Renske Bouwer, Maarten Goossens, Anneleen Viona Mortier, Marije Lesterhuis en Sven De Maeyer

Een comparatieve aanpak voor peer assessment: leren door te vergelijken

Peer assessment in het hoger onderwijs

Peer assessment is een van de effectiefste manieren om complexe vaardigheden en competenties aan te leren (Boud, Cohen & Sampson, 1999; Dochy, Segers & Sluijsmans, 1999). Toch blijkt niet elke vorm van peer assessment even effectief te zijn. In dit hoofdstuk laten we zien hoe dat komt en introduceren een nieuwe en veelbelovende methode voor peer assessment: paarsgewijze vergelijking. We beschrijven waarom deze vergelijkende methode voor studenten gemakkelijker is dan bijvoorbeeld het werken met analytische scoringsvoorschriften, zoals rubrics, en hoe het vergelijken van elkaars werk aanzet tot leren en tot kwaliteitsvolle feedback. Tot slot bespreken we de rol van de docent bij deze vorm van peer assessment.

Eerder onderzoek heeft laten zien dat studenten bij peer assessment op twee manieren leren: door het *geven* en het *ontvangen* van feedback (Topping, 2009). Vooral het geven van feedback is een krachtig middel, doordat het aanzet tot een diepe cognitieve verwerking (Lundstrom & Baker, 2009). Bij het geven van feedback leren studenten namelijk kritisch het werk van medestudenten evalueren, waarbij ze verwoorden waarom iets goed of minder goed is en wat mogelijke verbeterpunten zijn. Op deze manier ontwikkelen ze evaluatievaardigheden, die ook nodig zijn om de kwaliteit van hun eigen werk in te schatten en te reguleren (Nicol & Macfarlane-Dick, 2006). Het is immers makkelijker om het werk van iemand anders te evalueren dan dat van jezelf (Orsmond, Merry & Reiling, 2002). Daarnaast ontvangt iedere student feedback op zijn eigen werk, waarmee hij inzicht krijgt in waar hij op dat moment staat in zijn ontwikkeling. Deze feedback kan hij vervolgens gebruiken om zijn werk te verbeteren. In een peer assessment zijn studenten dus een bron van instructie en terugkoppeling voor elkaar. Deze wisselwerking tussen studenten leidt niet alleen tot betere prestaties, maar ook tot een verhoogde motivatie. Daarnaast verlaagt het de werkdruk van docenten, omdat zij niet al het werk zelf van commentaar hoeven te voorzien (Boud et al., 1999).

Peer assessment is echter lang niet altijd effectief (Falchikov & Goldfinch, 2000). Zo blijkt er een matchingeffect te zijn, waarbij de effectiviteit afhangt van de samenstelling van de feedbackpartners (Patchan & Schunn, 2016). Waar sterke studenten leren van zowel goed als slecht presterende peers, leren de minder sterke studenten vooral wanneer ze gekoppeld worden aan peers die wat niveau betreft met hen over-

eenkomen. Daarnaast valt of staat de kracht van peer assessment met de kwaliteit van de feedback die studenten elkaar geven (Patchan, Schunn & Correnti, 2016). Volgens Hattie en Timperley (2007) sluit goede feedback aan bij de individuele behoeften van studenten en is het gericht op de taak of op het onderliggende proces om tot de taak te komen, en niet op de student zelf. Het biedt studenten inzicht in waar ze staan ten opzichte van het gewenste leerdoel, met concrete aanwijzingen hoe ze dat doel (stapsgewijs) kunnen bereiken. De uitdaging hierbij is om niet te veel feedback te geven en een goede balans te creëren tussen positieve en negatieve feedback, zodat studenten gemotiveerd zijn om het de volgende keer (nog) beter aan te pakken. Dit is voor docenten al een lastige klus (Bouwer & Koster, 2017; Hattie & Timperley, 2007), laat staan voor studenten die de competentie nog onder de knie moeten krijgen. Het is daarom belangrijk om de studenten ondersteuning te bieden.

Rubrics bieden te weinig houvast voor studenten

In een overzichtsstudie naar peer assessment in het hoger onderwijs wordt aanbevolen om studenten van tevoren goed te instrueren over de beoordelingscriteria (Dochy et al., 1999). Vaak wordt daarbij gebruikgemaakt van rubrics, waarbij een complexe vaardigheid is opgedeeld in deelvaardigheden en niveaus, en kwaliteitscriteria zijn beschreven voor elk van de deelvaardigheden en niveaus. Maar geven deze beschrijvingen wel voldoende houvast aan studenten? Bekijk eens het volgende voorbeeld uit een rubric voor het schrijven van een onderzoeksverslag in het tweede jaar van een bacheloropleiding. In deze rubric worden de volgende prestatieniveaus onderscheiden voor het correct formuleren van de probleem- en vraagstelling:

- *Onvoldoende:* De probleem- of vraagstelling ontbreekt, is niet helder, niet afgebakend, of niet haalbaar.
- *Voldoende:* De probleem- of vraagstelling is concreet, maar kan beter afgebakend en/of scherper geformuleerd worden.
- *Goed:* De probleem- of vraagstelling is concreet, goed afgebakend en scherp geformuleerd.

Het is maar de vraag of deze beschrijving studenten echt helpt om de kwaliteit van het werk van hun medestudent goed te beoordelen. Want wanneer is een probleem- of vraagstelling concreet genoeg? En wat wordt er precies bedoeld met 'goed afgebakend' of 'scherper geformuleerd'? Zolang studenten geen concreet voorbeeld hebben bij deze definities, blijven het abstracte beschrijvingen en kunnen studenten er niet goed mee uit de voeten (Nicol & Macfarlane-Dick, 2006; Orsmond et al., 2002; Sadler, 2009).

Er zijn daarom onderzoekers die stellen dat studenten betrokken moeten worden bij het formuleren van de criteria (Fraile, Panadero & Pardo, 2017). Op deze manier worden de criteria in de woorden van de studenten zelf geformuleerd en kunnen zij ze beter internaliseren. Cruciaal daarbij is dat zij voorbeelden zien en met elkaar in discussie gaan over wat goed werk onderscheidt van minder goed werk. Het is

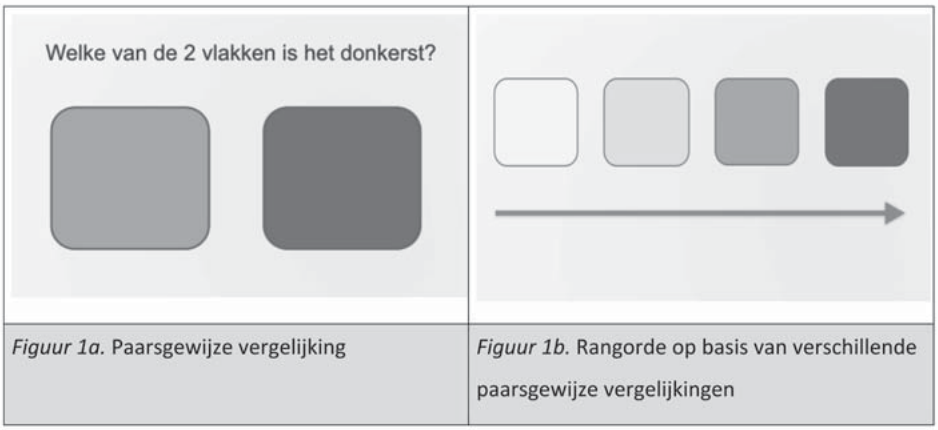
echter niet vanzelfsprekend om een proces in gang te zetten waarbij alle studenten actief betrokken worden in het formuleren en bediscussiëren van de criteria (Carless & Kam Ho Cham, 2016). Ook lopen docenten er vaak tegenaan dat studenten voorbeelden klakkeloos kopiëren zonder zich de kwaliteitscriteria eigen te maken.

De kracht van paarsgewijs vergelijken

Sinds kort is de alternatieve methode van paarsgewijze vergelijking in opkomst, waarbij studenten het werk van hun medestudenten niet langer hoeven te vergelijken met abstracte kwaliteitscriteria in een rubric, maar direct met elkaar, in willekeurig samengestelde paren. Bij elk paar hoeft de student enkel aan te geven welk werk van hogere kwaliteit is en waarom. Doordat elk werk in meerdere vergelijkingen terugkomt en de vergelijkingen willekeurig worden verspreid over studenten, leiden deze vrij eenvoudige beslissingen uiteindelijk tot kwaliteitsvolle feedback. Ook is deze manier van peer assessment een belangrijke leerervaring voor studenten, omdat ze verschillende producten van uiteenlopende kwaliteit langs zien komen.

Het principe van paarsgewijs vergelijken is niet nieuw. In een publicatie uit 1927 beschreef psychofysicus Louis L. Thurstone al hoe objecten op basis van een serie van paarsgewijze vergelijkingen, bijvoorbeeld welk object donkerder is, op een (subjectieve) schaal kunnen worden geplaatst (zie Figuur 1a en 1b). Pas aan het begin van de 21e eeuw is deze methode ook met succes toegepast op het beoordelen van complexe vaardigheden (Pollitt, 2004). Zie Figuur 2a en 2b voor een voorbeeld van het paarsgewijs vergelijken van essays.

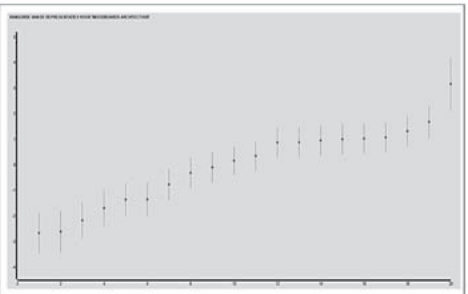
Het beoordelen van competenties door middel van paarsgewijze vergelijking is een stuk eenvoudiger dan absoluut beoordelen aan de hand van criteria in een rubric. Wanneer men de kwaliteit van een product evalueert, wordt immers altijd al (onbewust) een vergelijking gemaakt met een eigen interne standaard of met eerder gezien werk (Laming, 2004). Op basis van deze referentie bepaalt men vervolgens



Figuur 1a. Paarsgewijze vergelijking

Figuur 1b. Rangorde op basis van verschillende paarsgewijze vergelijkingen

Figuur 2a. Paarsgewijze vergelijking bij schrijfvaardigheid: welke tekst is beter?



Figuur 2b. Resultaat van alle vergelijkingen, waarbij de teksten zijn gerangschikt van lage naar hoge kwaliteit

of iets goed is of juist niet. De vergelijkende manier van beoordelen sluit daarmee dus goed aan bij de natuurlijke neiging van mensen om dingen met elkaar te vergelijken. Daarnaast blijkt dat mensen hier ook veel beter in zijn dan in het maken van absolute vergelijkingen (Gill & Bramley, 2013).

D-PAC: Een digitaal platform voor paarsgewijze vergelijking

Het willekeurig samenstellen van paren en het berekenen van een rangorde op basis van alle gemaakte vergelijkingen is een vrij ingewikkelde en tijdrovende klus. Gelukkig zijn er digitale tools beschikbaar waarin dit proces is geautomatiseerd. In Engeland zijn twee tools ontwikkeld: *Digital Assess* (www.digitalassess.com) en *No More Marking* (NMM; www.nomoremarking.com). Recentelijk hebben wij aan de Universiteit Antwerpen, in samenwerking met de Universiteit Gent en IMEC, een soortgelijke tool ontwikkeld: D-PAC, oftewel *Digital Platform for the Assessment of Competences* (www.d-pac.be).

De drie tools onderscheiden zich van elkaar door de algoritmes waarmee de paren worden samengesteld. Zo maakt *Digital Assess* gebruik van een adaptief algoritme, waarbij de informatie van elke vergelijking wordt gebruikt om de volgende paren samen te stellen. Het gevolg hiervan is dat er maar zo'n acht tot twaalf vergelijkingen per product nodig zijn om tot een stabiele rangorde te komen. Er zijn echter twijfels over de werkelijke betrouwbaarheid van deze adaptieve methode, aangezien de uiteindelijke rangorde sterk afhangt van de eerste reeks vergelijkingen die zijn gemaakt. Producten die in deze eerste ronde (ten onrechte) als slechtste zijn aangemerkt, zullen nooit meer in de top van de rangorde uitkomen. Dit leidt tot een vertekend beeld van de betrouwbaarheid (Bramley, 2015). NMM en D-PAC maken daarom gebruik van een random algoritme waarbij de paren steeds op willekeurige wijze worden samengesteld. Om dit toch zo efficiënt mogelijk te laten verlopen wordt er bij D-PAC wel rekening gehouden met het aantal keren dat een product al is vergeleken, zodat er alleen paren worden samengesteld van producten waarvoor nog extra informatie

nodig is. *D-PAC* heeft daarnaast nog extra features specifiek voor peer assessments. Zo kunnen studenten hun eigen werk anoniem uploaden en kan ervoor gekozen worden dat zij hun eigen werk niet terug zien komen in de vergelijkingen. Uit onderzoek van Jones en Alcock (2014) blijkt namelijk dat wanneer studenten hun eigen werk zien in een vergelijking, ze dit bijna altijd als beste kiezen. Studenten kunnen in de rangorde (het resultaat van alle paarsgewijze vergelijkingen) wel zien waar hun eigen werk is gepositioneerd. Ook kunnen studenten tijdens het vergelijken het werk van hun medestudenten van feedback voorzien, die vervolgens anoniem wordt teruggekoppeld.

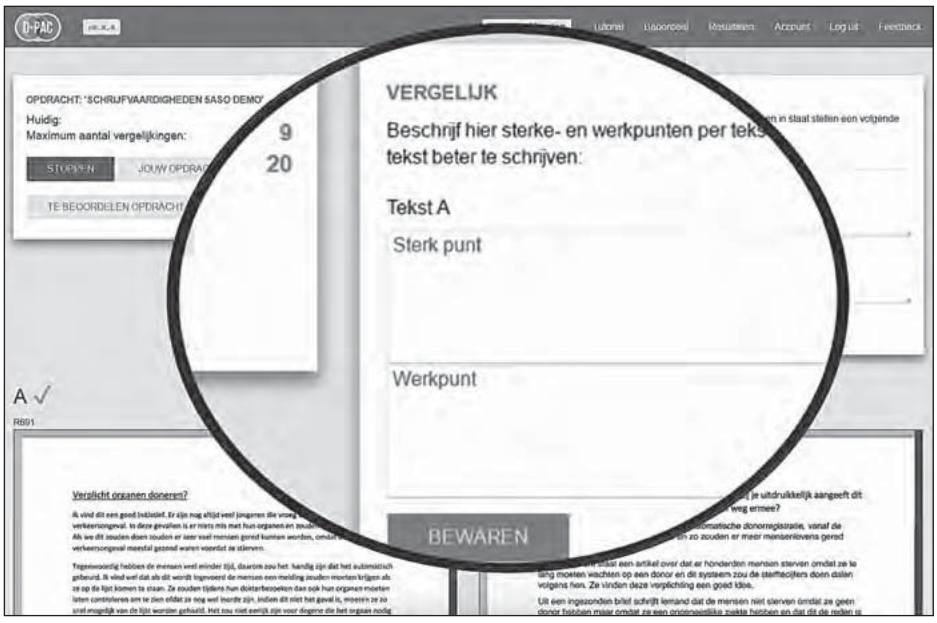
Leren door te vergelijken

Studenten komen door elkaars werk te vergelijken dus gemakkelijker tot een goed kwaliteitsoordeel. Daarnaast draagt het proces van vergelijken bij aan het leerproces (Gentner, 2010). In D-PAC lijkt dit leren door te vergelijken op twee manieren plaats te vinden. Ten eerste zien studenten in een reeks vergelijkingen een reeks van goede en slechte voorbeelden voorbijkomen. Hierdoor ontwikkelen zij stap voor stap hun eigen referentiekader voor kwaliteit. Ten tweede leren studenten van het actief vergelijken van deze verschillende voorbeelden. Onderzoekers uit Berlijn hebben in een reeks van experimenten laten zien dat vergelijken tot een dieper begrip leidt dan het een voor een bekijken van voorbeelden (Pachur & Olsson, 2012). Dit komt doordat studenten in een vergelijking actief op zoek gaan naar overeenkomsten en verschillen. Wanneer een student bijvoorbeeld een reeks essays vergelijkt, kan in de eerste vergelijking het verschil in structuur tussen de teksten opvallen, in een tweede vergelijking de overtuigingskracht van de argumenten en in een derde vergelijking spellings- of grammaticafouten. Op deze manier kunnen studenten zich een concreet beeld vormen bij abstracte of complexe kwaliteitscriteria en kunnen ze zich deze gaandeweg eigen maken. Orsmond en collega's (2002) hebben in eerder onderzoek inderdaad laten zien dat het beoordelen van voorbeelden van medestudenten leidt tot een beter begrip van kwaliteitscriteria. Ook bleek uit dit onderzoek dat studenten door concrete voorbeelden beter weten wat er van hen verwacht wordt en meer gerichte en relevante feedback geven aan hun medestudenten.

Leren door het geven van feedback

Bij peer assessments in D-PAC is het ook mogelijk dat studenten na elke vergelijking feedback geven op de plus- en minpunten van elk van de producten in die vergelijking. Zie Figuur 3 voor een voorbeeld van hoe deze feedbackmodule er in D-PAC uitziet. Van het expliciet formuleren van feedbackpunten zouden studenten ook kunnen leren. Het formuleren van feedback wordt immers verondersteld actief bij te dragen aan het leereffect van peer assessment (Lundstrom & Baker, 2009). Om erachter te komen of dit ook voor paarsgewijze vergelijking geldt, hebben we een

onderzoek opgezet met 96 studenten van de master Opleidings- en Onderwijswetenschappen aan de Universiteit Antwerpen. Deze studenten vergeleken de kwaliteit van videofragmenten van wetenschappelijke interviews met elkaar. Ze werden hierbij in drie willekeurige groepen opgedeeld. De eerste groep hoefde enkel de interviews met elkaar te vergelijken, de tweede groep moest na elke vergelijking hun keuze beargumenteren en de derde groep moest na elke vergelijking de goede en slechte aspecten van de interviews benoemen. Daarnaast beschreven alle studenten vooraf en achteraf wat zij onder een goed wetenschappelijk interview verstonden. Uit de resultaten bleek dat de studenten in alle drie de groepen na de paarsgewijze vergelijkingen significant meer kwaliteitsaspecten van een wetenschappelijk interview konden benoemen dan voor de vergelijkingen. De kennistoename verschilde echter niet tussen de drie condities, wat in dit geval betekent dat op korte termijn het expliciteren van feedback niet direct heeft geleid tot meer expliciete kennis dan enkel het vergelijken op zich. Uiteraard kan de peer feedback wel een effect hebben op de feedbackontvanger.

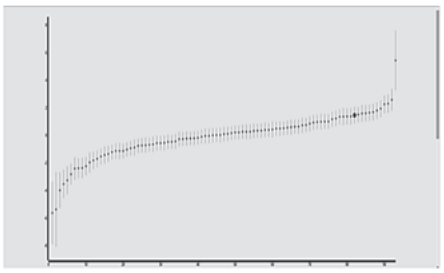


Figuur 3. De manier waarop studenten in D-PAC feedback geven op twee producten in een vergelijking

Welke feedback levert het op en wat leren studenten daarvan?

Na het voltooien van een reeks paarsgewijze vergelijkingen in D-PAC ontvangen studenten feedback over hun eigen werk. In D-PAC zijn er twee verschillende vormen van feedback: feedback over het huidige prestatieniveau door een positie in de rangorde (zie Figuur 4a) en feedback over specifieke plus- en verbeterpunten (zie

Figuur 4b). Door middel van de feedback in de rangorde krijgen studenten inzicht in het niveau van hun huidige prestatie ten opzichte van die van de medestudenten en/of ten opzichte van het gewenste doel (als er bijvoorbeeld voorbeelden in de vergelijking zijn meegenomen die de verschillende prestatieniveaus representeren, de zogenoemde ankers). Wanneer studenten op hun werk in de rangorde klikken, krijgen ze de peer feedback over hun specifieke plus- en verbeterpunten te zien. Deze peer feedback wordt anoniem weergegeven en bestaat uit geschreven commentaar bij de taken van de studenten. Deze combinatie van feedback, waarin zowel feedback over het huidige prestatieniveau ten opzichte van een gewenst doel wordt gegeven als concrete verbeterpunten om dichter bij het gewenste doel te komen, blijkt cruciaal te zijn voor het leren van feedback (Hattie & Timperley, 2007).



Figuur 4a. Feedback in de rangorde: hoe presteert de student ten opzichte van zijn peers en ten opzichte van het gewenste doel? Het rode bolletje geeft aan waar het werk van de student door de paarsgewijze vergelijkingen is geplaatst.

Figuur 4b. Feedback over specifieke plus- en verbeterpunten: wat gaat er al goed, en wat kan er beter? Deze peer feedback wordt geanonimiseerd weergegeven.

Hoe waardevol is de feedback over het prestatieniveau?

Bij peer assessments komt het geregeld voor dat studenten twijfelen aan de feedback die ze van medestudenten krijgen. Dit kan ook optreden bij deze holistische methode van paarsgewijs vergelijken, aangezien studenten nogal vrij worden gelaten bij het maken van de vergelijkingen. Dat dit tot onzekerheid bij studenten kan leiden, zagen we bijvoorbeeld bij een peer assessment in de master Opleidings- en Onderwijswetenschappen aan de Universiteit Antwerpen. Hier vergeleken studenten de kwaliteit van elkaars statistiekopdrachten. Studenten gaven na afloop aan te hebben geleerd van het vergelijken en van de feedback in de rangorde. Er waren echter ook studenten die gedurende het vergelijken hadden gezien dat niet alle medestudenten de opdracht hadden begrepen. Deze studenten vonden de ontvangen

feedback over hun positie in de rangorde daarom niet geloofwaardig en gaven aan meer behoefte te hebben aan feedback van de docent.

Bovenstaande bevindingen stimuleerden ons om de kwaliteit van de comparatieve peer feedback verder te onderzoeken. Zo vond er een peer assessment plaats met dertig studenten van de opleiding Interieur, Architectuur, Stedenbouw aan de Universiteit Antwerpen, waarbij we hebben gekeken naar de betrouwbaarheid van de vergelijkingen door studenten. Hierbij vroegen we ons af of studenten in paarsgewijze vergelijkingen hetzelfde werk als beter bestempelen. Daarnaast zijn we nagegaan in hoeverre studenten en ervaren docenten hierin overeenkomen: komen zij tot een gelijkaardige rangorde, en letten zij daarbij op dezelfde aspecten? Deze informatie geeft ons inzicht in de validiteit van het peer assessment, althans als we aannemen dat docenten goede beoordelingen maken. Voor deze peer assessment hadden studenten de opdracht gekregen om een moodboard te ontwikkelen rondom een bepaalde emotie, en deze via D-PAC paarsgewijs te vergelijken. Alle moodboards werden in totaal achttien keer vergeleken en van feedback voorzien. De groep studenten werd in twee groepen van vijftien studenten gesplitst, zodat we de resultaten van deze twee groepen met elkaar konden vergelijken. De betrouwbaarheid van de oordelen in beide groepen was hoog: .81 voor groep 1 en .73 voor groep 2. Ook was er een sterke correlatie van .75 tussen de oordelen van de twee groepen. Studenten kwamen dus sterk overeen in welke moodboards zij de betere vonden. Dezelfde moodboards werden ook door een groep van vijf ervaren docenten beoordeeld. Zij kwamen tot een betrouwbaarheid van .71. De rangordes van de beide studentengroepen correleerden .65 met de rangorde van de docenten.

Ook bij de opleiding Industriële Ingenieurswetenschappen aan de KU Leuven hebben we de betrouwbaarheid en validiteit van de vergelijkende peer feedback onderzocht. Hierbij waren 25 studenten betrokken, die elk een Entity-Relationship (ER)-schema hadden ontwikkeld. In ER-schema's wordt de relatie tussen verschillende entiteiten in een database grafisch weergegeven. Elk van deze schema's werd in totaal zeventien keer paarsgewijs vergeleken door de groep studenten. Dezelfde ER-schema's werden ook paarsgewijs vergeleken door een groep van vier ervaren docenten. De resultaten waren ook hier positief: de betrouwbaarheid van de oordelen van zowel de studenten als docenten was hoog, respectievelijk .73 en .77, en de oordelen van studenten en die van docenten kwamen overeen ($r = .62$, $p < .001$).

Collega-onderzoekers Jones en Alcock (2014) en Jones en Wheadon (2015) uit Engeland vonden vergelijkbare resultaten in peer assessments naar het wiskundig begrip van studenten. Ook hier bleek de betrouwbaarheid van de studentenoordelen hoog, namelijk .72 en .85, en kwamen ze sterk overeen met die van de experts (correlatie respectievelijk .77 en .72). In de laatstgenoemde studie werd ook gekeken naar de betrouwbaarheid van oordelen wanneer de peers elkaars werk op een absolute manier beoordelen, dus zonder dat het werk onderling vergeleken werd. Het bleek dat de betrouwbaarheid van de absolute beoordelingswijze nagenoeg nul was, zowel voor de studenten als voor de docenten. Ook deden de studenten er langer over om tot een absoluut oordeel te komen.

Kortom, studenten voorzien elkaar via paarsgewijze vergelijkingen van kwaliteitsvolle feedback. Deze feedback geeft echter enkel inzicht in hoe ze op dat moment presteren, en dat ook nog eens relatief ten opzichte van elkaar. Wat leren ze daar dan precies van? De docent kan ervoor kiezen om het leereffect van deze feedback in peer assessments te versterken door voorbeeldproducten van verschillende kwaliteitsniveaus aan de vergelijkingen toe te voegen. Dat kunnen producten zijn van studenten uit het voorgaande jaar die volgens docenten representatief zijn voor het gemiddelde niveau of voor bijvoorbeeld de grens tussen goed en excellent. Als ook deze voorbeelden (of ankers) op de rangorde worden geplaatst, kunnen studenten nagaan of hun eigen werk voldoet aan de minimale doelstellingen van het vak. Extra toelichting bij de voorbeelden waarin de docent beschrijft waarom dat product van dat specifieke kwaliteitsniveau is, geeft studenten concrete aanknopingspunten om hun eigen werk te verbeteren. Ook kan de docent ervoor kiezen om in een peer assessment meerdere werken van dezelfde student mee te nemen. Daarmee krijgen studenten inzicht in de mogelijke leerwinst die ze hebben gemaakt over tijd. De focus ligt dan niet meer op hoe de student presteert ten opzichte van zijn medestudenten, maar juist hoe de student presteert ten opzichte van zichzelf. Daarnaast kunnen studenten het werk van hun peers (anoniem) bekijken, om inzicht te krijgen welke werken een beetje beter zijn dan dat van hen en welke het beste beoordeeld zijn. Dat zijn immers goede voorbeelden om van te leren.

Hoe waardevol is peer feedback over plus- en verbeterpunten?

Naast feedback in de rangorde over het niveau van de prestatie is het ook mogelijk om feedback te ontvangen over specifieke plus- en verbeterpunten van de geleverde prestatie. Studenten blijken vooral deze feedback relevant te vinden, aangezien ze de inhoudelijke suggesties kunnen gebruiken voor het verbeteren van hun werk (Mortier, Lesterhuis, Vlerick & De Maeyer, 2015). Maar hoe kwaliteitsvol is deze feedback? Om dit te onderzoeken hebben we gekeken naar de verschillen tussen comparatieve feedback en die van de traditionele manier van feedback waarbij producten een voor een worden becommentarieerd. In de lerarenopleiding Nederlands aan de Odisee Hogeschool in Sint-Niklaas gaven elf tweedejaars studenten feedback aan eerstejaarsstudenten op de door hen geschreven formele brieven (Mortier, Lesterhuis, Vlerick, Donche & De Maeyer, 2016). De tweedejaars studenten werden geïnstrueerd de feedback zo te formuleren dat de eerstejaarsstudenten het konden gebruiken om hun brief aan te passen. Van de elf tweedejaarsstudenten startten zes studenten met de comparatieve feedback en de andere vijf met het geven van feedback via opmerkingen in de tekst, waarna ze van aanpak wisselden. De resultaten laten zien dat de feedback in de comparatieve conditie gericht was op zowel hogereordeaspecten van de brief (inhoud, structuur en stijl) als op lagereordeaspecten (taalfouten en conventies), in tegenstelling tot de conditie waarbij feedback in de tekst werd gegeven: daarbij was de feedback bijna volledig gericht op lagereordeaspecten. Deze resultaten laten zien dat door steeds twee producten met elkaar te

vergelijken en studenten niet de mogelijkheid te geven om opmerkingen in de tekst te plaatsen, studenten zich meer richten op de essentiële en complexe aspecten van tekstkwaliteit. Het is immers vooral de hogere orde feedback die studenten vooruit helpt bij het schrijven (Patchan, Schunn & Correnti, 2016).

Ook bij de opleiding Kinesitherapie aan de Universiteit Hasselt hebben we onderzocht in welke mate een comparatieve peer assessment waardevolle feedback oplevert voor studenten. Bij deze opleiding schrijven studenten elk jaar meerdere patiëntendossiers. De zeventig studenten in dit onderzoek kregen op twee momenten in het jaar feedback op hun patiëntendossiers door middel van paarsgewijze peer assessment. De docent van deze opleiding was benieuwd in hoeverre de peer feedback overeenkwam met haar oordeel en of ze daarbij nog extra feedback moest geven. Haar reactie was positief: ze vond de rangorde uit het peer assessment een goede weergave van de oplopende kwaliteit van de patiëntendossiers. Ook vond ze de peer feedback over specifieke plus- en verbeterpunten relevant en zeer behulpzaam voor studenten met het oog op vervolgopdrachten. Uit verder onderzoek bleek ook dat meer dan de helft van de studenten aangaf de peer feedback succesvol te hebben gebruikt voor de tweede opdracht later in het jaar. Dit hing echter wel af van kenmerken van de studenten. Studenten die intrinsiek gemotiveerd waren om de feedback te begrijpen, scoorden ook hoger op de tweede taak later in het jaar. Studenten die de feedback niet hadden meegenomen naar de tweede taak gaven als reden dat ze de feedback tegenstrijdig of onduidelijk vonden, of ze waren het al vergeten.

De ondersteunende rol van de docent

Dat feedback met deze beoordelingsmethode als tegenstrijdig kan worden ervaren is bijna niet uit te sluiten. Het is immers zo dat elk product in meerdere vergelijkingen terugkomt, waardoor verschillende studenten feedback geven op hetzelfde product. Deze studenten kunnen dezelfde aspecten net anders waarderen, waardoor tegenstrijdige feedback ontstaat. Deze variatie in feedback kan echter ook juist aanzetten tot leren. Bij competenties of vaardigheden wordt de kwaliteit van een prestatie namelijk niet bepaald door één enkel aspect, maar juist door een combinatie van aspecten, en er zijn nu eenmaal verschillen tussen personen in wat zij goed of minder goed vinden. Voor studenten betekent dit dat zij de feedback dus niet klakkeloos kunnen overnemen, maar zelf moeten nagaan welke feedback ze relevant vinden en kunnen gebruiken voor het verbeteren van hun werk. Dit vraagt wel veel van studenten, en de docent moet dan ook extra ondersteuning bieden bij het begrijpen en implementeren van de feedback. Zo gaf de docent van de opleiding Kinesitherapie aan de Universiteit Hasselt zelf geen feedback meer, maar monitorde ze wel de kwaliteit van de gegeven peer feedback. Waar nodig wees ze de studenten in colleges op problemen of tegenstrijdigheden die ze tegenkwam in de peer feedback.

Docenten van de opleiding Meertalige Professionele Communicatie aan de Universiteit Antwerpen boden studenten op een soortgelijke manier ondersteuning tij-

dens een peer assessment. In deze opleiding werd de methode van paarsgewijze vergelijking ingezet voor peer feedback op slechtnieuwsbrieven. De docenten kwamen er bij de resultaten achter dat studenten tijdens het vergelijken niet op alle relevante aspecten hadden gelet, waardoor de studenten tot een net wat andere rangorde kwamen dan de docenten. In de daaropvolgende les hebben de docenten de rangorde en feedback besproken, waarbij ze duidelijk maakten wat er van de studenten verwacht wordt en wat de kwaliteitsaspecten zijn van een slechtnieuwsbrief. In het komende collegejaar willen deze docenten graag dezelfde werkwijze toepassen, en na het feedbackcollege nog een peer assessment uitvoeren om te kijken of dit de kwaliteit van de feedback heeft bevorderd.

Discussie

In dit hoofdstuk hebben we met verschillende praktijkvoorbeelden en onderzoeksbevindingen laten zien dat paarsgewijze vergelijking het leren van studenten kan stimuleren. Het vergelijken zet aan tot een dieper leerproces, en het zien van een grote reeks aan voorbeelden van peers kan inspirerend werken. Daarnaast resulteert de vergelijkende methode in verschillende vormen van kwaliteitsvolle feedback, die gebruikt kunnen worden voor formatieve doeleinden. Zo geeft feedback over de rangorde studenten inzicht in het niveau van hun prestaties en kunnen studenten met de feedback over specifieke plus- en verbeterpunten hun prestaties gericht verbeteren. Uit verschillende onderzoeken is gebleken dat de peer feedback op basis van paarsgewijs vergelijken grotendeels overeenkomt met de feedback van docenten, wat de feedback nuttig en relevant maakt.

Vervolgonderzoek is noodzakelijk om beter te begrijpen hoe en wanneer paarsgewijze vergelijking het effectiefst kan worden ingezet voor peer assessment. Zo weten we dat het actief vergelijken zorgt voor meer inzicht in de belangrijkste kwaliteitscriteria, maar de precieze werking hiervan is nog een *black box*. Daarnaast hebben we tot op heden peer assessment enkel toegepast in de loop van een cursus, wanneer studenten al een eigen product hebben gemaakt. Studenten zouden echter ook veel kunnen leren van het zien en vergelijken van voorbeelden voordat ze aan hun eigen product beginnen. Wat het mogelijke leereffect is van een peer assessment in D-PAC aan het begin van een cursus onderzoeken we momenteel in samenwerking met Artevelde Hogeschool. Tot slot is het nog onduidelijk wat de leereffecten zijn van paarsgewijze vergelijking op korte en lange termijn, en in hoeverre het tot betere prestaties leidt dan reguliere vormen van peer assessment.

Wat in eerder onderzoek ook nog onvoldoende aan bod is gekomen, is voor welk type studenten deze vorm van peer assessment het beste werkt. Leren sterke en minder sterke studenten evenveel van het zien en vergelijken van het werk van medestudenten? Het observeren van voorbeelden kan een belangrijke boost geven aan het zelfvertrouwen van studenten, want als ze ervaren dat peers tot goede producten kunnen komen, dan zouden zij dat ook moeten kunnen (Bandura, 1986; Zimmerman, 2000). Tegelijkertijd kan het ook demotiverend werken om het werk van peers

te zien, zeker wanneer studenten onzeker zijn over hun eigen kunnen en helemaal onder aan de rangorde uitkomen. Verder onderzoek naar de rol van individuele verschillen in het leereffect van comparatieve peer assessment is daarom noodzakelijk.

Implicaties voor de praktijk

Ondanks de behoefte aan verder onderzoek naar de leereffecten van comparatieve peer feedback kunnen we toch een paar belangrijke implicaties voor de onderwijspraktijk meegeven. D-PAC biedt een online platform waarin studenten eenvoudig hun werk kunnen uploaden, vergelijken en rijke peer feedback krijgen. Docenten besparen hiermee kostbare tijd, die ze kunnen steken in het ondersteunen van studenten tijdens het geven en ontvangen van feedback. Deze ondersteuning lijkt essentieel: verschillende praktijkvoorbeelden in dit hoofdstuk hebben laten zien dat studenten het soms moeilijk vinden om de ontvangen feedback te plaatsen of te accepteren, helemaal als peers tegenstrijdige feedback geven over mogelijke verbeterpunten.

Welke ondersteuning kunnen docenten het beste bieden? Allereerst is het belangrijk om studenten vooraf goed te instrueren over de methode. Denk aan vragen als: waarom ga je vergelijken, wat doe je als twee producten erg op elkaar lijken, waar resulteren de vergelijkingen in, hoe geef je effectieve feedback op zowel plus- en verbeterpunten? Maar ook achteraf hebben studenten behoefte aan begeleiding bij het interpreteren van de feedback, bijvoorbeeld bij de vraag hoe ze de informatie in de rangorde effectief kunnen gebruiken en hoe ze kunnen omgaan met (tegenstrijdige) feedback. Goede instructie voor- en achteraf zal naast het leren ook de motivatie tot deelname van studenten verhogen.

Docenten kunnen ook een belangrijke rol spelen in het optimaliseren van de kwaliteit van de feedback. Zo kan een docent vooraf voorbeelden selecteren die representatief zijn voor de verschillende kwaliteitsniveaus (onvoldoende, gemiddeld, goed) en deze meenemen in een peer assessment. Hierdoor krijgen deze voorbeelden een ankerfunctie in de rangorde en krijgen studenten niet enkel feedback over hoe ze zich ten opzichte van hun peers verhouden, maar ook over hun positie ten opzichte van het gewenste doel.

Ook kunnen docenten in D-PAC de kwaliteit van de peer feedback monitoren. Hierbij zijn vooral de volgende informatiebronnen van waarde:

1. De betrouwbaarheid van de rangorde: dit geeft inzicht in hoeverre studenten het eens zijn over welk product van hogere kwaliteit is. Als de betrouwbaarheid laag is, betekent dit dat niet iedereen hetzelfde werk goed of slecht vindt. Met behulp van extra analyses is het mogelijk om erachter te komen welke studenten afwijken van de gemiddelde groepsconsensus. Zijn dit de studenten die zelf ook minder goed presteren? Hebben zij wellicht extra uitleg nodig?
2. De rangorde: deze geeft informatie over welke producten studenten beter of minder goed vinden. Komt dit overeen met de mening van docenten?

3. De specifieke feedback over plus- en minpunten: dit geeft informatie over aspecten waar studenten tijdens het vergelijken op letten. De docent kan deze aspecten vergelijken met de vooropgestelde kwaliteitscriteria: komen deze overeen, of zijn er aspecten waar studenten nog onvoldoende naar kijken? Ook kan de docent de gegeven feedback gebruiken om een indruk te krijgen van wat studenten volgens henzelf al wel goed doen (pluspunten) en wat nog beter kan (verbeterpunten).
4. De feedback gegroepeerd per student: door na te gaan welke student welke feedback heeft ontvangen en gegeven, krijgt de docent inzicht in het prestatieniveau van individuele studenten en de aspecten waar zij op letten tijdens het vergelijken van elkaars werk. Deze informatie kunnen docenten gebruiken om individuele studenten extra uitleg of begeleiding te geven.

Wanneer de docent op basis van deze informatie concludeert dat studenten nog onvoldoende zicht hebben op de kwaliteitscriteria en op wat er van hen verwacht wordt, kunnen zij tijdens de colleges met studenten in gesprek gaan, met de voorbeelden in de rangorde en de geschreven feedback als vertrekpunt. Onderzoek van Fraile en collega's (2017) toont het belang aan van het co-creëren en bediscussiëren van de beoordelingscriteria.

Kortom, D-PAC biedt veel mogelijkheden voor kwaliteitsvolle peer assessments. Een belangrijke rol is weggelegd voor de docent in het instrueren en begeleiden van studenten, zowel vooraf bij het maken van de vergelijkingen als achteraf bij het interpreteren van de feedback en het monitoren van de kwaliteit ervan. Verder onderzoek in de praktijk is nodig om de mogelijkheden van deze methode nog verder te exploreren.

Referenties

Bandura, A. (1986). *Social Foundations of Thought and Action: A Social Cognitive Theory.* Englewood Cliffs, NJ: Prentice Hall.

Boud, D., Cohen, R., & Sampson, J. (1999). Peer learning and assessment. *Assessment & Evaluation in Higher Education, 24*(4), 413-426.

Bouwer, R., & Koster, M. (2017). *Bringing Writing Research into the Classroom. The Effectiveness of Tekster, a Newly Developed Writing Program for Elementary Students* (proefschrift). Utrecht: Universiteit Utrecht.

Bramley, T. (2015). *Investigating the Reliability of Adaptive Comparative Judgment.* Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.

Carless, D., & Kam Ho Cham, K. (2016). Managing dialogic use of exemplars. *Assessment & Evaluation in Higher Education, 42*(6), 930-941.

Dochy, F., Segers, M., & Sluijsmans, D.M.A. (1999). The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher Education, 24*(3), 331-350.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*(3), 287-322.

Fraile, J., Panadero, E., & Pardo, R. (2017). Co-creating rubrics: The effects on self-

regulated learning, self-efficacy and performance of establishing assessment criteria with students. *Studies in Educational Evaluation, 53,* 69-76.

Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science, 34*(5), 752-775.

Gill, T., & Bramley, T. (2013). How accurate are examiners' holistic judgements of script quality? *Assessment in Education, 20*(3), 308-324. doi:10.1080/0969594X.2013.779229.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112.

Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education, 39*(10), 1774-1787.

Jones, I., & Wheadon, C. (2015). Peer assessment using comparative and absolute judgement. *Studies in Higher Education, 39*(10), 1774-1787.

Laming, D.R.J. (2004). *Human Judgement: The Eye of the Beholder.* Londen: Thomson Learning.

Lundstrom, K., & Baker, W. (2009). To give is better than to receive: The benefits of peer review to the reviewer's own writing. *Journal of Second Language Writing, 18*(1), 30-43.

Mortier, A.V., Lesterhuis, M., Vlerick, P., Donche, V., & Maeyer, S. De (2016, juni). *Comparative Judgment-Based Feedback versus the Common Practice: Similarities and Differences.* Paper gepresenteerd op de conferentie Assessment in Higher Education, Manchester, UK.

Mortier, A.V., Lesterhuis, M., Vlerick, P., & Maeyer, S. De (2015). Comparative judgment within online assessment: Exploring students feedback reactions. *Proceedings of Communications in Computer and Information Science, 571,* 69-79.

Nicol, D.J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199-218.

Orsmond, P., Merry, S., & Reiling, K. (2002). The use of exemplars and formatie feedback when using student derived making criteria in peer and self-assessment. *Assessment & Evaluation in Higher Education, 27*(4), 309-323.

Pachur, T., & Olsson, H. (2012). Type of learning task impacts performance and strategy selection in decision making. *Cognitive Psychology, 65*(2), 1-34.

Patchan, M.M., & Schunn, C.D. (2016). Understanding the effects of receiving peer feedback for text revision: Relations between author and reviewer ability. *Journal of Writing Research, 8*(2), 227-265.

Patchan, M.M., Schunn, C.D., & Correnti, R.J. (2016). The nature of feedback: How peer feedback features affect students' implementation rate and quality of revisions. *Journal of Educational Psychology*. Advance online publication, doi:10.1037/edu0000103.

Pollitt, A. (2004, juni). *Let's stop marking exams*. Paper gepresenteerd op de conferentie International Association of Educational Assessment, Philadelphia, PA.

Sadler, D.R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education, 34*(2), 159-179.

Thurstone, L.L. (1927). A law of comparative judgment. *Psychological Review, 34*(4), 273-286.

Topping, K.J. (2009). Peer assessment. *Theory Into Practice, 48*(1), 20-27.

Zimmerman, B.J. (2000). Self-efficacy: An essential motive to learn. *Contempory Educational Psychology, 25*(1), 82-91.