Gaan **rubrics** ons helpen om **beter te beoordelen?**

Gerard J.J.M. Straetmans

Sinds de nadruk in veel opleidingen is verschoven van de verwerving van kennis naar die van vaardigheden en competenties, zien we in het gebruik van toetsvormen een daarmee synchroon lopende verschuiving optreden van theorie- naar praktijktoetsen. Voor docenten die als assessor zijn betrokken bij deze praktijktoetsen betekent dit dat er een groot beroep zal worden gedaan op hun observerend vermogen en professionele oordeelsvorming. Rubrics kunnen de assessor daarbij helpen. Althans, dat lijkt een aannemelijke verklaring voor de enorme populariteit van dit hulpmiddel. Maar is die aanname wel juist?

Inleiding

In de afgelopen decennia is in alle sectoren van het onderwijs de nadruk binnen de opleidingsdoelen verschoven van kennisreproductie naar toepassing van kennis. Een logische consequentie daarvan is dat docenten meer belangstelling hebben gekregen voor zogeheten praktijktoetsen. 'Praktijktoets' is een overkoepelende term voor methodes waarmee de bekwaamheid van personen om authentieke taken uit te voeren wordt vastgesteld in een (gesimuleerde) werksituatie (Straetmans, 2014). Praktijktoetsen leveren twee soorten responsen op: de taakuitvoering (het proces) en het, al dan niet blijvende, resultaat van die taakuitvoering (het product).

Uitzonderingen daargelaten is bij praktijktoetsen de inzet van een assessor vereist om de kwaliteit van processen of producten te waarderen. In de onderzoeksliteratuur bestaat consensus dat dit een gevoelig nadeel is. Assessoren zijn vaak behept met eigenschappen die de betrouwbaarheid van hun beoordelingen aantasten. Lang geleden identificeerde en benoemde De Groot (1961) een aantal zogenoemde beoordelingseffecten om dit probleem in het onderwijs onder de aandacht te brengen.

In de afgelopen decennia zijn diverse maatregelen gepropageerd om deze beoordelingseffecten te bestrijden, zoals professionalisering en training van assessoren. Maar de belangrijkste maatregel is de toepassing van beoordelingsinstrumenten die het oordeel kunnen objectiveren. Deze beoordelingsinstrumenten, zoals check lists, rating scales en rubrics, bieden een alternatief voor de in het onderwijs frequent toegepaste glance-and-grade methode. Een lastig te vertalen term die misschien nog het beste benaderd wordt met uitdrukkingen als 'klinische blik' of 'timmermansoog'. De methode houdt in dat een assessor na observatie van het te beoordelen proces/product op een zelf bepaalde, niet nader te verantwoorden wijze tot een oordeel komt over de kwaliteit. Het gebrek aan standaardisatie dat deze methode kenmerkt, leidt gemakkelijk tot verschillen tussen beoordelaars bij de waardering van een bepaalde prestatie, maar ook tussen herhaalde oordelen van dezelfde beoordelaar. Op basis van dergelijke onbetrouwbare beoordelingen kunnen geen deugdelijke conclusies worden getrokken over de beheersing van leerdoelen. Reden waarom assessmentexperts voor de beoordeling van vaardigheden en competenties het gebruik van beoordelingsinstrumenten propageren. Waar aanvankelijk vooral check lists en rating scales de discussie over de inzet van beoordelingsintrumenten domineerden, staan de laatste tijd vooral rubrics in de belangstelling. Dit artikel beoogt de lezer te behoeden voor een blindelings vertrouwen in deze nieuwe beoordelingsmethodiek.

Een rubric is een beoordelingsschema om de prestaties van kandidaten te beoordelen

Rubric: kenmerken, voor- en nadelen

Een rubric is een beoordelingsschema om de prestaties van kandidaten te beoordelen die zijn uitgelokt door een assessmenttaak of open antwoordvraag. Het beoordelingsschema bestaat ten minste uit criteria en ondersteunende hulpmiddelen (AERA, 1999). Met 'criteria' worden de beoordelingsaspecten of -dimensies bedoeld waarop de kwaliteit van proces en/of product beoordeeld wordt (zie de cellen in de eerste kolom van Figuur 1). 'Hulpmiddelen' verwijst naar de aanwezigheid van beoordelingsschalen (zie de cellen in de eerste rij van Figuur 1) en de beschrijving van de diverse schaalpunten daarop aan de hand van zogeheten indicatoren (zie de teksten in alle overige cellen van Figuur 1). Een indicator is een operationalisatie van de vereiste kwaliteit zoals gespecificeerd door het

	Exemplary 4	Accomplished 3	Developing 2	Beginning 1
Comprehensibility	Listeners accustomed to the speech of learnes are able to understand all of the presentation.	Listeners accustomed to the speech of learners are able to understand most of the presentation.	Listeners accustomed to the speech of learners are able to understand the main ideas and some details of the presentation.	Listeners accustomed to the speech of learners are able to understand isolated bits of the presentation.
Text Type	Describes, narrates , and/or expresses own thoughts in paragraph-level discourse.	Describes, narrates, and/or expresses own thoughts in connected strings of sentences.	Speaks in loosely connected sentences.	Speaks in unconnected sentences and phrases.
Language Control	High degree of accuracy in grammar and word choice in connected, rehearsed and occasionally complex discourse. Little or no interference from first language.	Usually accurate grammar and word choice in connected, rehearsed discourse. Occasional interference from first language.	Frequent, but usually minor grammar and word choice errors in rehearsed, sentence-level discourse. Significant interference from first language.	Comprehension is impeded by frequent grammar and word choice errors in rehearsed discourse. High degree of interference from first language.
Vocabulary Use	Uses an broad range of familiar and new words, phrases and idioms so that expression is highly varied and non-repetitive.	Uses an adequate range of familiar and new words, phrases and idioms so that expression is varied and only occasionally repetitive.	Uses familiar words, phrases and idioms and rarely attempts to go beyond basic vocabulary. Speech is repetitive and lacks variety.	Uses very basic vocabulary and memorized phrases. Speech is limited and highly repetitive.
Communication Strategies	Always maintains communication. Able to circumlocute and self-correct when needed. Use of memory aids enhances presentation.	Very few breaks in communication. Sometimes able to circumlocute and self-correct. Effective use of memory aids.	Frequent breaks in communication. Rarely able to circumlocute or self-correct. Use of memory aids sometimes detracts from presentation.	Generally unable to maintain communication. Overreliance on memory aids detracts from presentation.

Figuur 1. Analytische, generieke rubric voor het beoordelen van een mondelinge presentatie in een vreemde taal (ACTFL, 1999).

Generieke rubrics zijn van toepassing op een hele klasse van assessmenttaken

betreffende beoordelingsaspect en het betreffende prestatieniveau.

Er bestaan verschillende soorten rubrics. In globale zin zijn ze langs twee dimensies te categoriseren. De ene dimensie betreft de detaillering van het te geven oordeel: holistisch versus analytisch. De andere gaat over de vraag waarop de beoordeling gericht moet zijn: op het te meten construct of op de specifieke assessmentopdracht (generiek versus taakspecifiek).

Een *holistische rubric* is erop gericht om de prestatie van een kandidaat op een assessmenttaak uit te drukken in één score, door alle beoordelingsdimensies tegelijkertijd te evalueren. Het voordeel hiervan is uiteraard dat het snel gaat, maar de keerzijde hiervan is dat één totaalscore de kandidaat weinig informatie geeft over wat er verbeterd moet worden. Voor formatieve doelen is dit type rubric niet geschikt.

Een *analytische rubric* beoordeelt prestaties van kandidaten afzonderlijk aan de hand van meerdere beoordelingsdimensies. De voordelen hiervan zijn dat de kandidaat gerichte feedback krijgt voor verbetering van zijn prestatie. Dit type rubric is zowel voor formatieve als summatieve assessmentdoelen geschikt. Nadelig is dat het beoordelen meer tijd vergt.

Een *generieke rubric* bevat criteria die niet exclusief geformuleerd zijn bij een bepaalde assessmenttaak maar bij het onderliggende construct waarvan de beheersing moet worden beoordeeld. Het voordeel van generieke criteria is dat ze niet slechts toepasbaar zijn op één specifieke assessmenttaak, maar op een hele klasse van assessmenttaken. Een ander voordeel is dat generieke rubrics gedeeld kunnen worden met studenten om bijvoorbeeld de doelen van het onderwijs te verhelderen of voor self-assessment doeleinden. Het nadeel van generieke criteria is dat ze meer training vergen van assessoren om voldoende betrouwbare scores te produceren.

Taakspecifieke rubrics bevatten criteria die speciaal geformuleerd zijn voor de beoordeling van een bepaalde assessmenttaak. Deze criteria kunnen in meer concrete termen geformuleerd worden en leveren daardoor betrouwbaardere scores op. De kwaliteit van de feedback profiteert eveneens van criteria die nauwkeurig aansluiten op de uit te voeren assessmenttaak. Nadeel van een taakspecifieke rubric is dat het veel werk is om voor elke assessmenttaak een aparte rubric op te stellen. Verder kunnen dergelijke rubrics niet gedeeld worden met studenten op straffe van het weggeven van de correcte responsen.

Er worden doorgaans twee argumenten gegeven voor de inzet van rubrics in het onderwijs. Het eerste argument houdt verband met de veronderstelde grotere betrouwbaarheid en validiteit van beoordelingen van processen en producten. Het tweede argument gaat over de veronderstelde potentie van rubrics om het leren van studenten te bevorderen. In dit artikel wordt alleen ingegaan op het eerste argument.

Leidt het gebruik van rubrics tot meer betrouwbare en valide beoordelingen?

De professionele uitstraling van rubrics maakt dat veel assessoren vertrouwen hebben in dergelijke instrumenten en geneigd zijn deze te verkiezen boven andere instrumenten. Een typisch geval van face-validity (Holden, 2010) dat wellicht heeft bijgedragen aan de snelle opmars van rubrics in het onderwijs. Maar hoe zit het werkelijk met de beoordelingskwaliteit van rubrics?

Jonsson & Svingby (2007) bestudeerden een groot aantal, voor het merendeel na 2000 gepubliceerde artikelen om na te gaan of het gebruik van rubrics de betrouwbaarheid en de validiteit bevordert. Hieronder wordt achtereenvolgens op hun onderzoeksresultaten ten aanzien van beide kwaliteitsaspecten ingegaan.

Betrouwbaarheid

Van de 75 geanalyseerde artikelen rapporteerden er 70 over de interbeoordelaarsbetrouwbaarheid.

Tabel 1 geeft een samenvatting van de interbeoordelaarsbetrouwbaarheden die Jonsson en Svingby vonden. Deze overziend concludeerden de auteurs dat beoordelen met een rubric waarschijnlijk betrouwbaarder is dan zonder. Het is echter lastig om hun conclusie te accepteren. In de eerste plaats omdat on-

Taakspecifieke rubrics bevatten criteria voor de beoordeling van een bepaalde assessmenttaak

duidelijk blijft hoeveel studies precies interbeoordelaarsbetrouwbaarheden rapporteerden die groter of gelijk zijn aan de in de literatuur vermelde ondergrenzen. De auteurs hebben het over 'meerderheid van de studies', maar meerderheid is een rekkelijk begrip dat bijvoorbeeld 75 procent van de studies kan betreffen, maar ook 51 procent. Een groter probleem is echter dat nergens in het artikel expliciet gesproken wordt over studies waarin vergelijkend onderzoek zou zijn gedaan naar de interbeoordelaarsbetrouwbaarheid van beoordelingen met en zonder rubric. Veel kunnen het er niet geweest zijn, want dat soort studies is schaars. Eén zo'n studie, niet in de review study van Jonsson en Svingby opgenomen, is die van Rezaei en Lovorn (2010). Het gaat om een omvangrijke (n = 361), experimentele studie waarin geen ondersteuning werd gevonden voor de hypothese dat het gebruik van een rubric de betrouwbaarheid van de beoordelingen zou verhogen. Juist het omgekeerde effect werd gevonden, namelijk dat het gebruik van de rubric de betrouwbaarheid van de beoordelingen verlaagde. De onderzoekers vertrouwden deze onverwachte uitkomst omdat het gevonden effect bij elk van de vier onderzochte groepen proefpersonen (71 ≤ *n* ≤ 108) werd gevonden.

Validiteit

De validiteit van onderwijskundige meetinstrumenten heeft altijd minder aandacht gekregen van onderzoekers dan betrouwbaarheid. Dat komt ook tot uiting in de review study van Jonsson en Svingby waar slechts 25 van de 75 opgenomen artikelen rapporteren over één of meer van de aspecten (content, generalizability, external, structural, substantive en

Tabel 1. Samenvatting resultaten review study Jonsson & Svingby (2007).

Methode[1]	Aantal onderzoeken	Range alle (i) en range meerderheid van studies (ii)	Acceptabele waarden volgens literatuur
Consensus estimates	27	i 4 – 100 (%) .20 - .63 (Cohen's Kappa) ii 55 – 75 (%) .40 - .75 (Cohen's Kappa)	≥ 70 (%) ≥ .40 (Kappa)
Consistency estimates	24	i .27 - .98 (Pearson corr.) .50 - .92 (Coëff. alpha) ii .55 - .75 (Pearson corr.) > .70 (Coëff. alpha)	≥ .70 (Pearson corr.) ≥ .80 (Coëff. alpha)
Measurement estimates	19	i .15 - .98 (Generaliseerbaarheidscoëff.)[2] ii .50 - .80 (Generaliseerbaarheidscoëff.)	≥ .80 (General.coëff)

[1]Er zijn diverse statistische methodes voor het bepalen van de interbeoordelaarsbetrouwbaarheid, die grofweg in de volgende categorieën zijn in te delen (Stemler, 2004):
Consensus estimates: Deze maten drukken in een getal de exacte overeenstemming uit tussen de beoordelingen van onafhankelijke beoordelaars. Percentage overeenstemming en Cohen's Kappa zijn bekende voorbeelden.
Consistency estimates: Deze maten geven de samenhang weer tussen de beoordelingen van onafhankelijke beoordelaars. Voor een hoge samenhang is het niet per se nodig dat beoordelaars exact dezelfde scores toekennen. De Pearson correlatie coëfficiënt en Cronbach's alpha zijn bekende voorbeelden.
Measurement estimates: Binnen deze categorie vallen verschillende maten waarvan de meest gebruikte het resultaat zijn van een generaliseerbaarheidsanalyse. Die maakt het mogelijk om rekening te houden met verschillen in beoordelingen van onafhankelijke beoordelaars, zoals bijvoorbeeld veroorzaakt door verschillen in mildheid en strengheid. Ook is het mogelijk om met deze maten de effecten van minder of meer beoordelaars op de hoogte van de beoordelaarsovereenstemming te kwantificeren.

[2]Een generaliseerbaarheidscoëfficiënt is een maat voor de betrouwbaarheid van een toets waarin verschillende variantiebronnen verwerkt kunnen zijn. Een variantiebron die vrijwel altijd in de coëfficiënt verwerkt is, is de mate van overeenstemming tussen beoordelaars.

Rubrics worden gekenmerkt door een grotere mate van standaardisatie

consequential) die volgens de definitie van validiteit onderscheiden kunnen worden aan dit begrip (AERA/APA/NCME, 1999). Terecht vragen Jonsson en Svingby zich af wat het betekent voor de validiteit van een rubric als bijvoorbeeld het content aspect positief geëvalueerd is, maar er over de andere validiteitsaspecten (nog) geen informatie beschikbaar is. Door het ontbreken van onderzoeken die alle aspecten van validiteit bestudeerden en de niet meer dan matige resultaten van de onderzoeken die zich op slechts één of twee aspecten richtten, durven Jonsson en Svingby, anders dan bij het kwaliteitsaspect betrouwbaarheid, aan rubrics niet het voordeel van de twijfel te geven, door te redeneren dat rubrics vanwege hun standaardisatie en gereglementeerd gebruik leiden tot meer valide scores.
Dat deze voorzichtigheid niet onterecht is blijkt uit de al eerder aangehaalde studie van Rezaei & Lovorn (2010), waarin de onderzoekers constateren dat de beoordelingen die de proefpersonen gaven sterk beïnvloed waren door de meer technische kenmerken (zoals grammaticale en spellingsfouten) van het te beoordelen construct (schrijfvaardigheid), terwijl deze kenmerken in de gebruikte rubric een ondergeschikte rol speelden (tien procent van de toe te kennen punten). Anders gezegd: ondanks de standaardisatie en het voorgeschreven gebruik van de rubric hanteerden beoordelaars toch eigen prestatiestandaarden bij de beoordeling van essays. Voor de onderzoekers was dit een reden om vraagtekens te zetten bij de veronderstelling dat gebruik van de rubric zou leiden tot meer valide beoordelingen.

Rubrics: wel of niet inzetten?

De conclusie die uit het hierboven gepresenteerde onderzoek getrokken moet worden, is dat het voorbarig is om te beweren dat beoordelingen op basis van rubrics beter van kwaliteit zijn dan beoordelingen zonder zo'n instrument. Moet daarom het gebruik van rubrics in het onderwijs ontraden worden? Nee, als het gaat om het nemen van eerlijke, controleerbare en deugdelijke beslissingen over het al dan niet voldoende zijn van complexe prestaties van studenten, lijkt de rubric goede papieren te hebben. In vergelijking met de glance-and-grade methodiek worden rubrics gekenmerkt door een grotere mate van standaardisatie, die er in principe voor kan zorgen dat elke kandidaat op identieke wijze beoordeeld wordt. Daarmee is aan een belangrijke voorwaarde voor betrouwbare scores en dus ook voor valide scores voldaan. Of een rubric werkelijk betrouwbare en valide scores produceert, hangt natuurlijk vooral af van de vraag of hij zorgvuldig ontworpen is en op de juiste wijze gebruikt wordt.

Wat het zorgvuldig ontwerp betreft, moet in de eerste plaats gedacht worden aan deugdelijke prestatiecriteria. Een belangrijke vraag in dit verband is of een te beoordelen complexe vaardigheid wel beoordeeld kan worden aan de hand van een standaard set van prestatiecriteria. Dwingen deze prestatiecriteria de assessor niet teveel in de richting van het herkennen van gedrag of eigenschappen van producten die beschreven zijn in de prestatiecriteria ten koste van een betekenisvolle interpretatie die een deskundige assessor kan geven aan de prestatie van een kandidaat (Delandshere & Petrosky, 1998)? Zeker vanuit het oogpunt van validiteit is dit een belangrijk punt. Het kan daarom niet genoeg benadrukt worden hoe belangrijk het is om veel aandacht te besteden aan het afleiden en formuleren van beoordelingsdimensies en prestatie-indicatoren die (a) een adequate operationalisatie zijn van de te beoordelen eigenschap, (b) waar mogelijk in observeerbare en objectieve termen geformuleerd zijn en (c) geaccepteerd worden door alle betrokkenen.

Assessorentraining is een belangrijk punt als het om het juiste gebruik van rubrics gaat. Gegeven de onmogelijkheid om prestatiecriteria te formuleren die geschikt zijn om een groot aantal verschillende prestaties op volstrekt objectieve wijze te beoordelen, is het nodig om assessoren te leren hoe ze de prestatiecriteria dienen te interpreteren. Hoe goed men daarin slaagt zal onderzocht moeten worden in een periodiek onderzoek naar de interbeoordelaarsbetrouwbaarheid die de betreffende assessoren weten te realiseren.

Ten slotte moet worden opgemerkt dat van zorgvuldig ontwerpen van rubrics geen sprake kan zijn als dit de slotactiviteit is van de ontwikkeling van een

Assessorentraining is een belangrijk aspect als het om het juiste gebruik van rubrics gaat

leerplan. Deugdelijke rubrics zijn het resultaat van een nauwkeurige afstemming tussen doelen, onderwijsleeractiviteiten, assessmenttaken en prestatiecriteria. Auteurs als Stiggins (1987) en Baldwin c.s. (2008) beschreven procedures om deze afstemming te realiseren.

Literatuur

American Council on the Teaching of Foreign Languages. (1999). *ACTFL performance guidelines for K-12 learners.* Yonkers, NY: ACTFL.

AERA/APA/NCME (1999). *Standards for educational and psychological testing.* Washington, CD: AERA.

Baldwin, D., Fowles, M., & Livingston, S. (2008). Guidelines for constructed-response and other performance assessments. Princeton (NJ): ETS.

Delandshere, G., & Petrosky, A. (1998). Assessment of complex performances: Limitations of key measurement assumptions. *Educational Researcher, 27*, 2, 14-25.

Groot, A.D. de (1961). *Methodologie: grondslagen van onderzoek en denken in de gedragswetenschappen.* 's-Gravenhage: Mouton.

Holden, R. B. (2010). Face validity. In: Irving B. Weiner, W. Edward Craighead, (Eds): *The Corsini Encyclopedia of Psychology* (4th ed.). Hoboken, NJ: Wiley. pp. 637–638.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. A research review. *Educational Research Review, 2*, 2, 130–144.

Rezaei, A.R. & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Wriging, 15*, 18-39

Stemler, S (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research, and Evaluation, 9*, 4, 1-19

Stiggins, R.J. (1987). Design and development of performance assessments. *Educational Measurement: Issues and Practice, 6*, 33-42.

Straetmans, G.J.J.M. (2014). Toetsen met performance assessment methodieken. In: H. van Berkel, A. Bax & D. Joosten-ten Brinke (Red.): *Toetsen in het hoger onderwijs (3e druk)* Houten: Bohn Stafleu van Loghum.

☐ Dr. Gerard J.J.M. Straetmans is lector Assessment bij het Kenniscentrum Onderwijsinnovatie van Saxion en medewerker van Cito, Arnhem. Email g.j.j.m.straetmans@saxion.nl

Gesignaleerd

De onderwijsinspecteur van nu is een stuk strenger dan vroeger. Toen keek hij vooral en deelde dan zijn oordeel uit, nu mag iedereen ook iets terug zeggen.

De onderwijsinspecteurs stellen vragen en maken aantekeningen. 'Je kijkt de klas rond om te zien of alle kinderen meedoen en betrokken zijn? Gaat de leraar niet te snel, legt hij wel goed uit?

Bron: De Volkskrant, 27 oktober 2015