Competenties kwaliteitsvol beoordelen met D-PAC

■ **Sven de Maeyer**
Sven de Maeyer is hoogleraar bij het departement Opleidings- en Onderwijswetenschappen aan de Universiteit Antwerpen en promotor van het D-PAC project.

■ **Renske Bouwer**
Renske Bouwer is onderzoekscoördinator en postdoc onderzoeker voor D-PAC bij het departement Opleidings- en Onderwijswetenschappen aan de Universiteit Antwerpen.

■ **Roos Van Gasse**
Roos Van Gasse is doctoraal onderzoeker voor D-PAC bij het departement Opleidings- en Onderwijswetenschappen aan de Universiteit Antwerpen.

■ **Maarten Goossens**
Maarten Goossens is wetenschappelijk medewerker voor D-PAC bij het departement Opleidings- en Onderwijswetenschappen aan de Universiteit Antwerpen.
Email: Maarten.Goossens@uantwerpen.be

Eind jaren 70 deed competentiegericht onderwijs zijn intrede in het onderwijs. In deze onderwijsvorm gaat het om de integratie van kennis, vaardigheden en houdingen en het kunnen toepassen hiervan in betekenisvolle en authentieke situaties. Het doel is om mensen beter voor te bereiden op de snel veranderende maatschappij. Een veelgehoorde kritiek op het competentiegericht onderwijs is echter dat de evaluatiemethodes blijven hangen op het niveau van eenvoudig te meten kennisvragen en geen rekening houden met de context waarin de competentie tot uiting komt. Op deze manier schiet competentiegericht onderwijs zijn doel voorbij. Want hoe weet je hoe competenties zich ontwikkelen als je ze niet op de juiste manier meet? In dit artikel gaan we in op de belangrijkste struikelblokken bij het beoordelen van competenties en laten we zien hoe D-PAC een veelbelovend alternatief is.

Mijn vijf is niet de vijf van mijn collega

Docenten maken bij het beoordelen van competenties meestal gebruik van een absolute criteriumgerichte beoordelingsaanpak, waarbij ze scores geven per leerlingproduct. Bij deze aanpak kunnen verschillende beoordelaarseffecten optreden: zowel tussen beoordelaars als binnen beoordelaars (Erkens, 2015). Het is moeilijk om een score toe te kennen omdat je de kwaliteit van een product van nature vergelijkt met producten die je al eerder gezien hebt, of met je idee over hoe een product er zou moeten uitzien. Als gevolg hiervan kunnen er volgorde-effecten optreden, waarbij het oordeel afhankelijk is van de volgorde waarin wordt beoordeeld (Meuffels, 1994). Zo kan een product van middelmatige kwaliteit bijvoorbeeld een lagere score krijgen wanneer het volgt na een reeks van goede producten dan wanneer het volgt na een reeks slechte producten. Daarnaast zijn oordelen sterk persoonsgebonden. Elke beoordelaar heeft zijn eigen interne standaard van wat goed of niet goed is en welke scores daarbij horen. Zo hebben sommige beoordelaars de neiging om vooral scores rond het gemiddelde te geven en zijn sommige beoordelaars strenger dan anderen (Erkens, 2015).

In de praktijk blijken analytische beoordelingsmethoden echter geen goede oplossing voor het probleem met absolute oordelen

Competenties zijn toch meer dan de som van criteria?

Voor meer overeenstemming tussen beoordelaars wordt in de praktijk vaak gebruik gemaakt van criterialijsten of rubrics. Deze analytische beoordelingsinstrumenten hebben als doel om voor zowel docenten als leerlingen duidelijk te maken welke criteria belangrijk zijn bij (het beoordelen van) een bepaalde competentie. Bij rubrics wordt daarnaast per criterium een beschrijving gegeven van de verschillende prestatieniveaus. Zo wordt bijvoorbeeld in een rubric voor argumentatief schrijven gedefinieerd wanneer de argumentatie onvoldoende, voldoende of goed is. Door dit ook voor andere criteria van de competentie te doen probeert men meer overeenstemming tussen beoordelaars te krijgen, en dus de betrouwbaarheid van de absolute oordelen te verhogen (Jonsson & Svingby, 2007).

In de praktijk blijken analytische beoordelingsmethoden echter geen goede oplossing voor het probleem met absolute oordelen. Ook wanneer criteria vooraf worden gespecificeerd zijn er nog grote verschillen tussen beoordelaars. Dit is met name het geval bij deelaspecten van complexe competenties als kritisch denken, creativiteit, of schrijfvaardigheid (Lumley, 2002). Want wat betekent creatief kleurgebruik precies en wanneer is de structuur van een tekst nu precies voldoende? Vanuit validiteitsoogpunt is het moeilijk om elke competentie in verschillende aspecten op te delen. Scores voor deelaspecten blijken dan ook onderling hoog te correleren (De Glopper, 1988). Daarnaast laat de generaliseerbaarheid van analytische oordelen naar de competentie in zijn geheel te wensen over (Van den Bergh, De Maeyer, Van Weijen, & Tillema, 2012).

Een alternatieve beoordelingsmethode: paarsgewijs vergelijken

Zelfs met analytische beoordelingsmethoden lijkt het dus onmogelijk om op een betrouwbare en valide manier een absoluut oordeel te vellen. Elke beoordeling is uiteindelijk een vergelijking met eerder gezien werk of met een interne standaard van wat goed of niet goed is. Een alternatief voor de absolute wijze van beoordelen zijn beoordelingen die uitgaan van dit principe van paarsgewijs vergelijken. In deze aanpak worden producten steeds willekeurig in paren met elkaar vergeleken. De beoordelaar hoeft enkel een uitspraak te doen over welk product beter is met betrekking tot de te beoordelen competentie. Beoordelaars kunnen vooraf wel instructie krijgen over welke criteria ten grondslag liggen aan een competentie, maar de beoordeling is altijd door de kwaliteit van het product in zijn geheel te beoordelen. Zo kan het bijvoorbeeld dat bij argumentatieve schrijfopdrachten de structuur bij de ene tekst heel goed is,

maar dat er wel inhoudelijke onjuistheden in de tekst staan, terwijl bij de andere tekst de inhoud klopt, maar de structuur slordig is. Het is dan aan de beoordelaar om een keuze te maken welke tekst overtuigender is. Het kan dan natuurlijk dat de ene beoordelaar meer op inhoud let dan de andere beoordelaar. Waar dit bij traditionele beoordelingsmethoden een probleem oplevert (want verschillen tussen beoordelaars leidt tot minder betrouwbare oordelen), zijn dit soort verschillen bij de methode van paarsgewijs vergelijken juist positief. De vergelijkingen worden namelijk niet door één beoordelaar, maar door meerdere beoordelaars gemaakt, waardoor de uiteindelijke plek op de rangorde wordt bepaald door het samenspel van alle criteria en niet door een criterium die in de rubric toevallig meer gewicht krijgt. Paarsgewijze vergelijking laat dus toe dat de beoordelaar de kwaliteit van de gehele competentie kan evalueren, zonder zich zorgen te maken over hoe de verschillende criteria geïnterpreteerd moeten worden. Bovendien hebben mensen van nature de neiging om te vergelijken, waardoor deze methode het beoordelen gemakkelijker maakt (Laming, 2003). Daarnaast verhoogt de methode de betrouwbaarheid van oordelen binnen en tussen beoordelaars (Thurstone, 1927; Pollitt, 2012). Een beoordelaar blijkt consistentere beslissingen te nemen over tijd (d.w.z. hetzelfde product in een vergelijking laten winnen, ongeacht het moment van de dag of welke producten je ervoor gezien hebt). Daarnaast zijn oordelen ook consistenter tussen beoordelaars. De kans is immers groter dat meerdere beoordelaars hetzelfde product in een vergelijking als beste aanwijzen dan dat ze allen dezelfde scores toekennen.

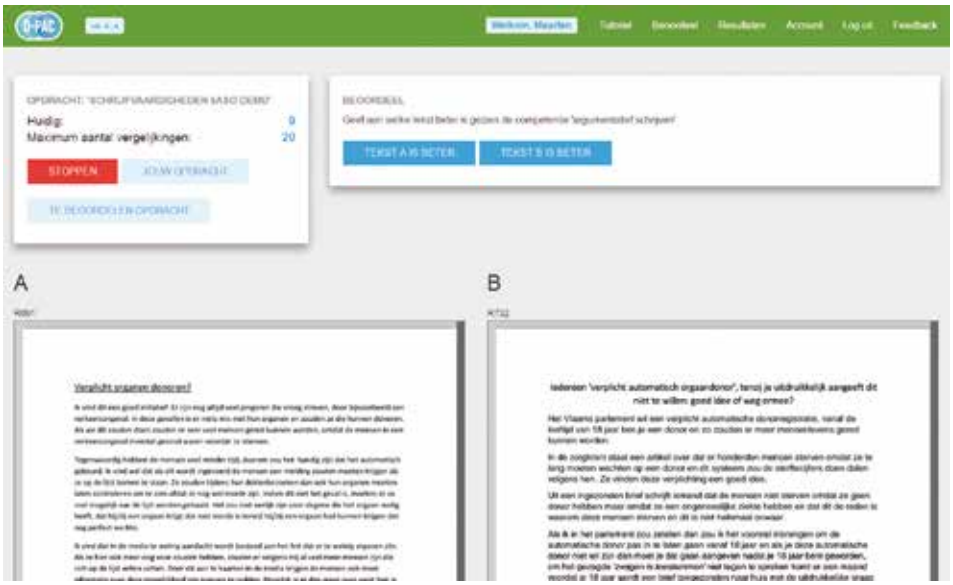
Paarsgewijs vergelijken met D-PAC

Het is echter niet praktisch haalbaar voor docenten om voor elke toets zelf paren willekeurig samen te stellen. Daarom hebben onderzoekers van de Universiteit Antwerpen, Universiteit Gent en iMec recent een online beoordelingsinstrument ontwikkeld dat gebruik maakt van het principe van paarsgewijs vergelijken. Dit instrument heet D-PAC, wat staat voor een 'Digital Platform for the Assessment of Competences'. D-PAC maakt het mogelijk om in een online omgeving snel en betrouwbaar producten paarsgewijs te vergelijken. Elke beoordelaar krijgt een reeks paren te zien die willekeurig zijn samengesteld en kiest steeds het beste product van de twee. Voor nog meer zekerheid over welke product nu de betere is, maakt D-PAC gebruik van meerdere beoordelaars. Alle beoordelingen samen resulteren in een kwaliteitsschaal die alle producten in een rangorde plaatst en daarbij de verschillen inzichtelijk maakt. Zie Figuur 1a t/m d voor een overzicht van de verschillende frames in D-PAC.

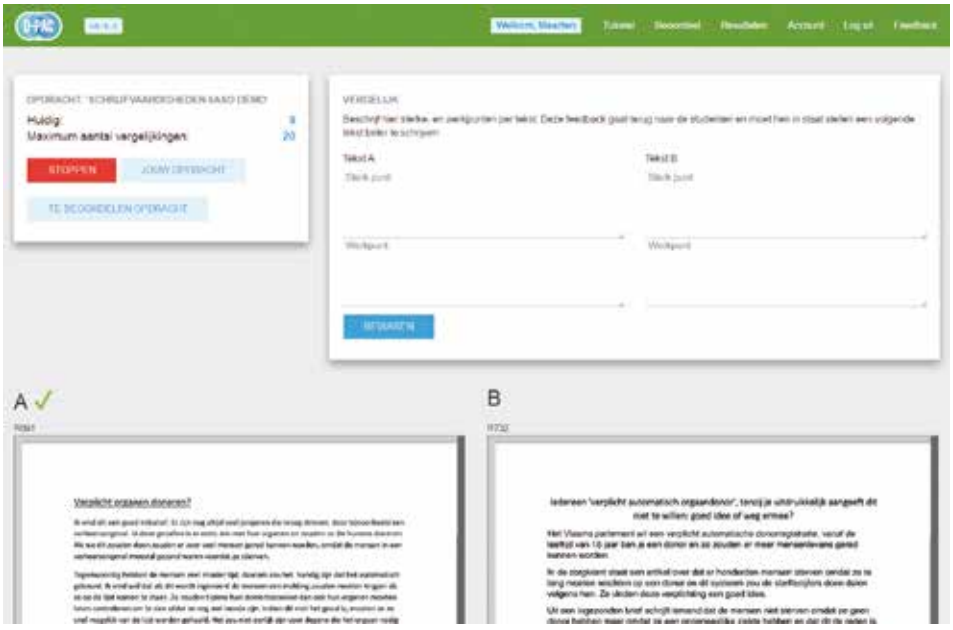
D-PAC genereert niet automatisch scores voor de producten. Bepalen wat de cesuur is voor een reeks producten wordt desgewenst achteraf door de beoordelaars bepaald. Hiervoor zijn verschillende mogelijkheden. Er kan bijvoorbeeld een cesuuroverleg plaatsvinden waarin een slaaggrens wordt gelegd bij het gemiddelde van de rangorde. Een andere optie is dat de producten met de laagste kwaliteit en de hoogste kwaliteit besproken worden om een score toe te kennen. Vervolgens kunnen de scores van de tussenliggende producten toegekend worden via de waarden in de rangorde. Een laatste optie is dat reeds gescoorde producten mee beoordeeld worden, zodat in de rangorde snel duidelijk wordt welke producten zich rond bepaalde scores situeren.

Hoe waardevol is D-PAC voor de praktijk?

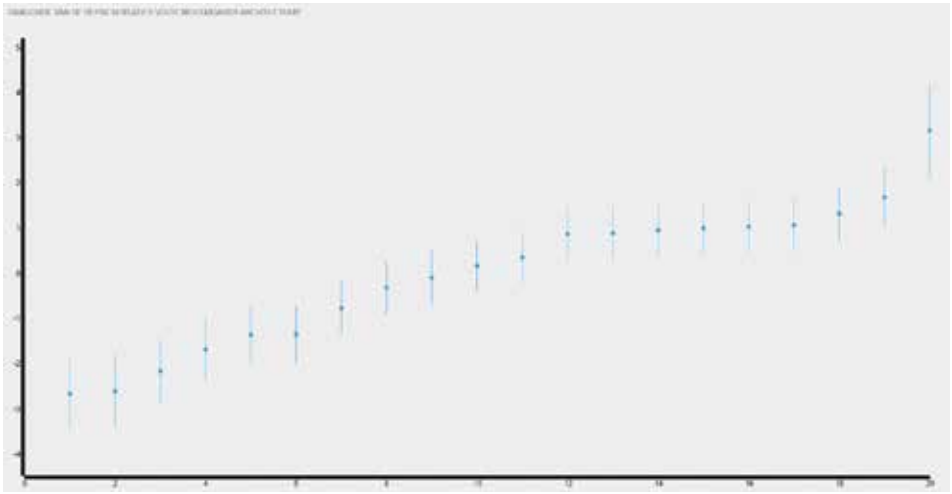
Voor de meest betrouwbare en valide resultaten met D-PAC is het raadzaam om (1) meerdere beoordelaars te gebruiken, en (2) elk product meerdere keren te vergelijken met willekeurige anderen. Als aan deze twee voorwaarden wordt voldaan kan deze beoordelingsmethode meer tijd en mankracht kosten dan traditionele methoden waarin een product vaak maar door één beoordelaar wordt beoordeeld. Is dat het wel waard?



Figuur 1a. Een voorbeeld van het basisscherm waarbij beoordelaars een keuze moeten maken voor de beste tekst in licht van de te beoordelen competentie (in dit geval argumentatief schrijven).



Figuur 1b. Een voorbeeld van hoe de feedbackoptie bij het beoordelen eruit kan zien.



Figuur 1c. Een voorbeeld van een rangorde op basis van alle vergelijkingen, inclusief algemene informatie over beoordelingen, zoals de betrouwbaarheid en het aantal vergelijkingen.

Figuur 1d. Een voorbeeld van hoe de feedback per product eruit kan zien.

Om te onderzoeken hoe waardevol D-PAC is voor de praktijk zijn er sinds de ontwikkeling al 27 try-outs geweest waarbij verschillende soorten competenties zijn beoordeeld, van schrijftaken en beeldende vaardigheden tot aan zelfreflecties, zie Figuur 2. De meeste try-outs waren met docenten en/of leerlingen uit het basisonderwijs, voortgezet onderwijs, en hoger onderwijs. Daarnaast waren er try-outs in de HR context, waarbij D-PAC werd gebruikt voor selectie en trainingsdoeleinden. Er zijn hierbij verschillende typen producten getest, zoals teksten (van een halve pagina tot wel 80 pagina's), plaatjes, maar ook audio en video. Gemiddeld bestond elke try-out uit 65 producten (min = 6, max = 201) en 24 beoordelaars (min = 4, max = 93).

Efficiëntie en betrouwbaarheid

In één van de try-outs hebben we de betrouwbaarheidswinst afgezet tegen de tijd dat het de beoordelaars kostte om tot een oordeel te komen. Hiermee krijgen we enig inzicht in de efficiëntie van D-PAC ten opzichte van criterialijsten. Voor de groep beoordelaars die met D-PAC werkten werd de interbeoordelaarsbetrouwbaarheid berekend aan de hand van de Separate Scale Reliability. Voor de groep beoordelaars die met criterialijsten werd de Intra Class Correlation berekend. Ondanks dat dit twee verschillende maten zijn voor de consistentie tussen beoordelaars, zijn ze wel onderling te vergelijken.

De uitkomsten laten zien dat voor de analytische oordelen meer dan 50% werd bepaald door zaken die niet gerelateerd zijn aan de kwaliteit van de producten (bv, de persoon die beoordeelt, de gebruikte criterialijst,…), ongeacht de tijd die in de beoordeling werd geïnvesteerd. Met andere woorden, meer tijd leverde voor deze groep beoordelaars geen betrouwbaarheidswinst op. Met D-PAC bleek dat het beoordelen wel sneller leidde tot een hogere betrouwbaarheid. Bij 17 vergelijkingen werd 80% van de verschillen bepaald door kwaliteit van het product, ongeacht de beoordelaar of de producten van anderen waarmee vergeleken werd. Ook bleek dat beoordelaars relatief snel vergelijkingen kunnen maken (Coertjens, Lesterhuis, Verhavert, & De Maeyer, 2016).

Validiteit van de oordelen

Doordat D-PAC gebruik maakt van het oordeel van meerdere beoordelaars, is de beoordeling niet meer afhankelijk van een individuele beoordelaar. Dit leidt uiteindelijk tot meer valide en dus kwaliteitsvolle oordelen (Van Daal, Lesterhuis, Coertjens, Donche, & De Maeyer, 2016). Bovendien genereert D-PAC feedback over de kwaliteit van de beoordelaars en hun oordelen. Deze informatie kan gebruikt worden in scholen om de discussie rond de gangbare beoordelingswijze aan te gaan om zo tot eenduidige evaluaties te komen.

Onderwijs	Verschillende schrijftaken Wiskundig probleemoplossend vermogen Beeldende vaardigheden Interpretatie statistiek Moodboards Zelfreflecties ER-schema's Evidence based practice
Selectie	CV's Portfolio's directies projectaanvragen
Training	Interviewvaardigheden Professionalisering examinatoren

Figuur 2. Overzicht van try-outs met D-PAC.

Feedback met D-PAC

D-PAC blijkt dus een efficiënt, betrouwbaar en valide instrument voor het beoordelen van competenties. In het onderwijs willen we echter meer dan alleen maar oordelen geven. Een bijkomend voordeel van D-PAC is dat het inzetten van meerdere beoordelaars uitgebreide en diepgaande feedback naar studenten mogelijk maakt. Dit vraagt natuurlijk wel wat van een school, want het is lang niet altijd praktisch haalbaar om meerdere docenten bij de beoordeling te betrekken. De feedbackfunctie leent zich echter uitermate goed voor peer assessments, waarbij leerlingen elkaars producten beoordelen en van feedback voorzien.

Meer weten?

Het "Digital Platform for the Assessment of Competences" (D-PAC) verricht onderzoek naar de kwaliteit van verschillende beoordelingsvormen, zowel in onderwijskundige als Human Resources contexten. D-PAC is een onderzoekssamenwerking tussen Universiteit Antwerpen, Universiteit Gent en iMec en wordt gesteund door het Agentschap voor Innovatie door Wetenschap en Technologie (IWT). Meer informatie kan u vinden op www.d-pac.be. ■

Referenties

- Coertjens, L., Lesterhuis, M., Verhavert, S., & Maeyer, S. De. (2016). Een afweging tussen de betrouwbaarheid en efficiëntie van twee beoordelingsmethoden: Paarsgewijze vergelijking en analytisch scoren vergeleken. Manuscript is aangeboden voor publicatie.
- Glopper, K. De. (1988). Schrijven beschreven. Inhoud, opbrengsten en achtergronden van het schrijfonderwijs in de eerste vier leerjaren van het voortgezet onderwijs [Proefschrift]. Den Haag: SVO.
- Erkens, S. (2015). Leerlingenevaluatie over scholen heen. Hoe verschillend beoordelen leraren uit verschillende scholen eenzelfde leerlingprestaties? [masterproef]. Ongepubliceerd manuscript, Universiteit Antwerpen, Instituut voor Onderwijs- en Informatiewetenschappen.
- Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. Educational Research Review, 2, 130-144.
- Laming, D. (2003). Human Judgment: The eye of the beholder. Andover, UK: Cengage Learning EMEA.
- Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? Language Testing, 19, 246-276.
- Meuffels, B. (1994). De verguisde beoordelaar: Opstellen over opstelbeoordeling. Amsterdam: Thesis Publishers.
- Pollitt, A. (2012) The method of adaptive comparative judgment. Assessments in Education: Principles, Policy & Practice, 19 (3), 281-300
- Thurstone, L. L. (1927). Psychophysical analysis. The American journal of psychology, 38(3), 368-389.
- Van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., & Maeyer, S. De. (in press). Validity of comparative judgement to assess academic writing: Examining implications of its holistic character and building rank-orders on a shared consensus. Assessment in Education: Principles, Policy & Practice.
- Bergh, H. Van den, Maeyer, S. De, Weijen, D. Van, & Tillema, M. (2012). Generalizability of text quality scores. In E. Van Steendam, M. Tillema, G. Rijlaarsdam, & H. Van den Bergh (Eds.)., Measuring writing: Recent insights into theory, methodology and practice (Vo. 27, pp. 23-32). Leiden: Brill.