Comproved: Waarom moeilijk beoordelen als het ook eenvoudig kan?

ECHO-tip juli 2021

ExpertiseCentrum Hoger Onderwijs *(Universiteit Antwerpen)*

I.s.m. Maarten Goossens (Opleidings- en Onderwijswetenschappen en co-founder Comproved, UAntwerpen), Karla Groen (E-campus, UAntwerpen) & Philip Lambrechts (Departement Onderwijs, UAntwerpen)

Doorgaans gebruiken lesgevers **criterialijsten** of **rubrieken** om de prestaties van studenten te beoordelen. Zij helpen hen op weg om op een **objectieve manier** een uitspraak te kunnen doen over de mate waarin studenten bepaalde competenties al dan niet hebben bereikt. Toch is beoordelen aan de hand van criteria of rubrieken niet altijd vanzelfsprekend. Misschien herken je wel (één van) volgende situaties? (1) Een collega-docent stelde een uitgekiende rubriek op voor de beoordeling van de schrijfopdrachten die jullie studenten uitwerkten. Fijn, zo hebben jullie beiden de nodige houvast voor het toekennen van scores. Hoewel, je vraagt je nu tijdens de correctie wel regelmatig af wat er precies wordt bedoeld met bepaalde competentiebeschrijvingen. (2) Je leest een opdracht van een student en volgens je ervaring schat je de waarde van het werk in op 12/20. Maak je echter de optelsom van de vooropgestelde criteria, dan kom je uit op 15/20. Een opvallend verschil! Hou je vast aan de objectieve criteria en geef je de student een 15, of pas je hier en daar wat aan zodat de optelsom toch 12 wordt? (3) Je geeft studenten een omvangrijke opdracht via dewelke je onder andere ook hun creativiteit wil stimuleren en beoordelen. Hoe definieer je echter een complexe competentie als creativiteit zodat iedereen er hetzelfde onder verstaat en anderzijds bepaalde nuances en/of het totaalbeeld toch niet verloren gaan? Kunnen alle aspecten van creativiteit wel gevat worden in een aantal vooropgestelde criteria (zie ook de ECHO-onderwijstip over het stimuleren en beoordelen van creativiteit)?

Het is niet zo vreemd als je jezelf herkent in (één van) bovenstaande situaties. Ook wetenschappelijk onderzoek toont immers aan dat er wel wat **valkuilen** zijn bij het **ontwikkelen en gebruiken van criteria** en **rubrieken** (Bloxham, 2009; Sadler, 2009). De **paarsgewijze vergelijkingsmethode** biedt kansen om de genoemde moeilijkheden weg te werken. In wat volgt verduidelijken we eerst het begrip paarsgewijze vergelijking. Vervolgens gaan we in op enkele concrete toepassingen aan de hand van de tool Comproved. We sluiten af met een overzicht van enkele belangrijke (aandachts)punten over paarsgewijs vergelijken en Comproved.

Paarsgewijze vergelijking

Laming (2014) stelt dat elk oordeel het gevolg is van een vergelijking. Als je bijvoorbeeld het werk van een student moet beoordelen, vergelijk je het werk ofwel met het werk van andere studenten, ofwel met je interne standaard. Ook als je gebruik maakt van objectieve standaarden, zoals criteria, ben je aan het vergelijken. Je vergelijkt ofwel met die standaard, ofwel met andere taken die je tegenover de standaard spiegelde. Met andere woorden, **vergelijken is impliciet aan beoordelen**. Paarsgewijs vergelijken maakt die impliciete vergelijking expliciet. De methode vindt zijn oorsprong in de wet van Thurstone (1927). Die stelt dat mensen beter en betrouwbaarder zijn in het vergelijken van twee objecten en in het aanwijzen welke meer/beter is, dan in het toekennen van absolute scores aan een enkel object. We illustreren dit met een voorbeeld:

Je krijgt een vreemd voorwerp in je handen en we vragen je om het gewicht te schatten. Een moeilijke taak, want je voelt dat je een stap in het ongewisse moet zetten en bovendien is de kans zeer klein dat je het exacte gewicht kan bepalen. Geven we je echter twee voorwerpen en vragen we je welk van de twee het zwaarst is, dan ervaar je dat als een zeer eenvoudige taak. Waarvan je het gevoel hebt dat je die tot een goed einde zal brengen, en de kans bovendien groot zal zijn dat je het bij het rechte eind hebt.

Bij de methode van paarsgewijze vergelijking maken **meerdere beoordelaars telkens keuzes uit paren van objecten**. Die paren worden **ad random samengesteld** uit alle te beoordelen objecten. In een onderwijssetting betekent dit dat meerdere lesgevers paren van taken van verschillende studenten met elkaar vergelijken en aangeven **welke van de twee nu de beste is**, in functie van een bepaalde competentie. Dit is duidelijk en eenvoudig. Aan de hand van de gemaakte keuzes kan er vervolgens een **betrouwbare rangorde** worden opgesteld, van de minst goede tot de beste taak.

De methode doet bewust een beroep op je expertise als beoordelaar. Hierbij word je vooraf niet gedwongen door een bepaalde bril te kijken (van bijvoorbeeld een criterialijst of rubriek). Doordat verschillende beoordelaars, in verschillende paren, meerdere keren eenzelfde taak bekijken, worden nagenoeg automatisch **alle aspecten van de te beoordelen competentie meegenomen** in de uiteindelijke scoring. Op die manier wordt de **validiteit** van de beoordelingsmethode gewaarborgd (zie ook de ECHO-onderwijstip 'meet wat u moet weten' van 2013).

Toepassingen binnen onderwijs

In het onderwijs kan paarsgewijs vergelijken gebruikt worden voor het beoordelen van producten van uiteenlopende aard en uit **verschillende contexten**, waaronder wiskundige redeneeropdrachten, schrijf- en ontwerpproducten, presentaties, etc. Een voordeel van paarsgewijze vergelijking is de **ruimte tot openheid van de taken**. Je bent bij de ontwikkeling van de opdracht niet gebonden door de vraag: "(Hoe) gaan we dit kunnen scoren?". Zo hoef je bijvoorbeeld niet te oordelen of een redenering juist of fout is, maar kan je aangeven welke van de twee redeneringen de sterkste is. Ook al leidt die misschien niet tot de juiste uitkomst, omdat er ergens een rekenfoutje werd gemaakt.

Paarsgewijs vergelijken kan je ook als **werkvorm** toepassen binnen je lespraktijk. Zo kan je twee voorbeeldteksten presenteren en je studenten aanzetten om op zoek te gaan naar waarom de ene beter is dan de andere. Doordat studenten mogen vergelijken, vallen hen de verschillende karakteristieken per tekst veelal goed op. Het is mogelijk dat de ene tekst een veel duidelijkere structuur heeft dan de andere. En in het volgende paar valt hen bijvoorbeeld op dat de argumentatie in tekst twee meer steek houdt dan die van tekst een. Op die manier **bouwen studenten zelf kwaliteitscriteria op**, op een eenvoudige manier.

Wil je paarsgewijs vergelijken inzetten als **beoordelingsmethode**, dan is een tooltechnische ondersteuning aangewezen. Die ondersteuning kan **Comproved** bieden. Het is een tool die is ontworpen door onderzoekers van UAntwerpen, UGent en Imec.

Concrete toepassingen a.d.h.v. Comproved

Omdat paarsgewijze vergelijking een relatief eenvoudige taak is, die zonder veel training en inleiding kan worden ingezet, leent het zich uitermate voor **peerfeedback/-assessment** (zie ook de ECHO-tip 'naar een betrouwbaar peer assessment' van 2017). Studenten hoeven immers **geen absolute uitspraken** te doen over het werk van medestudenten en alles verloopt **anoniem**. Hierdoor vermindert veelal de eventuele **weerstand** tegen peerfeedback/-assessment en het **gevoel van onveiligheid** dat er soms bij heerst .

Hieronder geven we twee concrete beschrijvingen van peerfeedback/-assessment, waarin gebruik werd gemaakt van **paarsgewijze vergelijking aan de hand van Comproved**.

(1) Peerfeedback van ‘mood boards’ in de opleiding Interieurarchitectuur – UAntwerpen

*Een **mood board** is een beeld waarin een bepaalde emotie wordt uitgedrukt, vaak gebruikt om op een visuele manier te communiceren met de klant over sferen, gevoelens, organisatiewaarden, etc.*

*Studenten Interieurarchitectuur van UAntwerpen krijgen jaarlijks de opdracht om in groep mood boards te ontwikkelen. Die werken worden **klassikaal besproken** en hebben als doel de betrokken **studenten te leren inschatten wanneer ze een werk kwalitatief mogen vinden**. Dit is doorgaans een **moeizaam en tijdrovend proces**.*

*Om dit proces **efficiënter en vlotter** te laten verlopen werd een **peerfeedback in Comproved** opgezet. Op die manier konden de studenten alle mood boards **thuis beoordelen en becommentariëren**.*

*In concreto werden de mood boards van de studenten opgeladen in Comproved en bracht het algoritme van de tool de opgeladen producten **ad random samen** in duo’s. Vervolgens moest elke student het beste van twee mood boards aanduiden en **sterke en werkpunten** per board aangeven. Ze kregen hiervoor een week de tijd. Als ze opnieuw aanmeldden gingen ze verder waar ze waren gebleven. Totdat het vooropgestelde aantal vergelijkingen was gemaakt.*

*Aan de hand van alle keuzes van de studenten, genereerde Comproved een **rangorde** van de opgeladen werken; van het minst kwalitatieve naar het meest kwalitatieve, en dit volgens de **inschatting van de studenten**.*

*De verkregen rangorde was voor de lesgever de basis om de mood boards in groep te bespreken. Hiervoor werden de **resultaten vanuit Comproved geprojecteerd en besproken in een feedbackcollege**. Zowel de rangorde zelf (de kwalitatieve inschatting van de studenten tegenover die van de lesgever) als de gegeven feedback (op welke aspecten letten de studenten en hoe beargumenteren ze die) werden hiervoor gebruikt.*

*Nadien werd de **feedback aan de respectievelijke groepen** bezorgd, zodat zij hiermee hun mood board konden **bijsturen**.*

(2) Reductie workload via peerassessment (summatief) – UHasselt

*In een zoektocht naar het **reduceren van verbeterlast** besloot een lesgever van UHasselt om Comproved in te zetten voor een **peerassessment**. De **honderd** betrokken **studenten** moesten een **paper** schrijven en die uploaden in Comproved. Vervolgens werden automatisch hieruit **random paren** geselecteerd, die de studenten moesten beoordelen. Ze moesten ook elke paper voorzien van feedback. Dit resulteerde enerzijds in een **rangorde** van de werken en anderzijds in **feedback** van een tiental medestudenten per paper.*

*Voordien beoordeelde de lesgever de papers met een **‘pass/fail’**-systeem. Op het moment van de formatieve beoordeling ging de lesgever na of de paper al voldeed aan de eindcompetentie. Zo ja, dan kreeg de paper een ‘pass’, indien niet kreeg de paper een ‘fail’. De lesgever gaf ook feedback op de paper.*

*In eerste instantie liep deze ‘pass/fail’ beoordelingsmethode **nog parallel met het peerassessment in Comproved**. Van de honderd papers kregen er veertien een ‘fail’. Die ‘fails’ vielen allen in de twintig laagst genoteerde papers op de rangorde die resulteerde uit de vergelijkingen van de studenten in Comproved. Met andere woorden die papers werden **door de studenten ook aanzien als minder kwalitatief**. Bovendien bleek de **feedback**, die de studenten op de papers hadden gegeven, **zeer kwalitatief** en vergelijkbaar met de feedback van de lesgever.*

*Op basis van deze resultaten besliste de lesgever om voortaan de papers enkel nog via peerassessment te beoordelen en te voorzien van feedback. Ter **controle** keek de lesgever alleen de **onderste veertig procent** van de rangorde na, om te zien of er **geen onterechte ‘fails’** tussen zaten en om de cesuur te bepalen. Zo leverde deze methode een **aanzienlijke tijdswinst** op voor de lesgever.*

Enkele punten op een rijtje

Hoe werkt comparatief beoordelen?

- Als **assessor** (lesgever, externe beoordelaar, student) maak je **meerdere vergelijkingen van te beoordelen opdrachten**.
- **Meerdere beoordelaars** maken vergelijkingen, waardoor elk werkstuk meerdere keren wordt vergeleken.
- Op basis hiervan wordt een **rangorde** van de beoordeelde opdrachten berekend.
- Het achterliggend **algoritme** van Comproved selecteert telkens een **nieuw vergelijkingspaar** en zorgt op die manier ervoor dat elk 'werkstuk' evenveel wordt vergeleken. Volgens de configuratie krijg je bij bv. twintig werkstukken en vijf beoordelaars een vergelijkingsgraad van vijftien keer per werkstuk en een totaal van dertig vergelijkingen per beoordelaar.
- Omdat de vergelijkingen holistisch gebeuren en elke vergelijking doorgaans **maximaal drie minuten** in beslag neemt, bekom je een totale tijdsbelasting van ca. negentig minuten per beoordelaar, voor een vergelijking van twintig werkstukken.

Wat is de kracht van comparatief beoordelen?

- Vergelijken is **makkelijk en snel**.
- Je maakt gebruik van de expertise van **meerdere beoordelaars**.
- Elk te beoordelen werk komt terug in **meerdere vergelijkingen**.
- Wetenschappelijk onderzoek toont aan dat je **betrouwbare en valide beoordelingen** verkrijgt.
- Een comparatieve beoordeling via Comproved kan worden georganiseerd binnen de **instelling of opleiding**, maar ook instellings- en opleidingsoverschrijdend.

Wanneer comparatief beoordelen?

- **(peer)assessment** (formatief en/of summatief), bv. moodboards, academisch schrijven, stage, (e-)portfolio, zelfreflectie, argumentatieve tekst, wiskundig probleemoplossend vermogen, interactieve installatie (live assessment), etc.
- **Selectie**, bv. cv-screening, projectvoorstel, etc. Hier gaat het over het verkiezen van een 'winnaar'.
- **Professionalisering**, bv. beoordelaarstraining examencommissie (schrijfvaardigheden). De resultaten uit Comproved kunnen gebruikt worden om de beoordelaars te professionaliseren en hun beoordelingspraktijk meer met elkaar in lijn te brengen.

Wat zijn de voordelen van Comproved bij peerassessment en -feedback?

- De student beoordeelt met Comproved comparatief het werk van anderen, waardoor hij/zij **gerichter feedback** leert geven.
- De student leert door taken te zien van **uiteenlopende kwaliteit** van zijn/haar medestudenten. De vele voorbeelden verschaffen volop inspiratie. Bovendien stimuleert het zien van het werk van anderen **reflectie over het eigen werk**. Het moeten geven van specifieke feedback aan peers helpt bij het **zich eigen maken van kwaliteitscriteria**.

- Zowel het **leren uit voorbeelden**, als het **vergelijken** en zoeken naar overeenkomsten en tegenstellingen, zijn bewezen leerprincipes (e.g., Carless & Cham, 2016; Pachur & Olsson, 2012).
- Zelfs al zijn medestudenten (nog) geen expert, ze kunnen vaak prima inschatten of een ander de opdracht goed of minder goed heeft uitgevoerd. Werken onderling vergelijken en de beste aanduiden is eenvoudig, er is dus **geen uitvoerige instructie** of training vooraf vereist.
- De **feedback** die studenten van hun peers ontvangen vinden ze **rijk en waardevol**. Ze aanvaarden de resultaten, doordat ze vertrouwen hebben in elkaars bekwaamheid. Ze zijn bereid om er **ook effectief iets mee te doen**, net doordat ze ook zelf feedback gaven (en verwachten dat dit peers aanzet tot reflectie en actie).
- Het **achteraf bespreken van de opdracht** en de gegeven **feedback is waardevol**, want hierdoor worden de criteria voor zowel lesgevers als studenten worden helder. Het mooie hierbij is dat **studenten mee mogen denken** over de **beoordelingscriteria**. Dit zorgt ervoor dat ze de criteria beter zullen begrijpen en eerder zullen accepteren.
- Comproved biedt **lesgevers veel extra informatie**, want ze kunnen monitoren in welke mate studenten al dan niet gelijkaardige opvattingen hebben over wat kwalitatief is. De lesgever kan ook opvolgen hoe lang studenten over de evaluatie hebben gedaan, om bv. zo de betrouwbaarheid van het assessment te analyseren.

Waarmee moet je rekening houden bij het gebruik van Comproved?

- Het is minder geschikt voor de beoordeling van **omvangrijke werkstukken** (bv. lange teksten van tientallen pagina's), want dan is de werklast van het vergelijken simpelweg te groot.
- Je kan ermee aan de slag vanaf twee beoordelaars, maar de voorkeur is **minimaal vier**.
- Je moet de **configuratie vooraf instellen** (bv. type werkstuk, vergelijkings-vragen (bv. Welke van de twee is beter?), aantal vergelijkingen, het al dan niet integreren van feedback…).
- De tool integreer je best in eerste instantie in een **formatieve evaluatie**, om er eventueel later ook summatief mee aan de slag te gaan.

Wat leren we uit de ervaringen van Comproved-gebruikers?

- Studenten ervaren het platform als een **'veilige' omgeving**: het beoordelen verloopt er geheel anoniem. Niemand weet welk werk van wie is.

Meer weten?

Criteria en rubrieken

ExpertiseCentrum Hoger Onderwijs (2013). *Vijftig onderwijstips*. Antwerpen-Apeldoorn: Garant. (Voor personeelsleden UAntwerpen hier online raadpleegbaar) > onderwijstip 'rubrieken als begeleidings-en beoordelingsinstrument'

Bloxham, S. (2009). Marking and moderation in the UK: false assumptions and wasted resources. *Assessment & Evaluation in Higher Education, 34*(2), 209-220. https://doi.org/10.1080/02602930801955978

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education, 34*(2), 159–179. https://doi.org/10.1080/02602930801956059

Paarsgewijze vergelijking

Coertjens, L. Lesterhuis, M., Verhavert, S., Gasse, R., & De Maeyer, S. (2018). *Teksten beoordelen met criterialijsten of via paarsgewijze vergelijking: een afweging van betrouwbaarheid en tijdsinvestering*. Geraadpleegd op 21 mei 2021, van https://www.kennisdelingtaalbeleid.org/wp-content/uploads/2018/01/Teksten-beoordelen-met-criterialijsten-of-via-paarsgewijze-vergelijking.-Een-afweging-van-betrouwbaarheid-en-tijdsinvestering.pdf

Mortier, A. V., Lesterhuis, M., Vlerick, P., & Maeyer, S. D. (2015). Comparative judgment within online assessment: exploring students feedback reactions. In E. Ras & D. J. Brinke (Eds.), Computer Assisted Assessment. Research into E-Assessment (pp. 69–79). Springer International Publishing.

Pollitt, A. (2012a). Comparative judgement for assessment. *International Journal of Technology and Design Education, 22*(2), 157–170. https://link.springer.com/content/pdf/10.1007/s10798-011-9189-x.pdf

Pollitt, A. (2012b).The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice, 19*(3), 281–300. https://doi.org/10.1080/0969594X.2012.665354

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*(4), 273-286. https://doi.org/10.1037/h0070288

Comproved

Comproved. (z.d.). *FAQ*. Comproved: beoordeel beter en makkelijker. https://comproved.com/faq

i-Learn Vlaanderen. (2021, 18 januari). *Onder de motorkap: Comproved*. i-Learn Vlaanderen. https://www.i-learn.vlaanderen/nieuws/edtech/onder-de-motorkap-comproved

Deneire, A. (2017, 10 september). *Zo laat je je studenten lessen trekken uit krachtige D-PAC-feedback*. Edubron blogt. http://www.edubronblogt.be/onderzoek/zo-laat-studenten-lessen-trekken-krachtige-d-pac-feedback

Willems, T. (2020, 30 november). *Laat je studenten elkaars werk rangschikken met Comproved*. Avans Hogeschool. https://tools.avans.nl/articles/laat-je-studenten-elkaars-werk-rangschikken-met-comproved

ExpertiseCentrum Hoger Onderwijs (ECHO)

Venusstraat 35

B - 2000 Antwerpen

echo@uantwerpen.be

www.uantwerpen.be/echo

Gebruikerservaringen met Comproved, via (sociale) media en/of contact met Comproved

Twitter. (z.d.). *Comproved*. https://twitter.com/comproved

Facebook. (z.d.). *Comparative judgement to the rescue*. https://www.facebook.com/groups/2393764744190491

De Wilde, B. (2019, 7 september). *Hoe evalueer je competenties?* Klasse. https://www.klasse.be/192665/hoe-evalueer-je-competenties/

ExpertiseCentrum Hoger Onderwijs: onderwijstips

Creativiteit stimuleren en beoordelen (november 2015)

Meet wat u moet meten (november 2013)

Peerassessment: studenten beoordelen elkaar (september 2013)

Naar betrouwbaar peerassessment (juni 2017)

Rubrieken als begeleidings- en beoordelingsinstrument (september 2017)

Algemeen

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2011). *Building a Validity Argument for the Test of English as a Foreign Language.* New York: Routledge.

Laming, D. (2003). *Human judgment: The eye of the beholder.* Andover: Cengage Learning EMEA.

McNamara, T. F. (1996). *Measuring second language performance.* Addison Wesley Longman.

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

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

UAntwerpen-specifieke inspiratiebronnen

UAntwerpen heeft een universiteitsbrede licentie voor Comproved. In BlackBoard vind je concrete (technische) informatie om met Comproved aan de slag te gaan. Er is ook een gids voor studenten beschikbaar.

Infocenter Onderwijs biedt je enkele good practices van peerassessment (ook d.m.v. paarsgewijze vergelijking).

Op de Pintra-pagina 'Onderwijs op de campus en online' vind je heel wat info over peer-assessment en -feedback.