door het internationaal hoog aangeschreven Cambridge-certificaat”.
De Volkskrant 19 januari 2016