Competenties kwaliteitsvol beoordelen : brengt een comparatieve aanpak soelaas?

Reference:

Lesterhuis Marije, Donche Vincent, de Maeyer Sven, Van Daal Tine, Van Gasse Roos, Coertjens Liesje, Verhavert Sam, Mortier Anneleen, Coenen Tanguy, Vlerick Peter.- Competenties kwaliteitsvol beoordelen : brengt een comparatieve aanpak soelaas?

Tijdschrift voor hoger onderwijs - ISSN 0168-1095 - 33:2(2015), p. 55-67

To cite this reference: http://hdl.handle.net/10067/1283920151162165141

Titel:
Competenties kwaliteitsvol beoordelen: brengt een comparatieve aanpak soelaas?

Auteurs:
Marije Lesterhuis
M. Lesterhuis (marije.lesterhuis@uantwerpen.be) is werkzaam bij het Instituut voor Onderwijs- en Informatiewetenschappen, Universiteit Antwerpen

Vincent Donche
Prof. dr. V. Donche (vincent.donche@uantwerpen.be) is werkzaam bij het Instituut voor Onderwijs- en Informatiewetenschappen, Universiteit Antwerpen

Sven De Maeyer
Prof. dr. S. De Maeyer, is werkzaam bij het Instituut voor Onderwijs- en Informatiewetenschappen, Universiteit Antwerpen

Tine van Daal
T. van Daal is werkzaam bij het Instituut voor Onderwijs- en Informatiewetenschappen, Universiteit Antwerpen

Roos Van Gasse
R. Van Gasse is werkzaam bij het Instituut voor Onderwijs- en Informatiewetenschappen, Universiteit Antwerpen

Liesje Coertjens
dr. L. Coertjens is werkzaam bij het Instituut voor Onderwijs- en Informatiewetenschappen, Universiteit Antwerpen

San Verhavert
S. Verhavert is werkzaam bij het Instituut voor Onderwijs- en Informatiewetenschappen, Universiteit Antwerpen

Anneleen Mortier
A. Mortier is werkzaam bij het de Faculteit Psychologie en Pedagogische Wetenschappen, Universiteit Gent

Tanguy Coenen
dr. T. Coenen is werkzaam bij, iMinds, iLab.o, Gent

Peter Vlerick
Prof. dr. P. Vlerick is werkzaam bij het de Faculteit Psychologie en Pedagogische Wetenschappen, Universiteit Gent

Jan Vanhoof
Prof. dr. J. Vanhoof is werkzaam bij het Instituut voor Onderwijs- en Informatiewetenschappen, Universiteit Antwerpen

Peter Van Petegem
Prof. dr. P. Van Petegem is werkzaam bij het Instituut voor Onderwijs- en Informatiewetenschappen, Universiteit Antwerpen

Titel:
Competenties kwaliteitsvol beoordelen: brengt een comparatieve aanpak soelaas?

Trefwoorden (5 engelse, 5 nederlandse)
Comparative judgement, performance assessment, scoring, reliability, validity
Comparatieve beoordelingen, performance assessment, scoren, betrouwbaarheid, validiteit

Abstracts
Samenvatting (max 200 woorden):
In het hoger onderwijs wordt steeds vaker met performance assessments gewerkt om competenties van studenten te evalueren. Docenten worstelen echter met hoe ze deze performance assessments het beste kunnen scoren. Meestal wordt hier een combinatie van criterialijsten en holistisch scoren voor gebruikt. Deze methode leidt echter niet altijd tot betrouwbare resultaten. Ook treden er problemen op met de validiteit, doordat in deze methode de competentie niet als geheel benaderd wordt. In dit artikel wordt ingegaan op deze problematiek en wordt een alternatieve benadering voorgesteld, de comparatieve beoordelingsmethode (CB). In CB wordt aan beoordelaars gevraagd prestaties van studenten te vergelijken en aan te geven welke het best presteert in termen van de te beoordelen competentie. Door meerdere vergelijkingen op te lossen kan een rangorde gegenereerd worden van beste naar minst goede prestatie. Betrouwbaarheid wordt nagestreefd door meerdere beoordelaars prestaties meerdere keren te laten beoordelen. Daarnaast stelt de methode meer valide te zijn, omdat de beoordelaars hun keuze baseren op basis van een holistische evaluatie van de prestaties en de taak voor studenten kan meer open geformuleerd worden. Alhoewel onderzoek laat zien dat het een veelbelovende alternatieve methode is, is meer onderzoek noodzakelijk naar de betrouwbaarheid, validiteit en praktische haalbaarheid.

Engelstalige titel:
Credible assessment of competences: is comparative judgement the way forward?

Engelse samenvatting (max 200 woorden):
Performance assessments are more and more used in higher education, to evaluate the competences of students. However, teachers struggle with the scoring of the assessments. Most teachers use combinations of analytical assessment forms, such as rubrics, and holistic scores. However, this method does not always result in reliable and valid scores. In this article an alternative method is introduced, comparative judgement (CJ). CJ asks judges to compare students' work and to decide which performance is best in terms of the competence under assessment. By making several comparisons, a rank order can be generated from best to worst performance. Reliability is pursued by asking more judges to make more comparisons. Thereby, the method should be more valid, because judges make the decisions based on a holistic evaluation of the performances and student assignments can be developed more open-ended. Although prior research shows that CJ is a promising method, more research is needed on the reliability, validity and feasibility of the method.

Inleiding

Competentiegericht onderwijs heeft haar volledige intrede in het hoger onderwijs gemaakt. Dit heeft gevolgen voor het evalueren van studenten. Het is niet alleen belangrijk om kennis en inzicht te toetsen, maar ook complexe competenties zoals bijvoorbeeld probleemoplossend denken, creativiteit, of samenwerken te evalueren. Voor het beoordelen van deze competenties zijn kennistoetsen vaak ontoereikend en wordt meer gebruik gemaakt van zogenaamde performance assessments (Lane & Stone, 2006). Bij performance assessments krijgen studenten opdrachten waarin ze moeten laten zien dat ze de juiste kennis, inzichten, attitudes en vaardigheden kunnen toepassen om in specifieke (praktijk-) situaties goed te functioneren (Baartman, 2008). Dergelijke opdrachten zijn vaak zeer open geformuleerd, waarin zowel het eindproduct als de wijze waarop studenten de opdracht aanpakken van belang zijn. Voor beoordelaars is het daardoor moeilijk of zelfs onmogelijk om van te voren vast te leggen wat goed of fout is. Het evalueren van competenties is hierdoor een stuk complexer dan het toetsen van kennis (Dierick & Dochy, 2001). De uitdaging voor de beoordelaar ligt erin dat de uitvoering van de taak omgezet moet worden naar één of meerdere scores die aangeeft/aangeven of en in welke mate de student de competentie beheerst (Baartman, 2008).

Er zijn verschillende methoden die de beoordelaar kan gebruiken om tot een score te komen. De twee meest gekende en toegepaste methoden zijn het analytisch scoren via criterialijsten en het holistisch scoren. Bij het analytisch scoren worden vooraf criteria opgesteld die gespecificeerde deelaspecten van de competentie aanhalen. Beoordelaars geven vervolgens per criterium aan in welke mate de studenten geslaagd zijn voor dit deelaspect. Door een (gewogen) som te nemen van de deelscores wordt het eindpunt bepaald (Van Petegem & Vanhoof, 2002). Bij holistisch scoren geven beoordelaars een punt op basis van een evaluatie van het geheel van het werk van de student. Dit punt reflecteert de mate waarin de student de competentie als geheel beheerst (Sadler, 2009).

Zowel analytisch als holistisch scoren kent echter nadelen en de toepassing ervan wordt lang niet door alle beoordelaars als bevredigend ervaren (Brooks, 2005; Sadler; 2009). Zowel de praktijk als de wetenschappelijke literatuur is daarom zoekende naar hoe performance assessments het beste gescoord kunnen worden. Vaak wordt er gekozen om analytisch en holistisch scoren te combineren of een tussenweg te formuleren. Beide methoden kennen echter een aantal kritiekpunten, wat de vraag doet rijzen of er alternatieven bestaan.

In dit artikel staan we stil bij enkele meetvraagstukken die komen kijken bij het scoren van performance assessments. Eerst wordt ingegaan op de kritieken op analytisch en holistisch scoren. Vervolgens wordt een alternatief voorgesteld, de comparatieve beoordelingsmethode (CB). In deze methode staat het onderling holistisch vergelijken van het werk van studenten centraal. Tot slot wordt ingegaan op hoe deze methode een antwoord kan bieden op de vaak aangehaalde kritiekpunten van de meer gekende methoden waarbij aandacht uitgaat naar de betrouwbaarheid, validiteit en praktische haalbaarheid van deze beoordelingsmethoden. Het wordt duidelijk dat CB een veelbelovend alternatief is, maar dat er nog veel onderzoek gedaan moet worden naar de voor- en nadelen.

Criteriagericht en holistisch scoren: enkele kritiekpunten

De meest beschreven en gekende methode om performance assessments te scoren is het gebruik van criterialijsten (Lane & Stone, 2006). Criterialijsten zorgen ervoor dat beoordelaars analytisch naar het werk van studenten kijken, omdat ze dit werk op verschillende deelaspecten moeten beoordelen en scoren. De optelling van deze scores leidt, volgens deze methode, tot een adequaat beeld van de mate waarin een student de competentie beheerst (Baird & Scharaschkin, 2002). Het gebruik van criterialijsten kan de beoordelaar genoeg houvast geven om elk werk onbevangen, dus los van ervaring en eerder beoordeeld werk, te beoordelen op de kwaliteit (Sadler, 2005). Deze methode maakt naar studenten toe duidelijk op welke aspecten zij wel en niet voldoende scoren (Jonsson & Svingby, 2007).

Een belangrijk kritiek op het gebruik van criterialijsten is dat in deze methode aangenomen wordt dat elke competentie uiteengerafeld kan worden in deelaspecten en dat de som van scores weergeeft of een student de competentie beheerst. Dit is een kritiek op de validiteit van de methode (Sadler, 2009). Doordat een criterialijst uit praktische overwegingen vaak enkel uit een selectie van de criteria bestaat, beslaat de beoordeling nooit het hele bereik van een competentie. Daarnaast kan er voor sommige opdrachten op voorhand niet genoeg informatie zijn om een volledige criterialijst te maken. Bijvoorbeeld als de uitkomsten zeer divers zijn, of als er om veel eigen input van de student wordt gevraagd (Kimbell et al., 2009). Een ander vaak gerapporteerde moeilijkheid is dat criterialijsten er zelden in slagen een evenwichtige aandacht voor proces en product te hebben. Zo is het moeilijk om een criterialijst te ontwerpen die de competentie creatief denken scoort, omdat in deze competentie het proces en het product onlosmakelijk met elkaar verbonden zijn (Jones, Swan & Pollitt, 2014). Kortom, een criterialijst doet niet altijd recht aan het integrale geheel van een competentie.

Ook blijkt dat beoordelaars het beoordelen via criterialijsten vaak als onbevredigend ervaren. Enerzijds kan er ongenoegen ontstaan wanneer zij het werk van studenten op zeer veel criteria moeten scoren, terwijl zij in één opslag zien of en waarom een student al dan niet geslaagd is. Anderzijds kan een beperktere criterialijst leiden tot de indruk dat niet alle relevante beoordelingscriteria worden geëvalueerd. Beoordelaars geven aan dat zij daarom tijdens het beoordelen ook met andere aspecten rekening houden dan vermeld in de criterialijst. Daarnaast hebben veel beoordelaars toch de behoefte om de werken van studenten met elkaar te vergelijken, om zo grip te krijgen op de kwaliteit. Een criterialijst biedt niet genoeg ondersteuning om elk werk onbevangen te kijken (Crisp, 2013). Hoe complexer de opdracht, hoe problematischer het werken met criterialijsten ervaren wordt (Heller, Sheingold & Myford, 1998; Schaaf, Stokking & Verloop, 2005).

Als alternatief maken beoordelaars vaak ook gebruik van holistisch scoren. Hierin bepalen beoordelaars op basis van hun algemene indruk van het werk of een student al dan niet geslaagd is en welk punt zij krijgen (Sadler, 2009). Een belangrijke kritiek op deze methode is dat het niet inzichtelijk is hoe en waarom de beoordelaar tot een bepaald punt komt. Op deze manier wordt een score afhankelijk van de focus van de specifieke beoordelaar. Vooral beginnende beoordelaars kunnen problemen ervaren bij het consistent beoordelen van het geheel van het werk, omdat zij weinig houvast hebben (Barkaoui, 2010). Hierdoor besteden beoordelaars vaak veel tijd en inspanning aan het ordenen en herbeoordelen van het werk van studenten om (mogelijke) inconsistentie te voorkomen (Crisp, 2010).

In de praktijk maken docenten vaak gebruik van een combinatie van beide beoordelingsmethoden. De aangehaalde probleempunten worden hierdoor echter maar gedeeltelijk ingelost. Het gebruik van meer comparatieve beoordelingsmethoden kan een mogelijke uitweg vormen.

Comparatief beoordelen als alternatief

Als reactie op deze beperkingen wordt sinds een aantal jaren onderzoek gedaan naar of en hoe je competenties kan meten via een systeem van comparatieve beoordelingen (CB), ofwel door prestaties van studenten te vergelijken. In deze beoordelingsmethode staat een holistische en zelfs intuïtieve aanpak centraal (Pollitt, 2012a). In het systeem van CB worden telkens prestaties van twee studenten onderling vergeleken, en geven beoordelaars aan welke beter is ten aanzien van de te beoordelen competentie. Figuur 1 licht via een abstract voorbeeld toe hoe de beoordelingstaak van beoordelaars kan leiden tot een rangorde. CB laat uiteraard toe om heel diverse en complexere prestaties te beoordelen dan het voorbeeld laat zien. Zo kunnen onder meer papers, filmpjes, foto's of portfolio's waarin studenten hun prestatie tonen, worden vergeleken.

Figuur 1: Comparatief beoordelen – van vergelijken naar een rangorde (naar idee van www.nomoremarking.com)

Om tot een uitspraak te komen over een reeks van prestaties, wordt aan meerdere beoordelaars gevraagd om een groot aantal vergelijkingen te maken. Achter de schermen worden de paren samengesteld en (random) verspreid over de verschillende beoordelaars. Op basis van deze vergelijkingen kan door middel van een statistisch model een rangorde worden gegenereerd die de prestaties ordent van minst goede naar beste prestatie. Dit statistisch model is het Bradley-Terry-Luce-Model, een specifieke vorm van het Rasch-model. Dit model vertaalt de uitkomst van de vergelijkingen naar de kans dat een prestatie van een andere wint (Pollitt, 2012b). Op deze manier ontstaat een schaal die uitdrukt hoe de ene prestatie zich verhoudt ten opzichte van alle andere prestaties (Bramley, 2007). Omdat het Rasch-model zeer robuust is, is het niet nodig dat alle mogelijke vergelijkingen gemaakt worden, om toch een betrouwbare rangorde te genereren (Whitehouse & Pollit, 2012).

De methode bouwt voort op Thurstones *"Law of comparative judgement"*, uit 1927. Hij stelt dat mensen betrouwbaarder zijn in het maken van een vergelijking, dan in het toekennen van een absolute waarde aan een prestatie. Laming (2004) stelt dat mensen niet alleen betrouwbaarder zijn in het vergelijken, maar dat *alle* beoordelingen ook vergelijkingen zijn. Een beoordelaar heeft namelijk een referentiepunt nodig om een waardeoordeel te kunnen geven. Dit maakt dat CB beter aansluit bij hoe mensen van nature beoordelingen maken (Greatorex, 2007). Door hen een referentiepunt te bieden en een holistische vergelijking te laten maken, zullen zij niet hun eigen referentiepunt gebruiken waardoor consistentie van een beoordelaar en over beoordelaars heen gewaarborgd kan worden (Pollitt & Crisp, 2004). Daarnaast vraagt de beoordeling minder cognitieve inspanning van beoordelaars, doordat zij enkel een vergelijking hoeven te maken (Jones et al., 2014). Ook voor beginnende beoordelaars zou het geven van een vergelijkingspunt genoeg ondersteuning moeten geven om een intuïtieve beoordeling te kunnen maken.

CB is niet alleen theoretisch beschreven, maar kent ook al enkele concrete toepassingen. Kimbell, Wheeler, Miller en Pollitt (2007) pasten de methode toe voor het vak design en technologie. 249 studenten hadden als opdracht om het doosje van een gloeilamp om te vormen zodat het een nieuwe functie kreeg. Zij maakten een portfolio waarin zij hun denkstappen weergaven met behulp van tekeningen, foto's en reflecties. Vervolgens bekeken zeven beoordelaars samen een aantal vergelijkingen, om zicht te krijgen op waar ze op moesten gaan letten tijdens het beoordelen. Daarna hebben ze elk individueel ongeveer 40 vergelijkingen gemaakt. Dit heeft geleid tot een rangorde van beste naar minst goede portfolio met een betrouwbaarheid van 0,93 (Kimbell et al., 2007). Deze betrouwbaarheidsmaat geeft aan in hoeverre de rangorde gebaseerd is op daadwerkelijke verschillen in kwaliteit van de prestaties ten opzichte van meetfouten.

Whitehouse en Pollitt (2012) pasten de methode toe op de antwoorden die studenten gaven op de stelling "*Soft engineering is a better river flood management strategy than hard engineering*" (Whitehouse & Pollitt, 2012, p. 5). Deze antwoorden waren ongeveer een tiental regels lang. 564 papers moesten beoordeeld worden door 23 docenten. Gemiddeld maakte elke docent 153 vergelijkingen. Dit leidde tot een rangorde met een betrouwbaarheid van 0,97. In deze studie laten de auteurs alle kansen zien die het werken met het Rasch-model biedt. Zo wordt ook aandacht besteed aan de beoordelaars en prestaties die zich niet volgens de verwachting van het model gedragen. Daarnaast geven Whitehouse en Pollitt (2012) inzicht in hoe de betrouwbaarheid stijgt naarmate er meer beoordelingen worden gemaakt.

In wat volgt lichten we verder toe hoe CB een antwoord kan bieden op enkele bekende problemen die zich voordoen wanneer competenties moeten worden gescoord.

Betrouwbaarheid

Betrouwbaarheid is een essentieel aspect van elk assessment. Betrouwbaarheid gaat onder meer over de vraag of de methode zorgt dat een waardering voor een prestatie niet wordt beïnvloed door de subjectiviteit van een beoordelaar of door het moment van beoordelen (Dierick & Dochy, 2001). In de analytische methode moeten gedetailleerde criterialijsten ervoor zorgen dat beoordelaars naar dezelfde aspecten kijken en daardoor tot eenzelfde punt komen. Zowel bij analytisch als bij holistisch scoren wordt door het vooraf trainen en gezamenlijk enkele beoordelingen maken getracht om de betrouwbaarheid te verhogen

(Harsch & Martin, 2013). Het blijkt echter vooral bij meer open taken moeilijk om een acceptabele betrouwbaarheid te realiseren (Jonsson & Svingby, 2007).

Bij CB hoeven beoordelaars enkel een vergelijking te maken en geen punten te geven. Hierdoor wordt ervoor gezorgd dat neigingen van beoordelaars om altijd streng of altijd soepel te scoren niet meer relevant zijn. Ook leidt dit ertoe dat het niet uitmaakt op welk moment je als student wordt beoordeeld, aan het begin of aan het eind van de reeks, op een zonnige of een regenachtige dag (Pollitt, 2012a).

In tegenstelling tot holistisch en analytisch oordelen wordt in deze methode niet van beoordelaars gevraagd om exact dezelfde aspecten te beoordelen. Zij mogen terugvallen op hun expertise en unieke perspectief op de competentie, omdat dit juist leidt tot een meer geïnformeerde score. *"Measuring the reliability of new forms of assessment stresses the need for more evidence in a doubtful case, rather than to rely on making inferences from a fixed and predetermined set of data"* (Martin, 1997 p. 342). Volgens Martin (1997) en Pollitt (2012b) moet betrouwbaarheid gaan over het gebruik maken van de diversiteit aan perspectieven om meer zicht te krijgen op de kwaliteit van een prestatie. Binnen CB wordt het werk van studenten altijd door meerdere beoordelaars bekeken. Een stijging in betrouwbaarheid betekent dat de score met meer precisie kan worden aangegeven. Deze kan worden verhoogd door het werk nog vaker te laten beoordelen (Pollitt, 2012a).

Door het gebruik van het statistisch Rasch-model waarin CB is ingebed, ontstaat inzicht in de verwachte en geobserveerde scores van prestaties. Hierdoor is het mogelijk om beoordelaars of prestaties die te sterk afwijken van de consensus te identificeren. Een mogelijke oorzaak is dat de beoordelaar een heel ander beeld heeft van wat de competentie behelst. Dit kan aanleiding zijn de beoordelaar extra te informeren of te trainen in hoe de beoordeling gemaakt moet worden. Wanneer er bij de beoordelaars onduidelijkheid is over de kwaliteit van bepaalde prestaties, wordt dit ook gesignaleerd. Deze prestatie kan dan vaker naar een (meer ervaren) beoordelaar worden uitgestuurd (Whitehouse & Pollitt, 2012). Op deze manier biedt CB verschillende mogelijkheden om de betrouwbaarheid van scores te verbeteren, na aanvang van het beoordelingsproces.

Validiteit

De aanhangers van CB stellen dat de analytische meetmethoden minder valide zijn (Jones & Alcock, 2013; Pollitt, 2012ab; Whitehouse & Pollitt, 2012). Een belangrijk aspect van validiteit is of de methode ervoor zorgt dat je meet wat je wilt meten (Whitehouse, 2012). Zoals eerder gesteld kan het opdelen van een competentie ervoor zorgen dat er geen recht meer gedaan wordt aan het integrale geheel van kennis, attitude en vaardigheid die het omvat. Bij het gebruik van criterialijsten wordt gesteld dat de criteria uitsluitend zijn, terwijl zij vaak echter een selectie van alle mogelijk criteria zijn (Sadler, 2009). Een beoordelaar kan net daardoor vertekend antwoorden, doordat hij of zij op basis van expertise andere (impliciete) criteria hanteert. Bij holistisch beoordelen valt dit bezwaar weg. Hierin hangt de validiteit echter sterk samen met de beoordelaar die de beoordeling maakt en waar hij of zij de keuze voor een bepaalde score op baseert (Harsch & Martin, 2013).

Het belangrijkste argument voor de validiteit van CB is dat deze holistische methode het integraal geheel van de competentie als voorwerp neemt (Sadler, 2009). Op deze manier wordt niet alleen naar de kwaliteitsaspecten gekeken die geëxpliciteerd kunnen worden,

maar naar het geheel aan impliciete en expliciete kenmerken van kwaliteit. Beoordelaars zijn goed in staat om op basis van hun expertise kwaliteit te herkennen (Pollitt, 2012a). Doordat de expertise van verschillende beoordelaars leidt tot een uitspraak van de score, wordt ervoor gezorgd dat de score geïnformeerd is en de rijkheid van een competentie omvat.

Ten tweede kan deze methode voor een valide meting van competenties zorgen, doordat een breed scala aan type taken kan worden beoordeeld (Jones & Alcock, 2013). Doordat de assessmentontwikkelaar niet gebonden is aan een vooraf op te stellen criterialijst, wordt het mogelijk meer open taken te ontwikkelen met bijvoorbeeld meer oog voor de praktijk. Hierdoor kan de focus van de beoordeling verplaatst worden van de criterialijst naar de kwaliteit van de prestatie (Whitehouse & Pollitt, 2012). Vervolgens kunnen studenten door middel van verschillende formats, bijvoorbeeld film, geschreven werk en foto's, bewijslast aanvoeren voor de verworven competentie. De cognitieve inspanning die op deze manier van studenten gevraagd wordt, sluit hierdoor meer aan bij de (beroeps)praktijk. Dit verhoogt de validiteit van het assessment (Linn, Baker & Dunbar, 1991).

Hoe beoordelaars het werk beoordelen speelt een grote rol in de validering van een methode (Bejar, 2012). Waar je met criterialijsten beoordelaars kan sturen in waar ze op letten en zo de validiteit kan verhogen, heb je bij CB deze middelen niet. In CB is psychologische validiteit, de vraag of beoordelaars wel naar de juiste kenmerken van prestaties kijken, cruciaal (Bramley, 2007). Zo moet er bijvoorbeeld in een assessment van argumentatief schrijven niet naar het handschrift gekeken worden maar wel naar hoe de tekst is opgebouwd. In een studie naar de beoordeling van een aardrijkskundeopdracht waarin studenten een stelling moeten beargumenteren, komt Whitehouse (2012) tot de conclusie dat beoordelaars hoofdzakelijk relevante argumenten aandragen voor hun keuze voor een van de twee prestaties. Het vragen naar een korte argumentatie voor de keuze, zorgt ervoor dat de validiteit van het assessment makkelijk na kan worden gegaan.

Praktische haalbaarheid van de meetmethoden

Beoordelingsmethoden dienen niet alleen betrouwbaar en valide te zijn. Om in de praktijk te worden gebruikt, moeten methoden ook zo praktisch en efficiënt mogelijk zijn (Dierick & Dochy, 2001). Zowel het gebruik van criterialijsten als holistisch scoren kan door een individuele docent met weinig middelen toegepast worden. Wanneer er met meer beoordelaars gewerkt wordt, is echter meer tijdinvestering in training en voorbereiding nodig.

CB is moeilijk om op grote schaal toe te passen zonder computer en internet. Op dit moment wordt gewerkt aan een aantal tools die het assessment via deze methode ondersteunen. Zo zijn in Engeland twee tools ontwikkeld, E-Scape (Kimbell et al., 2009) en NoMoreMarking (www.nomoremarking.com). Ook in Vlaanderen wordt gewerkt aan een digitaal platform die het comparatief beoordelen van competenties in onderwijs en organisaties kan ondersteunen, D-PAC (www.d-pac.be). In deze tools kunnen studenten hun werk uploaden in vorm van bijvoorbeeld een paper of filmpje. Vervolgens organiseert de tool het beoordelingsproces door het werk van studenten in koppels aan beoordelaars te tonen en bij te houden welke van welke prestatie gewonnen heeft.

Het is nog niet duidelijk of en in hoeverre CB tijdswinst oplevert ten opzichte van het werken met criterialijsten. Enerzijds kan het ontwerpen van criterialijsten en het trainen van beoordelaars in het gebruik ervan zeer tijdsintensief zijn. In CB wordt aan beoordelaars enkel

de competentie-omschrijving meegegeven, vervolgens wordt vertrouwd op de expertise van beoordelaars. Deze aanpak sluit aan bij hoe de methode betrouwbare en valide resultaten wil nastreven, en levert tegelijkertijd tijdswinst op (Whitehouse & Pollitt, 2012). Anderzijds wordt een prestatie veel vaker beoordeeld, wat uiteraard meer tijd kost.

Pollitt (2012b) gebruikt binnen E-Scape verschillende methoden die het assessment efficiënter maken. Het basisidee is dat prestaties niet puur random aan elkaar worden toegewezen, maar dat er enkel vergelijkingen worden gemaakt die informatief zijn. Zo kunnen er vanaf de eerste ronde enkel papers aan elkaar gekoppeld worden die even vaak gewonnen hebben. Wanneer er genoeg vergelijkingen zijn gemaakt, kan op basis van de plaats op de rangorde een paar worden samengesteld. Tot slot kan het systeem ervoor zorgen dat beoordelaars regelmatig dezelfde prestaties zien, waardoor zij de tweede keer minder tijd nodig hebben om een beeld te krijgen van de kwaliteit van het werk.

Het is in deze methode echter noodzakelijk om over meerdere beoordelaars te beschikken. Dit neemt niet weg dat bij criterialijsten doorgaans nauwelijks aandacht besteed wordt aan het berekenen van de betrouwbaarheid, omdat met het terugvallen op één beoordelaar geen interbeoordelaarsbetrouwbaarheid berekend kan worden. Wanneer er wel meerdere beoordelaars betrokken zijn bij het beoordelen, is de tijdsinvestering van het dubbel-beoordelen zeer hoog.

Tot slot

In het hoger onderwijs neemt de populariteit van competentiegericht onderwijs toe, hiermee groeit ook de aandacht voor hoe je competenties meer kwaliteitsvol kan meten. Daarbij is er ruimte ontstaan om alternatieve beoordelingsmethoden te onderzoeken en te kijken hoe en wanneer zij kunnen bijdragen aan betrouwbare, valide en efficiënte evaluaties. CB biedt een alternatief perspectief op het scoren van competenties, dat beter aansluit bij hoe mensen van nature beoordelen.

In CB wordt aan meerdere beoordelaars gevraagd om op basis van een holistische vergelijking aan te geven welke van twee prestaties beter is in termen van de te beoordelen competentie. Deze methode ziet betrouwbaarheid als de precisie waarmee de kwaliteit aangegeven kan worden. Validiteit wordt gewaarborgd doordat beoordelaars een uitspraak doen over het geheel van de prestatie, waarin ruimte is voor de expertise van de beoordelaar. Daarnaast hoeft de opdracht niet meer aan te sluiten bij een van te voren gedefinieerde criterialijst.

Alhoewel deze beoordelingsmethode vanaf 2007 wordt toegepast voor het evalueren van competenties, is het empirisch onderzoek naar de meerwaarde beperkt. Er zijn nog veel vragen. Allereerst is het aantal studies naar hoe en wanneer betrouwbare en valide scores worden verkregen nog nauwelijks onderzocht. Ten tweede is er nog geen duidelijkheid over de efficiëntie van deze methode. Mogelijk is deze methode niet in alle situaties praktisch haalbaar vanwege een tekort aan tijd en beoordelaars. CB kan echter een goed alternatief zijn, vooral voor het beoordelen van meer complexe competenties die steeds meer aandacht krijgen in de opleidingen van het hoger onderwijs. Het biedt de mogelijkheid taken te ontwikkelen waarin studenten grote vrijheid hebben in de uitvoering en uitgedaagd worden na te denken over aanpak en eindproduct. Hierdoor zou de methode goed kunnen worden toegepast om opdrachten die binnen een stageomgeving of in groepen worden uitgevoerd te

beoordelen. Vervolgonderzoek is noodzakelijk om de meerwaarde en beperkingen van deze methode voor de evaluatiepraktijk van het hoger onderwijs in kaart te brengen.

Referenties

Baartman, L. K. (2008). *Assessing the assessment: Development and use of quality criteria for competence assessment programmes.* Dissertation, Universiteit Utrecht.

Baird, J-A. & Scharaschkin, A. (2002). Is the whole worth more than the sum of the parts? Studies of examiners' grading of individual papers and candidates' whole a-level examination performances. *Educational Studies*, *28* (2), 143-162, DOI: 10.1080/03055690220124588.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7* (1), 54-74.

Bejar, I.I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice, 31* (3), 2-9.

Bramley, T. (2007). Paired comparison methods. In P. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms (eds.), *Techniques for monitoring the comparability of examination standards* (pp. 246–294). London: QCA.

Brooks, V. (2009). Marking as judgment. *Research Papers in Education, X* (X), 1-34, DOI: 10.1080/02671520903331008.

Crisp, V. (2010). Judging the grade: Exploring the judgement processes involved in examination grading decisions. *Evaluation & Research in Education, 23* (1), 19-35, DOI: 10.1080/09500790903572928.

Crisp, V. (2013). Criteria, comparison and past experience: How do teachers make judgements when marking coursework? *Assessment in Education: Principles, Policy & Practice, 20* (1), 127-144, DOI: 10.1080/0969594X.2012.741059.

Dierick, S. & Dochy, F. (2001). New lines in edumetrics: New forms of assessment lead to new assessment criteria. *Studies in Educational Evaluation, 27,* 307-329.

Greatorex, J. (2007). Conteporary GCSE and A-level awarding: A psychological perspective on the decision-making process used to judge the quality of candidates' work. Paper presented at BERA 2007.

Harsch, C. & Martin, G. (2013). Comparing holistic and analytic scoring methods: Issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice, 20* (3), 281-307.

Heller, J.I., Sheingold, K. & Myford, C.M. (1998). Reasoning about evidence in portfolios: Cognitive foundations for valid and reliable assessment. *Educational Assessment, 5* (1), 5-40, DOI: 10.1207/s15326977ea0501_1.

Jones, I. & Alcock, L. (2013). Peer assessment without assessment criteria. *Studies in Higher Education, X* (X), 1-14.

Jones, I., Swan, M. & Pollitt, A. (2014). Assessing mathematical problem solving using comparatieve beoordeling. *International Journal of Science and Mathematics Education, X*(X), xx-xx.

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

Kimbell, R., Wheeler, T. Miller, S. & Pollitt, A. (2007). *E-scape portfolio assessment: Phase 2 report*. Goldsmiths College: University of London.

Kimbell, R., Wheeler, T., Stables, K., Shepard, T., Davies, D., Martin, F., Pollitt, A., & Whitehouse, G. (2009). *E-scape portfolio assessment: Phase 3 report*. Goldsmiths College: University of London.

Laming, D. R. J. (2004). *Human judgment: The eye of the beholder.* London, Thomson.

Lane, S. & Stone, C. A. (2006). Performance Assessments. In B. Brennan (Ed.), *Educational Measurement.* American Council on Education & Praeger: Westport, CT.

Linn, R.L., Baker, E.L. & Dunbar, S.B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20* (8), 15-21.

Martin, S. (1997). Two models of educational assessment: a response from initial teacher education: if the cap fits… *Assessment & Evaluation in Higher Education, 22* (3), 337-343, DOI: 10.1080/0260293970220307.

Pollitt, A. (2012a). The method of adaptive comparatieve beoordeling. *Assessment in education: principles, policy, & practice, 19 (3),* 281-300, DOI:10.1080/0969594X.2012.665354.

Pollitt, A. (2012b). Comparatieve beoordeling for assessment. *International Journal of Technology and Design Education, 22*, 157–170.

Pollitt, A. & Crisp, V. (2004) Could comparative judgements of script quality replace traditional marking and improve the validity of exam questions? Paper presented at the BERA Annual Conference, UMIST Manchester, September 2004.

Sadler, D.R. (2005). Interpretations if criteria-based assessment and grading in higher education. *Assessment & Evaluation in Higher Education, 30* (2), 175-194.

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

Schaaf, M.F. van der, Stokking, K.M. & Verloop, N. (2005). Cognitive representations in raters' assessment of teacher portfolios. *Studies in Educational Evaluation, 31,* 27-55.

Thurstone, L. L. (1927). The law of comparatieve beoordeling. *Psychology Review, 34*, 273-286.

Van Petegem, P. & Vanhoof, J. (2002). *Evaluatie op de testbank. Een handboek voor het ontwikkelen van alternatieve evaluatievormen.* Mechelen: Wolters Plantyn.

Whitehouse, C. (2012). Testing the validity of judgements about geography essays using the Adaptive Comparatieve beoordeling method. *Centre for Education Research and Policy.*

Whitehouse, C. & Pollitt, A. (2012). Using adaptive comparatieve beoordeling to obtain a highly reliable rank order in summative assessment. *Centre for Education Research and Policy.*