Paarsgewijs beoordelen van portfolio's beeldende vorming

Hugo Gitsels, Marjolein van Eck en Hans Kuhlemeier

Wat is de beste manier om de portfolio's met beeldend werk van vmbo-leerlingen voor hun Centraal Praktisch Examen (CPE) te beoordelen? Cito-onderzoekers Hugo Gitsels, Marjolein van Eck en Hans Kuhlemeier vergeleken de betrouwbaarheid en bruikbaarheid van paarsgewijs beoordelen met globaal beoordelen en een rubricmodel. In dit artikel beschrijven ze hun bevindingen.

Tussen 2014 en 2016 zijn in een reeks onderzoeken binnen Cito verschillende beoordelingsvormen voor portfolio's vergeleken, zoals het huidige, globale beoordelingsmodel van het CPE vmbo voor beeldende vorming (BV) en beoordeling met een analytisch rubricmodel. Ook zijn de verschillen tussen beoordeling van fysieke en gedigitaliseerde portfolio's onderzocht (Gitsels, Knüppe-Hüsken, Van Beukering, & Kuhlemeier, 2014). Daarnaast is er een literatuurstudie gedaan naar voor- en nadelen van verschillende vormen van beoordeling van beeldend werk (Copini, Van Beukering, & Gitsels, 2016): de holistische of globale benadering, de analytische aanpak en het beoordelen met (visuele) prestatieankers.

Uit deze onderzoeken kwam naar voren dat elke vorm (globaal, analytisch, fysiek en digitaal) eigen voor- en nadelen kent. Vaak gaat het om een spanningsveld tussen validiteit (en efficiëntie) enerzijds en betrouwbaarheid en transparantie anderzijds. In 2017 is daarom een nieuwe beoordelingsvorm onderzocht: paarsgewijs beoordelen van portfolio's. Alvorens nader in te gaan op dit onderzoek beschrijven we eerst de voor- en nadelen van de oudere beoordelingsvormen.

Voor- en nadelen oudere beoordelingsvormen

De globale beoordeling van fysieke portfolio's biedt ruimte aan beoordelaars om alle relevante criteria mee te nemen. De keerzijde is dat niet inzichtelijk is op welke criteria iemand beoordeelt, waardoor de mogelijkheid ontstaat dat niet alle scores op dezelfde gronden zijn gebaseerd (Van Berkel & Bax, 2006).

Het beoordelen van de fysieke werken biedt alle gelegenheid om de schetsen, studies en eindwerkstukken volledig te bestuderen (van afstand, in detail en vanuit verschillende gezichtshoeken). Daar staat tegenover dat deze vorm van beoordelen praktische beperkingen met zich meebrengt die beoordelaarseffecten, zoals normverschuiving in de hand kunnen werken. Bekend is dat beoordelaars geneigd zijn de strengheid van hun beoordelingen aan te passen aan het gemiddelde prestatieniveau van een groep leerlingen.

Het beoordelen volgens gedetailleerde analytische of rubricmodellen maakt de beoordelingscriteria inzichtelijk en zorgt er (in theorie) voor dat alle scores op dezelfde gronden zijn gebaseerd. Ook zorgt het ervoor dat alle relevante aspecten worden meegenomen (Van Berkel & Bax, 2006). Het nadeel is dat het een tijdrovende klus is (Kimbell, 2007) en dat de beoordelaar alleen van tevoren vastgestelde criteria kan meenemen (Van Berkel & Bax, 2006; Van Daal, Lesterhuis, Coertjens, Donche, & De Maeyer, 2016). Ook zijn er kanttekeningen te plaatsen bij de mate waarin beoordelaars eenzelfde interpretatie hebben van de criteria uit het model (Van Daal et al., 2016; Straetmans 2015).

Digitaal beoordelen heeft verschillende voordelen: digitale portfolio's van beeldend werk zijn beter hanteerbaar, waardoor de beoordeling minder tijdrovend is. Ze zijn snel en eenvoudig naast elkaar te zetten, waardoor ze beter

te vergelijken zijn (Dorn & Sabol, 2006). Deze voordelen kunnen helpen om beoordelingseffecten, waaronder de al genoemde normverschuiving, tegen te gaan. Voor het CPE vmbo BV zouden gedigitaliseerde portfolio's in wisselende, niet klas- of schoolgebonden samenstellingen gepresenteerd en beoordeeld kunnen worden (Gitsels et al., 2014).

Uit het onderzoek naar het CPE vmbo BV (Gitsels et al., 2014) bleek dat scores via een gedetailleerd rubricmodel lager uitvallen dan met een globaal model. Ook bleek het gebruik van zo'n rubricmodel een hogere inter-beoordelaars-overeenstemming op te leveren. De Cito-onderzoekers zagen geen verschillen in scores tussen een beoordeling van fysieke en digitale portfolio's. Uit eerder onderzoek (Dorn & Sabol, 2006) bleek dat de scores bij digitale portfolio's consistent zijn met die bij fysiek werk. Ook bleek de beoordelaarsovereenstemming in beide gevallen acceptabel, waarbij de beoordeling van fysiek werk iets meer overeenstemming liet zien.

Paarsgewijs beoordelen

Paarsgewijs beoordelen zou een alternatief kunnen vormen voor de eerdere onderzochte vormen van beoordelen voor het CPE vmbo BV. De methode van paarsgewijs beoordelen werd in 1995 geïntroduceerd door Pollitt en Murray en is gebaseerd op Thurstones wet van vergelijkend oordelen (Lesterhuis, Verhavert, Coertjens, Donche, & De Maeyer, 2017). Uitgangspunt van deze wet is dat een object of een representatie (bijvoorbeeld een essay of kunstwerk) een psychologische indruk maakt op de beschouwer (of beoordelaar) en dat deze indruk per moment kan wisselen. Dit betekent dat ook het oordeel per moment kan verschillen. Thurstone (in Verhavert, De Maeyer, Donche, & Coertjens, 2017) nam aan dat de psychologische indruk niet direct waarneembaar is en dat als objecten geordend kunnen worden op basis van een bepaald kenmerk, deze ordening de psychologische indruk zou volgen. Daarom is deze indruk alleen te meten door een beoordelaar te vragen om twee objecten te vergelijken en te laten bepalen welke het beste is, gelet op een bepaald kenmerk.

Bij paarsgewijs beoordelen krijgt de beoordelaar telkens nieuwe paren voorgelegd. Niet elke beoordelaar hoeft elke vergelijking te maken en niet alle vergelijkingen hoeven te worden gemaakt, zolang elk portfolio maar hetzelfde aantal keer is beoordeeld (Lesterhuis et al., 2017). Uiteindelijk kan op basis van alle beoordelingen een rangordening van de portfolio's worden gemaakt. Paarsgewijs beoordelen is relatief in plaats van absoluut. Er is geen absolute schaal (een beoordelingsmodel) op basis waarvan het portfolio wordt beoordeeld.

Bij complexe taken is de betrouwbaarheid van paarsgewijs beoordelen hoger dan die van het toekennen van scores (Thurstone, in Verhavert et al., 2017). Maar die betrouwbaarheid is sterk afhankelijk van het aantal vergelijkingen waarin een portfolio is betrokken. Uit eerder onderzoek kwamen de volgende

resultaten naar voren: vanaf negen vergelijkingen kwam de betrouwbaarheid uit op .70 of hoger, bij zeventien tot twintig vergelijkingen tussen .80 en .90 en voor een betrouwbaarheid van .98 waren maar liefst 69 vergelijkingen nodig (Lesterhuis et al., 2017). Vanwege het grote aantal benodigde beoordelingen geldt de methode al vanaf het ontstaan als inefficiënt (Verhavert et al., 2017). Maar in vergelijking met het scoren van tal van deelaspecten wordt de methode juist wel als efficiënt bestempeld (Kimbell, 2007). Paarsgewijs oordelen kan snel en intuïtief geschieden en is daardoor minder tijdrovend (Lesterhuis et al., 2017).

De belangrijkste reden om voor paarsgewijs vergelijken te kiezen is de veronderstelde validiteit van deze methode. Alastair Pollitt (2009, p. 2) noemt het een veel natuurlijkere en directere wijze van beoordelen: 'What could be more valid than judging that one piece of work is more creative than another? Or more effective? Or just better? And if many judges agree that the same one is better, isn't that the best evidence for validity we could ask for?' Vooraf opgestelde beoordelingscriteria, zoals bij een analytische beoordeling, kunnen beperkend werken en de beoordeling reductionistisch maken. Dit geldt zeker voor creatieve opdrachten waarbij de uitkomsten 'open' en moeilijk te voorspellen zijn. Hierdoor kan de validiteit in gevaar komen en daarmee ook de betrouwbaarheid (Lesterhuis et al., 2017).

Opzet van het onderzoek

Onderzoeksvragen

In ons onderzoek stonden de volgende vragen centraal:

1. Welke betrouwbaarheid levert paarsgewijs beoordelen op voor het CPE vmbo BV? En hoe verhoudt deze betrouwbaarheid zich tot de beoordeling volgens het huidige en het rubricmodel?
2. In hoeverre zijn de scores van (digitaal) paarsgewijs beoordelen vergelijkbaar met de (fysieke) beoordeling volgens het huidige en het rubricmodel?
3. Wat zijn voor beoordelaars doorslaggevende criteria in hun afweging bij het paarsgewijs beoordelen?
4. Hoe ervaren beoordelaars de methode van paarsgewijs beoordelen?

Op basis van eerder onderzoek (Lesterhuis et al., 2017) verwachtten we dat de betrouwbaarheid (alpha/SSR) hoger uitvalt dan bij de huidige beoordeling. We gingen daarbij uit van zeventien tot twintig inschattingen per representatie (dat wil zeggen een set van twee portfolio's) in het experiment paarsgewijs beoordelen. Verder verwachtten we op basis van onderzoek van Dorn en Sabol (2006) dat scores bij digitale portfolio's consistent zijn met die bij fysiek werk.

Het was moeilijk om op voorhand de door beoordelaars genoemde relevante criteria in te schatten. Uit de literatuur blijkt dat een veelheid aan criteria een rol speelt bij de beoordeling van beeldend werk, hoewel deze wel terug te brengen zijn tot een beperkt aantal dimensies. Groenendijk, Damen, Haanstra en Van Boxtel (2015) onderscheiden in totaal elf dimensies, waarbij de volgende het vaakst voorkomen in beoordelingsmodellen: originaliteit/creativiteit, beeldaspecten, onderzoek/experiment, techniek, concept/idee en zelfreflectie. Gezien de centrale opdrachtstelling van het CPE vmbo BV, met het accent op proces (onderzoek en reflectie) én product (uitvoering) was onze verwachting dat beoordelaars vooral criteria zouden noemen voor de dimensies onderzoek/experiment, originaliteit/creativiteit en techniek.

Ten slotte verwachtten we dat de beoordelaars de methode als passend (valide) beoordelen, maar misschien ook als arbeidsintensief (op basis van Verhavert et al., 2017; Politt, 2009).

Keuze voor tool

Voor de uitvoering van het experiment zijn verschillende aanbieders van tools voor paarsgewijs beoordelen vergeleken. We kozen uiteindelijk voor de tool van D-Pac vanwege de gebruiksvriendelijkheid ervan. Bovendien bood deze tool ruimte aan beoordelaars om hun keuze kort toe te lichten, wat voor beantwoording van onze de derde onderzoeksvraag van belang was.

Selectie portfolio's en beoordelaars

In ons experiment betrokken we dezelfde 45 porfolio's als in ons eerdere onderzoek (Gitsels et al., 2014). Hiervoor zijn destijds twaalf scholen benaderd met het verzoek portfolio's – en de daaraan toegekende scores – van het CPE vmbo BV 2012 te leveren.

Aan het experiment deden zeventien beoordelaars mee. Het betrof docenten die werden geworven via de website van Digischool in de Vakcommunity beeldende vakken. Ook stuurden we een e-mail aan docenten beeldende vorming uit het adressenbestand van Cito. In totaal hebben we zo'n driehonderd docenten benaderd. Voorwaarden om deel te nemen waren een ruime ervaring als beoordelaar van beeldend werk van leerlingen in het voortgezet onderwijs en bekendheid met het CPE vmbo BV.

De beoordelingssessie

In de digitale tool kregen de beoordelaars allereerst de opdracht getoond behorende bij het CPE vmbo BV uit 2012. Deze opdracht konden ze tijdens het beoordelen nog eens teruglezen. Daarnaast wezen we de beoordelaars erop dat ze in de beoordeling de volgende fases uit de opdracht dienden te betrekken: oriënteren, onderzoeken, vaststellen en uitvoeren.

Elk portfolio bestond uit een pdf-bestand met foto's van het beeldend werk van leerlingen: schetsen, studies, proefjes en eindwerkstukken, aangevuld met de door de leerlingen ingevulde opdrachtboekjes.

Elke beoordelaar diende dertig vergelijkingen uit te voeren, hiervoor hadden zij een maand de tijd. Ze konden de beoordeling gespreid uitvoeren. Behalve het kiezen van de beste uit twee portfolio's dienden ze bij elke vergelijking aan te geven wat het doorslaggevend criterium was om het ene portfolio boven de andere te verkiezen. De toewijzing van portfolio's gebeurde automatisch en aselect en was zo georganiseerd dat elke portfolio uiteindelijk in evenveel vergelijkingen werd betrokken.

Na afloop van de beoordelingssessie werd de beoordelaars per e-mail gevraagd naar hun ervaring van het experiment en deze wijze van beoordelen.

Uitkomsten van het onderzoek

Aantal beoordelingen

Uiteindelijk zijn 44 van de 45 portfolio's betrokken, één portfolio bleek incompleet en viel om die reden uit. Zeventien beoordelaars traden aan, maar niet allemaal voltooiden ze de volledige sessie van dertig vergelijkingen. Uiteindelijk zijn er 423 vergelijkingen voorgelegd en hebben er 417 beoordelingen plaatsgevonden. Dit betekent dat elke beoordelaar gemiddeld 24,5 beoordeling heeft uitgevoerd.

Betrouwbaarheid

De beoordelingen zijn geanalyseerd volgens het Bradly-Terry-Luce model dat resulteert in een logit schatting voor de portfolio's. De Scale Separation Reliability (SSR) van Rasch (ook bekend als Rasch alpha) wordt gerapporteerd als maat voor de betrouwbaarheid. De SSR kan binnen de methode van paarsgewijs beoordelen dienen als maat voor de interne consistentie en betrouwbaarheid, en als maat voor de inter-beoordelaarsovereenstemming (Verhavert et al., 2017). De SSR werd in dit experiment geschat op 0,76. Gedurende het experiment zijn ook tussentijdse SSR-waarden genoteerd. In tabel 1 zijn deze weergegeven.

Tabel 1. SSR-waarden tijdens het experiment

Aantal beoordelingen	**SSR**
240	0,68
286	0,70
380	0,73
417	**0,76**

Vergelijken we deze uitkomst met de geschatte interne betrouwbaarheid van het CPE vmbo BV uit 2012, 0,59 (Cronbachs alpha, Examenverslag (Cito, 2012)), dan zien we dat beoordelen volgens paarsgewijs beoordelen hoger uitkomt, dus betrouwbaarder is.

Qua beoordelaarsovereenstemming zit paarsgewijs beoordelen tussen het huidige beoordelingsmodel (ICC 0,60) en het rubricmodel (ICC 0,83) (Gitsels et al., 2014).

Een kritische kanttekening die we hierbij moeten maken, is dat de opzet van de beoordeling in deze onderzoeken verschilt: per portfolio geldt een verschillend aantal beoordelingen en dat leidt tot verschillen in betrouwbaarheid. Zo is de betrouwbaarheid van 0,60 bij het huidige beoordelingsmodel gebaseerd op twee beoordelingen per portfolio. De betrouwbaarheid van 0,83 bij het rubricmodel is eveneens gebaseerd op twee beoordelingen per portfolio. De digitale beoordeling van het rubricmodel (ICC 0,92) is gebaseerd op negen à tien beoordelingen per portfolio (Gitsels et al., 2014). Bij het paarsgewijs beoordelen was sprake van 417 beoordelingen. Bij elk van deze beoordelingen waren telkens twee portfolio's betrokken. Er is dus 834 keer een portfolio beoordeeld en er waren 44 portfolio's in totaal, gemiddeld is elk portfolio negentien keer beoordeeld.

Correlatie van scores

Om uitspraken te kunnen doen over de vergelijkbaarheid van scores bij paarsgewijs (digitaal) beoordelen, het rubricmodel en de huidige vorm van (fysiek) beoordelen bij het CPE vmbo BV hebben we de correlaties tussen de totaalscore per leerling van de drie beoordelingsmethoden berekend (zie tabel 2). Deze geven weer in hoeverre de drie methoden kandidaten op dezelfde wijze ordenen van laag naar hoog vaardig.

Als scores zijn de percentages van de maximumscore gebruikt zoals toegekend aan de portfolio's in het onderzoek uit 2014 volgens het huidige en het rubricmodel. Deze maximumscore kwam tot stand door de toegekende deelscores per onderdeel bij elkaar op te tellen. Voor de scores bij paarsgewijs beoordelen is uitgegaan van de gerapporteerde latente vaardigheidscores.

Tabel 2. De correlaties tussen de totaalscore per kandidaat

	Score huidig	**Score rubric**	**Score paarsgewijs**
Score huidig	1	0,739	0,96
Score rubric	0,739	1	0,672
Score paarsgewijs	0,96	0,672	1

Uit de tabel is af te lezen dat de correlatie tussen scores met het rubricmodel en paarsgewijs beoordelen 0,67 bedraagt. Voor het huidige model en paarsgewijs beoordelen bedraagt deze 0,96. Deze correlaties zijn respectievelijk middelmatig en zeer hoog te noemen. De gevonden correlaties zijn significant op 1%-niveau.

Gehanteerde criteria

Beoordelaars dienden bij elke vergelijking het doorslaggevend criterium te vermelden. Om hierover uitspraken te kunnen doen is op de notities van de beoordelaars een inhoudsanalyse uitgevoerd. We hebben deze gecodeerd naar de al eerder genoemde dimensies: originaliteit/creativiteit, beeldaspecten, onderzoek/experiment, techniek, concept/idee en zelfreflectie (Groenendijk et al., 2015).

Beoordelaars noteerden vaak meer criteria, ondanks de opdracht om alleen het doorslaggevende criterium te noteren (zie tabel 3).

Tabel 3. Het aantal notities van beoordelaars per dimensie

Dimensie	**Aantal notities**	**Percentage**
Originaliteit/creativiteit	74	13%
Beeldaspecten (in het eindwerkstuk)	141	25%
Onderzoek/experiment (proces)	264	48%
Techniek	27	5%
Concept/idee	19	3%
Zelfreflectie	18	3%
Overig	12	2%
Totaal	**555**	**100%**

Bij bijna de helft van alle beslissingen was een criterium uit de dimensie onderzoek/experiment (proces) doorslaggevend. In negatieve formuleringen gebruikten beoordelaars de volgende kwalificaties: 'oppervlakkig onderzoek', 'mager proces', 'beperkt onderzoek', 'veilige keuzes', 'proces ontbreekt' en 'weinig ontwikkeling'. Positieve kwalificaties en criteria waren onder meer: 'meer technisch en vormgevend onderzoek', 'variërend', 'betere schetsen', 'ijverig', 'diepgaand', 'interessant proces', 'durf' en 'doelgericht'.

In een kwart van de gevallen was een criterium uit de dimensie beeldaspecten (in het eindwerkstuk) doorslaggevend. In deze dimensie gebruikten beoordelaars overwegend positieve kwalificaties, zoals: 'beeldend technisch interessant', 'aspecten kloppen', 'sterk beeld', 'meer ruimtelijkheid', 'interessant geschilder[d] (gevarieerder in techniek en compositie en afwisseling van motieven)', 'bewuste toepassing beeldaspecten' en 'beeldende aspecten als kleur, ruimte en uitsnede zijn op originele en sterke wijze toegepast'.

In ruim tien procent van de gevallen ging het om de dimensie originaliteit/creativiteit. Beoordelaars gebruikten vaak termen als 'spannend' en 'origineel' of, in negatieve formuleringen, 'cliché' of 'voorspelbaar'.

Vergelijken we deze uitkomsten met de mate waarin onderdelen uit het beoordelingsmodel van het CPE gewicht in de schaal leggen, dan zien we in grote lijnen overeenkomsten. Zo bepaalt het onderdeel 'Uitvoeren werkstuk' voor 50 procent de totaalscore in het huidige beoordelingsmodel. Bij paarsgewijs beoordelen wordt 43 procent van de uitkomst bepaald door de hiermee corresponderende dimensies originaliteit/creativiteit, beeldaspecten (in het eindwerkstuk) en techniek.

In het huidige beoordelingsmodel vallen de volgende onderdelen onder de dimensie onderzoek: oriënteren, beeldend onderzoeken en vaststellen ontwerp. Samen bepalen deze onderdelen 28 procent van de totaalscore. Bij paarsgewijs is dat 48 procent en lijkt onderzoek dus een grotere rol te hebben in het geheel.

Ervaringen van beoordelaars

Ervaringen van beoordelaars zijn per e-mail geïnventariseerd. Alle beoordelaars ontvingen na afloop een standaardmail met de vraag hoe ze het experiment hadden ervaren. Daarbij vroegen we hoeveel tijd iemand had geïnvesteerd en hoe zij het paarsgewijs beoordelen als methode vonden. Ook tijdens het proces van beoordelen bereikten ons soms berichten van beoordelaars. Opmerkingen uit deze e-mails hebben we meegenomen in het totaal.

Uit alle reacties bleek dat ruim de helft van de beoordelaars de tijdsinvestering fors vond. Dit is in lijn met de bevindingen uit eerder onderzoek waarin de inefficiëntie van de methode wordt benoemd (Verhavert et al., 2017). De tool hield de geïnvesteerde tijd automatisch bij en kwam per beoordelaar uit op gemiddeld zo'n elf uur. Dit betreft de brutotijd, de tijd dat een sessie actief was. Hierbij is dus geen rekening gehouden met de mogelijkheid dat de beoordelaar de tool open liet staan, maar niet actief aan het beoordelen was. Op basis van de mediaan gaf de beheerder van de tool een schatting van de actief geïnvesteerde tijd. De meest voorkomende tijdsinvestering per vergelijking is twee minuten en 45 seconden. Uitgaande van deze tijdsinvestering voor alle 24,5 beoordelingen van een beoordelaar bedraagt de geïnvesteerde tijd per beoordelaar 67 minuten. In de reactie achteraf zeiden beoordelaars gemiddeld zo'n tweeëneenhalf tot drie uur bezig te zijn geweest. Dit komt neer op afgerond zeven minuten per beoordeling.

Daarnaast maakten enkele beoordelaars een opmerking over verschillen in ordening in de portfolio's. Om de beoordeling te optimaliseren zou het goed zijn als alle pdf's op een gelijke wijze zijn geordend en allemaal compleet zijn, zo merkten ze op. Het betrof hier verschillen in ordening van de opgavenboekjes, waarin de leerlingen hun proces bewaken en keuzes verantwoorden.

De beoordelingstool vond men prettig werken en zeer gebruiksvriendelijk. Over de methode van het paarsgewijs beoordelen toonden de beoordelaars zich positief, ze noemden deze 'geschikt' en 'interessant'.

Samenvatting, interpretatie en kanttekeningen

Hieronder vatten we per onderzoeksvraag de resultaten samen.

1. Welke betrouwbaarheid levert paarsgewijs beoordelen op voor het CPE vmbo BV? En hoe verhoudt deze betrouwbaarheid zich tot de beoordeling volgens het huidige en het rubricmodel?
   De betrouwbaarheid van paarsgewijs beoordelen in dit onderzoek werd geschat op 0,76. Dat is hoger dan de geschatte interne betrouwbaarheid van het CPE vmbo BV uit 2012 (0,59). Qua beoordelaarsovereenstemming zit paarsgewijs beoordelen tussen het huidige model (0,60) en het rubricmodel (0,83) in. Hierbij dient opgemerkt te worden dat per methode het aantal beoordelingen per portfolio verschilt en dit is van invloed op de betrouwbaarheid.

2. In hoeverre zijn de scores van (digitaal) paarsgewijs beoordelen vergelijkbaar met de (fysieke) beoordeling volgens het huidige en het rubricmodel?
   De correlatie tussen scores met het rubricmodel en paarsgewijs beoordelen is middelmatig (0,67) en zeer hoog voor het huidige model en paarsgewijs beoordelen (0,96).

   Een verklaring voor deze zeer hoge correlatie is dat in beide gevallen de beoordeling niet geschiedt volgens gespecificeerde criteria. Er is dus een vergelijkbare wijze van (globaal) beoordelen, terwijl het rubricmodel juist analytisch is. Overigens leidt het verschil in conditie (digitaal of fysiek) in dit geval nauwelijks tot verschillen in scores.

   Een mogelijke verklaring voor de lagere correlatie tussen paarsgewijs beoordelen en het rubricmodel is de verschillende mate waarin iemand onderdelen of dimensies meeweegt in de beoordeling. Bij rubrics moet de beoordelaar alle opgegeven onderdelen meewegen, bij paarsgewijs beoordelen is hij daarin vrij en blijken beoordelaars zich hoofdzakelijk te richten op drie dimensies.

3. Wat zijn voor beoordelaars doorslaggevende criteria in hun afweging bij het paarsgewijs beoordelen?
   Bij bijna de helft van alle beslissingen was een criterium uit de dimensie onderzoek/experiment (proces) doorslaggevend. De volgende criteria werden daarbij gehanteerd: diepgang, breedte, variatie, durf, doelmatigheid en ontwikkeling.

   In een kwart van de gevallen was een criterium uit de dimensie beeldaspecten (in het eindwerkstuk) doorslaggevend. Beoordelaars spraken hier vooral in algemene zin over het wel of niet geslaagd toepassen van beeldende aspecten (kloppend, sterke toepassing, bewust gebruik, originele toepassing van beeldaspecten). In iets meer dan tien procent van de gevallen ging het om een criterium in de dimensie originaliteit/creativiteit, met termen als ‘spannend’ en ‘origineel’ of juist ‘cliché’ of ‘voorspelbaar’.

   Er zijn grote inhoudelijke overeenkomsten met de beoordelingen volgens het huidige model. Alleen de dimensie onderzoek lijkt bij paarsgewijs beoordelen een grotere rol te spelen.

4. Hoe ervaren beoordelaars de methode van paarsgewijs beoordelen?
   De beoordelaars vonden de tijdsinvestering fors. Op basis van de door hen gerapporteerde tijdsinvestering vergde de beoordeling (van een duo) gemiddeld zeven minuten. Omdat voor voldoende betrouwbaarheid een substantieel aantal beoordelingen nodig is, blijkt de methode van paarsgewijs beoordelen arbeidsintensief. Ter vergelijking: bij de reguliere afname en beoordeling van het CPE vmbo BV zijn per school twee beoordelaars actief. Voor de portfolio’s in de dataset van het experiment (afkomstig van vier scholen) waren dat in totaal acht beoordelaars. Voor het experiment paarsgewijs beoordelen was het betrokken aantal beoordelaars twee keer zo groot.

   De beoordelingstool vond men prettig werken en de methode van het paarsgewijs beoordelen werd ‘geschikt’ en ‘interessant’ genoemd.

Al met al mogen we concluderen dat de methode van paarsgewijs beoordelen, inclusief een kort commentaar of motivatie voor de beslissing, perspectief biedt voor de toekomst. Het commentaar bij de beslissing is wellicht ook in te zetten voor formatieve evaluatie. Op dit punt is verder onderzoek nodig: hoe dient een beoordelaar commentaar te geven, zodat dit bruikbare feedback is voor leerlingen? En wat is het effect van deze feedback op een volgende prestatie? Daarnaast verdient het aanbeveling te onderzoeken hoe de tijdsinvestering bij het paarsgewijs beoordelen zich precies verhoudt tot die bij het beoordelen met een globaal en/of analytisch model.

Hugo Gitsels en **Marjolein van Eck** werken als toetsdeskundige bij Cito.
E Hugo.Gitsels@cito.nl

Hans Kuhlemeier werkt als onderzoeker bij de afdeling psychometrisch onderzoek van Cito.

Literatuur

Cito. (2012). *Examenverslag vmbo gl/tl, beeldende vakken.* www.cito.nl/onderwijs/voortgezet%20onderwijs/centrale_examens/examenverslagen/oude_verslagen, geraadpleegd op 22 augustus 2019.

Copini, H., Van Beukering, A., & Gitsels, H. (2016). *Rapport beoordeling Centraal Praktisch Eindexamen – vmbo Beeldend*. Ongepubliceerd onderzoeksverslag Cito.

Dorn, C. M., & Sabol, F. R. (2006). The effectiveness and use of digital portfolios for the assessment of art performances in selected secondary schools. *Studies in Art Education, 47*(4), 344-362.

Gitsels, H., Knüppe-Hüsken, M., Van Beukering, A., & Kuhlemeier, H. (2014). Maken en meten: de beoordeling van het CPE beeldende vakken vmbo. *Cultuur+Educatie, 14*(41), 26-42.

Groenendijk, T., Damen, M-L., Haanstra, F., & Van Boxtel, C. (2015). *Assessment in kunsteducatie.* Eindrapport NWO Review studie 411-12-228.

Kimbell, R. (2007). E-assessment in Project e-scape. *Design and Technology Education, 1*2(2), 66-76.

Lesterhuis, M., Verhavert, S., Coertjens, L., Donche, V., & De Maeyer, S. (2017), Comparative judgement as a promising alternative to score competences. In E. Cano, & G. Ion (Eds.), *Innovative Practices for Higher Education Assessment and Measurement* (pp. 119-138). Hershey, PA: IGI Global.

Pollitt, A. (2009). *Abolishing marksism and rescuing validity*. Paper presented at the IAEA Conference, Brisbane, Australia.

Straetmans, G. J. J. M. (2015). Gaan rubrics ons helpen om beter te beoordelen? *Examens,* (4), 20-25.

Van Berkel, H., & Bax, A. (2006). *Toetsen in het hoger onderwijs.* Houten: Bohn Stafleu van Loghum.

Van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., & De Maeyer, S. (2016). Validity of comparative judgement to assess academic writing: examining implications of its holistic character and building on a shared consensus. *Assessment in Education: Principles, Policy & Practice, 26*(1), 59-74.

Verhavert, S., De Maeyer, S., Donche, V., & Coertjens, L. (2017). Scale separation reliability: What does it mean in the context of comparative judgment? *Applied Psychological Measurement*, *42*(6), 428-445.