Hoe betrouwbaar is peer-assessment?

Twee empirische studies naar student-beoordelingen

Drs. F.R. Kappe (Rutger.Kappe@inholland.nl) is werkzaam als onderzoeker en onderwijskundig adviseur bij de Dienst Onderwijs, Kwaliteit, Research & Development van de Hogeschool INHOLLAND.

In een assessment center context blijken HRM-studenten van Hogeschool INHOLLAND in staat om de competenties van medestudenten betrouwbaar te beoordelen. In een presentatiecontext blijken zij daarentegen daartoe niet in staat, met uitzondering van het gezamenlijk eindoordeel dat zij geven, in de vorm van een eindcijfer. Deze 'overall' beoordeling blijkt sterk overeen te komen met het oordeel van de docent. De resultaten uit beide deelstudies duiden erop dat studenten niet in staat zijn om zonder training en uitgewerkte beoordelingscriteria tot betrouwbare beoordelingen te komen. Wel zijn zij in staat om een betrouwbare en valide overall beoordeling, zoals een eindcijfer, te geven. De resultaten van beide studies bieden aanknopingspunten voor ontwerpers en gebruikers van peer-assessment in het onderwijs.

Inleiding

Met competentiegericht leren hebben nieuwe toetsvormen als self-, peer- en co-assessment zich aangediend. Van studenten wordt verwacht dat zij kunnen (of leren) reflecteren op hun eigen prestaties, leren omgaan met feedback en feedback leren geven aan anderen. Daarvoor dienen zij in staat te zijn zichzelf en/of een ander objectief en eerlijk – in toetstermen betrouwbaar – te beoordelen. Onderzoek waarbij studenten betrokken worden in de beoordeling van het werk van medestudenten (Beatty, Haas & Sciglimpaglia, 1996; Sherrard & Raafat, 1994) toont aan dat deze methode van beoordelen als eerlijk (McIlveen, Greenan & Humphreys, 1997; McDowell, 1995; Strachan & Wilcox, 1996) en leerzaam (Orsmond, 1996) wordt ervaren en een positief effect op het leren heeft (Boud, 1988; Falchikov, 1986). Op basis van onderzoek in het Nederlandstalig hoger onderwijs concluderen Dochy, Admiraal en Pilot (2003) dat het betrekken van studenten bij het beoordelingsproces (assessment) kan leiden tot meer ownership, hogere motivatie, diepgaand leren, betere transfer en betere academische, communicatieve en reflectievaardigheden. De Volder en Kappe (2007) geven aan dat peer-assessment de mogelijkheid biedt om gebruik te maken van de *neglected teaching resource of the students themselves*. Sluijsmans (2002) merkt in haar proefschrift over peer-assessment op dat een training in het werken met peer-assessment de reflectie- en beoordelingsvaardigheden van studenten vergroot.

Dat studenten peer-assessment eerlijk vinden en als leerzaam beschouwen is geen garantie dat zij betrouwbare en valide oordelen leveren. Belangrijk onderzoek rond deze kwestie is uitgevoerd door Falchikov en Goldfinch (2000). In hun meta-analyse

naar de betrouwbaarheid en validiteit van studentbeoordelingen vinden zij onder meer een aanzienlijke correlatie (r= .69) tussen student- en expertbeoordelingen.
Zijn er ook bezwaren tegen het inzetten van studenten als beoordelaars? Onderzoek van McIllveen e.a. (1997) toont bijvoorbeeld aan dat bedrijfskundestudenten zich niet comfortabel voelen bij peer-assessment. Een ander punt is de mate waarin studenten in staat zijn tot objectieve oordeelsvorming (Sluijsmans, 2002). Vooral de toegeeflijkheidfout in de zin dat ze te lief voor elkaar zouden zijn, wordt vaak als tegenargument gebruikt (Pond, 1995).

Onderzoeksvraag
Dit artikel levert een empirische bijdrage aan de discussie over het betrekken van studenten bij het beoordelingsproces. De centrale vraag is of studenten in staat zijn de prestaties van medestudenten op een betrouwbare wijze te beoordelen. In twee studies wordt de mate van overeenstemming berekend tussen de studentbeoordelaars, ook wel de interbeoordelaarsovereenstemming genoemd.
In studie 1 staat de betrouwbaarheid van studentbeoordelingen centraal. De verwachting is dat de studentbeoordelingen in studie 1 betrouwbaar zullen blijken. De mate van betrouwbaarheid zal enigszins verschillen per competentie, afhankelijk van de kwaliteit van de operationalisering van de competentie.
In studie 2 staat eveneens de betrouwbaarheid van studentbeoordelingen centraal. Daarnaast wordt in deze studie de validiteit van studentbeoordelingen bestudeerd middels het vergelijken van hun beoordelingen met die door een docent. In dit valideringsdeel wordt tevens de aanwezigheid van de toegeeflijkheidfout binnen studentbeoordelingen onderzocht.

Methode

Onderzoekspopulatie (studie 1 en 2)
Voor deze studie zijn de beoordelingen die zijn gegeven door vierdejaars HRM-studenten van Hogeschool INHOLLAND gebruikt. In studiejaar 2005-2006 participeerden 34 studenten als assessor, van wie 35% man en 65% vrouw; in studiejaar 2006-2007 participeerden 26 studenten als assessor, van wie 42% man en 58% vrouw.

Onderzoekscontext en instrumenten (studie 1 en 2)
In beide studies gebruikten de assessoren een gestandaardiseerd beoordelingsformulier voorzien van een vijfpunts beoordelingsschaal. Het beoordelingsproces in beide studies is zo ingericht dat de studentassessoren eerst onafhankelijk van elkaar beoordelen, daarna in overleg treden en gezamenlijk het beoordelingsformulier invullen. Hierna worden specifieke kenmerken van studie 1 en studie 2 beschreven.

Studie 1
In studie 1 zijn beoordelingen verzameld binnen een assessment center context. (Voor een uitgebreide beschrijving van dit assessment center wordt verwezen naar Reichter & Rotteveel, 2004). De vierdejaarsstudenten beoordeelden, in elke keer random toe-

gewezen duo's, eerste- en derdejaarsstudenten HRM op een aantal competenties (onder andere adviseren, analyseren, oordeelsvorming). In totaal hebben de 60 assessoren 34 eerstejaars- en 59 derdejaarsstudenten beoordeeld.
Het assessment center voor de eerstejaars, met een duur van een dag, bestaat uit een individueel adviesgesprek, een teamoverleg, een tweede groepsopdracht en een zelfassessment. Het assessment center voor de derdejaars met dezelfde duur bestaat uit een individueel adviesgesprek, een postbakopdracht, een teamoverleg en een zelfassessment. De studentassessoren hebben voorafgaand aan hun beoordelingstaak een intensieve vijfdaagse training gevolgd, gericht op consensusvorming. In de training is gewerkt met videofragmenten van de daadwerkelijke assessmentopdrachten. De beoordeling van de competenties is gegeven op een gestandaardiseerd beoordelingsformulier, waarbij elke competentie is voorzien van een vijfpunts Likert-schaal met gedragsankers per schaalpunt (BARS: Behavioral Anchored Rating Scale, zie bijvoorbeeld Guion, 1998). Voor de berekening van de interbeoordelaarsbetrouwbaarheid zijn de individuele beoordelingsgegevens van de studentassessoren per geassesste student gebruikt.

Studie 2
In studie 2 zijn beoordelingen verzameld van groepspresentaties. Vierdejaarsstudenten HRM beoordeelden hun klasgenoten. De beoordeling is gegeven op een gestandaardiseerd beoordelingsformulier, waarbij per beoordelingsaspect (structuur, inhoud, techniek) een aantal algemene gedragsankers is vermeld. Voor de berekening van de betrouwbaarheid zijn de individuele beoordelingsgegevens van de studentbeoordelaars gebruikt. In deze beoordelingscontext beoordeelt ook een docent dezelfde groepspresentaties. Daarmee is het mogelijk om naast de betrouwbaarheid een indruk te krijgen van de validiteit van de studentbeoordelingen.

Analyse
De mate van overeenstemming tussen de beoordelaars wordt vaak berekend met behulp van een Pearson's correlatie of Cohen's Kappa. Een afgeleide daarvan, de Fleiss Kappa (κ), is voor dit onderzoek de meest geschikte maat, daar deze rekening houdt met het feit dat er met meerdere beoordelaars wordt gewerkt (Fleiss, 1971). Daarnaast houdt de Fleiss Kappa, net als de Cohen's Kappa, rekening met de kans dat beoordelaars bij toeval tot eenzelfde beoordeling komen. De Fleiss Kappa is interpreteerbaar als een reguliere correlatiecoëfficiënt waarbij waarden tussen .61 en .80 als substantieel worden beschouwd en waarden >.80 als acceptabel dan wel zeer acceptabel (Landis & Koch, 1977; Marcoulides, 1989).
In studie 2 wordt eveneens de Fleiss Kappa gerapporteerd als overeenstemmingsmaat tussen de beoordelingen door de studenten en de docent. Voor de beantwoording van de vraag of studenten toegeeflijker zijn in hun oordelen dan de docent wordt de effectgrootte (d) berekend.

Resultaten

Studie 1

De Fleiss Kappa is per competentie, als overeenstemmingsmaat tussen de assessoren, berekend. De resultaten per competentie van het assessment center van jaar 1 en het assessment center in jaar 3 staan in tabel 1.

Tabel 1 Fleiss Kappa coëfficiënten per competentie per assessment center

	Assessment Center jaar 1	**Assessment Center jaar 3**
	κ	κ
Analyseren	.81	.84
Communiceren	.89	.87
Adviseren	.88	.82
Rapporteren	.91	.91
Alternatieven aandragen	.82	.73
Samenwerken	.86	.75
Verantwoorden	.89	.78
Presenteren	.87	.74
Leerdoelen formuleren	.95	.84
Positie bepalen	n.b.	.71
Oordeelsvorming	n.b.	.73
Besluitvaardigheid	n.b.	.84
Relatie aangaan	n.b.	.85

De Fleiss Kappa coëfficiënten zijn alle significant met $\alpha<0.01$, $34<n<59$. n.b. = niet beoordeeld

Uit tabel 1 blijkt dat alle gevonden κ-waarden van de competenties, beoordeeld in het assessment center van jaar 1, hoger zijn dan de .80 norm (Landis & Koch, 1977; Marcoulides, 1989). In het assessment center van jaar 3 voldoet een aantal competenties net niet aan deze norm maar liggen wel boven de .70. Een gemiddelde Fleiss Kappa is berekend als overall maat voor de betrouwbaarheid van beide assessment centers. Voor jaar 1 is deze .88, voor jaar 3 is deze .81.

Studie 2

In studie 2 is eveneens de betrouwbaarheid van studentbeoordelingen onderzocht. Ook hier is de Fleiss κ per beoordelingsaspect berekend. De gevonden lage coëfficiënten van .22 voor het beoordelingsaspect structuur (p=.112), .29 voor het aspect inhoud (p=.037) en .33 voor presentatietechniek (p=.015) duiden op een geringe mate van overeenstemming tussen de studentbeoordelaars. De betrouwbaarheid van het eindcijfer, dat studenten geven op een tienpuntsschaal als een overall oordeel van de groepspresentatie, is met een waarde van .64 wel als substantieel te beschouwen.

In tabel 2 staan Fleiss κ-coëfficiënten die de mate van overeenstemming weergeven tussen de gezamenlijke studentbeoordeling en de docentbeoordeling. In tabel 2 staan

coëfficiënten met waarden van .45 voor het beoordelingsaspect techniek, .47 voor structuur en .61 voor inhoud en .68 voor het eindcijfer. Deze resultaten geven aan dat deelbeoordelingen (3 aspecten) minder valide zijn dan een overall oordeel.

Tabel 2 Beschrijvende en toetsende statistieken voor de relatie tussen student- en docentbeoordelingen

	Studenten	**Docent**	κ	**d**	**t**	**df**	**p**
Structuur	3.31	3.22	.47	.18	1.176	52	.245
	(SD .462)	(SD .547)					
Inhoud	3.27	3.25	.61	.04	.472	52	.639
	(SD .551)	(SD .588)					
Techniek	3.46	3.30	.45	.32	1.624	52	.110
	(SD .664)	(SD .697)					
Eindcijfer	6.76	6.78	.68	-.02	-.149	29	.882
	(SD .573)	(SD .738)					

p=2-zijdig

In tabel 2 staan tevens per beoordelingsaspect de effectgrootte (d) en de resultaten van een paired sample t-test. Een kleine effectgrootte duidt in dit verband op een klein verschil tussen de variabelen (beoordelaars), waarbij een positieve waarde duidt op *overmarking* door studenten en een negatieve waarde op *undermarking* (Falchikov & Goldfinch, 2000, p. 293). In combinatie met de resultaten van de t-test (kolom 6-8) kan worden geconcludeerd dat er geen significante verschillen bestaan tussen de gezamenlijke studentbeoordelingen en de docentbeoordelingen. Hoewel de gemiddelden (kolom 2 en 3) anders doen vermoeden, geeft de d-waarde aan dat studenten geneigd zijn wat toegeeflijker te zijn op deelaspecten, maar juist strenger bij het toekennen van een eindcijfer. Aangezien de verschillen niet significant zijn, moet de nulhypothese dat er geen verschillen tussen de beoordelaars zijn, worden aangenomen. Er is dus geen sprake dat studenten toegeeflijker zijn voor hun medestudenten dan de docent. Een aanvullende analyse van de gegevens is uitgevoerd voor het eindcijfer middels kruistabellen. Resultaat van deze analyse is dat in 23% van de gevallen de gezamenlijke studentbeoordeling en docentbeoordeling qua eindcijfer precies overeenkomt, in 37% van de gevallen beoordelen studenten hoger en in 40% van de gevallen beoordeelt de docent hoger. Wanneer het eindcijfer afgerond wordt op een heel cijfer (wat vaak de praktijk is), dan veranderen de percentages als volgt: overeenkomstig: 77, studenten hoger: 10, docent hoger: 13.
De resultaten van studie 2 tonen aan dat hoewel studenten het onderling niet direct met elkaar eens zijn, zij na overleg in staat zijn tot een redelijk valide oordeel te komen als het een overall oordeel, zoals een eindcijfer, betreft.

Discussie

In dit artikel staat de betrouwbaarheid van studentbeoordelingen centraal. Middels twee studies in deels overeenkomstige, deels verschillende beoordelingscontexten is de

betrouwbaarheid van studentbeoordelingen onderzocht. In deze discussiesectie worden eerst beide studies separaat besproken.

Studie 1: betrouwbaarheid en validiteit studentbeoordelingen

De resultaten van studie 1 wijzen erop dat getrainde studenten in staat blijken de competenties van medestudenten betrouwbaar te beoordelen (.71<κ<.95). De vergelijking van de overall betrouwbaarheidscoëfficiënten van beide assessment centers (.88 voor jaar 1 en .81 voor jaar 3) duidt erop dat de moeilijkheidsgraad en het type assessmentopdrachten geen sterk bepalende factoren zijn voor wat betreft de mate waarin assessoren in staat zijn tot overeenstemming te komen. Dat er in het assessment center van jaar 3 enkele competenties net niet aan de .80 norm voldoen, kan veroorzaakt zijn doordat de opdrachten van dit assessment center minder (of minder relevant) competentiegerelateerd gedrag, dat overeenkomt met de gedragsankers van de beoordelingsschalen, uitlokken en daardoor de beoordeling bemoeilijken. Het kan overigens ook zijn dat de opdrachten juist meer competentiegerelateerd gedrag uitlokken, wat eveneens een beoordeling kan bemoeilijken doordat de assessoren meer gedrag moeten interpreteren en wegen. Aanvullend kwalitatief onderzoek onder de assessoren kan inzicht op dit punt verschaffen.

Een zeer hoge mate van overeenstemming is gevonden voor de competentie leerdoelen formuleren, gemeten in het assessment center van jaar 1 (κ=.95 voor jaar 1; κ=.84 voor jaar 3). De (gedrags)ankers bij deze competentie zijn zeer helder en concreet geformuleerd en beperkt wat betreft aantal.

Uit eerder evaluatieonderzoek van Kappe (2005) onder de assessoren en geassesste studenten bleek dat alle betrokkenen positief waren over de inzet van studenten als assessoren. De studentassessoren gaven daarbij aan dat zij het een leerzame en beroepsmatig relevante leeractiviteit vonden. Dit resultaat komt overeen met menige studie naar de inzet van studentassessoren en is in tegenspraak met het resultaat van McIlveen e.a. (1997), die vonden dat studenten zich niet comfortabel voelen bij peer-assessment. Uit onderzoek van Kappe en Scholten-Linde (2005) bleek tevens dat geassesste studenten gevolg gaven aan hun assessment uitslag en er actief mee aan de slag gingen in het kader van hun persoonlijk ontwikkelplan, de stage en het afstuderen. Op basis van dat onderzoek concludeerden de auteurs dat de assessment centers daarmee een zekere consequentiële validiteit bezitten.

Samengevat duiden de resultaten van studie 1 erop dat studenten in staat zijn om competenties van medestudenten op een betrouwbare wijze te beoordelen. Daarbij kan worden vastgesteld dat alle betrokken studenten de assessments als een waardevol leermoment ervaren en dat de geassesste studenten actief met de feedback aan de slag gaan. Deze positieve bevindingen sluiten aan bij eerdere studies van Dochy e.a. (2003) en onderschrijven de veronderstelling van De Volder en Kappe (2007) dat studenten beschouwd kunnen worden als een potentiële bron van feedbackgevers.

Kanttekening bij studie 1 is het ontbreken van een expertbeoordeling. Hoewel de beoordelingen van de studenten betrouwbaar zijn gebleken, bestaat de mogelijkheid dat deze niet valide zijn. Aanleiding om echter een zekere validiteit van de studentbeoordelingen te veronderstellen is de unaniem hoge beoordelingen door geassesste studenten betreffende de kwaliteit van het optreden van de assessoren, de waardering

voor het terugkoppelingsgesprek over de resultaten, en de waardering voor het gehele assessment center. Aangenomen mag worden dat als zij zich niet zouden herkennen in hun competentiebeoordelingen – wat een sterke aanwijzing is voor een niet valide meting –, dat dat een lage waardering van het assessment center tot gevolg zou hebben. Verder methodisch onderbouwd onderzoek, bijvoorbeeld op basis van het design van studie 2 (bij voorkeur met meer dan 1 expert), zou meer eenduidig uitsluitsel kunnen bieden.

Studie 2: betrouwbaarheid en validiteit student- en docentbeoordelingen

De resultaten van studie 2 wijzen niet in dezelfde richting als de resultaten van studie 1. De beoordelaars slagen er voor deze beoordelingstaak niet in op drie beoordelingsaspecten (structuur, κ=.22; inhoud, κ=.29; techniek, κ=.33) een acceptabele mate van overeenstemming te bereiken. In tegenstelling tot deze lage betrouwbaarheden blijkt het eindcijfer dat de studenten gezamenlijk geven wel betrouwbaar (κ=.64, p=.000). Een opvallend resultaat, omdat het eindcijfer min of meer beschouwd kan worden als een optelsom van de deelbeoordelingen en studenten bij het geven van een eindcijfer geen concrete maatstaven hebben om hun beoordeling aan te relateren. Mogelijke verklaring is dat studenten wel hetzelfde geobserveerd hebben, maar dat niet altijd aan hetzelfde beoordelingsaspect toekennen (classificeren). Bij het geven van het eindcijfer effenen de verschillen in classificeren zich uit, wat leidt tot een hogere mate van overeenstemming voor het eindcijfer. Uit de assessment center literatuur is bekend dat assessoren veelal moeite hebben met dit classificeren van geobserveerd gedrag bij de juiste competentie. Het gevonden resultaat heeft ertoe geleid dat de beoordelingsaspecten op het gebruikte beoordelingsformulier zijn aangescherpt. Vervolgonderzoek zal moeten aantonen of het aanscherpen van de beoordelingsaspecten leidt tot een verhoging van de betrouwbaarheid van de beoordelingsaspecten. De verwachting is dat er enig effect zal zijn, maar dat het eindoordeel de meest betrouwbare maat zal blijven. In zijn algemeenheid wordt namelijk aangenomen dat meer globale beoordelingsmaten meer variantie in metingen verklaren en betrouwbaarder zijn (Ones & Viswesvaran, 1996).

In studie 2 was het mogelijk om de studentbeoordelingen te relateren aan de beoordeling door een docent. Resultaat daarvan is een matige overeenstemming op de beoordelingsaspecten (.45<κ<.61) tot een acceptabele mate van overeenstemming voor het eindcijfer (κ=.68). Deze coëfficiënt is vergelijkbaar met resultaten uit de metastudie van Falchikov en Goldfinch (2000), die, op basis van 48 studies, een correlatie van .69 vinden tussen student- en expertbeoordelingen op globale beoordelingsmaten.

Gezien de ruime ervaring van docenten wordt vaak verondersteld dat hun beoordelingen betrouwbaar zijn. Bij de vergelijking tussen de student- en docentbeoordelingen in studie 2 is van die veronderstelling uitgegaan. Echter, het niet kunnen aantonen van de betrouwbaarheid van de docentbeoordeling kan als een zwak punt in het design van studie 2 beschouwd worden. Geadviseerd wordt om ook docenten gezamenlijk te laten beoordelen en daarvoor overeenstemmingsmaten te berekenen. Dergelijk onderzoek kan tot interessante inzichten leiden.

Wat betreft de beoordelingsfout dat studenten voor elkaar 'te lief' zouden zijn (Pond, 1995) is geen bewijs gevonden. Berekening van de effectgrootte (d) geeft aan dat

hoewel studenten geneigd zijn op de beoordelingsaspecten structuur, inhoud en techniek iets toegeeflijker te zijn, zij dat juist niet zijn bij het toekennen van het eindcijfer. De verschillen in beoordelingen tussen de beoordelaars blijken niet significant, wat betekent dat er feitelijk geen verschil is tussen het gezamenlijk studentoordeel en het docentoordeel. Dit resultaat duidt op een zekere mate van validiteit van studentbeoordelingen. Een aanvullende analyse middels kruistabellen toont aan dat er – in het hypothetische geval waarbij de beoordeling geheel door studenten zou zijn gegeven – drie presentaties 'bevoordeeld' zouden zijn omdat de docent lager zou hebben beoordeeld. In vier gevallen zouden de beoordeelden 'benadeeld' zijn omdat de docent in die gevallen hoger zou hebben beoordeeld.

Studie 1 en 2: training en beoordelingscriteria

Ondanks het verschil in design levert een vergelijking van de resultaten van beide studies enkele interessante gezichtspunten op. De mate van overeenstemming tussen beoordelaars in studie 1 is groter dan in studie 2. De hoogste mate van overeenstemming in de prestatiecontext (κ=.68 voor het eindcijfer) is nog lager dan de laagst gevonden waarde in de assessment center context (κ=.71 voor de competentie positie bepalen). Een nadere vergelijking van de resultaten uit beide studies is mogelijk voor het beoordelingsaspect presenteren. In studie 1 wordt de competentie presenteren beoordeeld en in studie 2 het aspect presentatietechniek, beide op een vijfpuntsschaal. Beide zijn echter in verschillende mate geoperationaliseerd (BARS versus aantal algemene ankers). Voor de competentie presenteren worden coëfficiënten van .87 (assessment center van jaar 1) en .74 (assessment center van jaar 3) gevonden, tegen .33 voor het beoordelingsaspect presentatietechniek. Dit aanzienlijke verschil kan verklaard worden door de mate van training die de assessoren hebben ontvangen en de mate waarin de beoordelingscriteria waarmee zij werkten, zijn geoperationaliseerd. Veel onderzoeken (Kolk, 2001; Sluijsmans, 2000; Lievens, 2001) tonen positieve effecten aan van assessorentraining.

Meer studies naar de betrouwbaarheid van peer-assessment dienen te worden uitgevoerd om inzicht te krijgen in de mate waarin de resultaten van dit onderzoek generaliseerbaar zijn naar andere studierichtingen en beoordelingscontexten. Falchikov en Goldfinch (2000) vinden overigens geen verschillen in de betrouwbaarheid en validiteit van peer-assessment tussen verschillende studierichtingen. Over de steekproef kan worden opgemerkt dat deze van redelijke omvang is voor een onderwijsgerelateerd onderzoek.

Conclusie

Dit onderzoek laat zien dat studenten in staat zijn betrouwbaar te beoordelen wanneer ze daarvoor zijn getraind en beoordelen aan de hand van uitgewerkte beoordelingsschalen (studie 1). Daarnaast is duidelijk geworden dat zij in staat zijn om een betrekkelijk betrouwbaar en valide overall oordeel (eindcijfer) te geven, zonder dat er specifieke ankers zijn. De conclusie dat studenten in staat zijn betrouwbaar te beoordelen is waardevol, aangezien het betrekken van studenten bij beoordelingen allerlei positieve leereffecten kan hebben (Dochy e.a., 2003; Sluijsmans, 2002). Voor alle studies die

opleiden tot managementfuncties waarbij beoordelingstaken moeten worden uitgevoerd, is een training beoordelingsvaardigheden en het opdoen van enige ervaring als beoordelaar nuttig.

In het geval van weinig specifieke (gedrags)ankers bij beoordelingscriteria blijken studenten beter in staat om eindcijfers te geven dan gedetailleerde beoordelingen. Bij zeer specifieke gedragsankers per beoordelingsaspect en schaalpunt (1-5), zogenoemde BARS, blijken studenten wel in staat om betrouwbare beoordelingen te leveren. Bij het inzetten van studenten in het beoordelingsproces is het dus van belang om het doel daarvan duidelijk voor ogen te hebben en op basis daarvan keuzes te maken voor wat betreft de beoordelingscriteria (breed versus specifiek) en de operationalisering ervan (beperkt versus uitgewerkt). Zoals aangegeven door Gueldenzoph en May (2002) is het effectief uitvoeren van peer-assessment een kwestie van maatwerk.

Noot

Dit artikel is mede tot stand gekomen met behulp van financiering door Hogeschool INHOLLAND en de Stichting LTP te Amsterdam. Met dank aan Jeroen Scholten-Linde voor zijn bijdrage aan de eerste versie van dit artikel.

Referenties

Beatty, J.R., Haas, R.W. & Sciglimpaglia, D. (1996). Using peer evaluations to assess individual performance in group class projects. *Journal of Marketing Education*, *8*, 17-28.

Boud, D. (Ed.) (1988) *Developing student autonomy in learning* (2nd ed.). London: Kogan Page.

Dochy, F., Admiraal, W. & Pilot, A. (2003). Peer- en co-assessment als instrument voor diepgaand leren: bevindingen en richtlijnen. *Tijdschrift voor Hoger Onderwijs*, *21*, *4*, 220-229.

Falchikov, N. (1986). Product comparisons and process benefits of collaborative self and peer group assessments. *Assessment and Evaluation in Higher Education,11*, 146-166.

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

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *5*, 378-382.

Gueldenzoph, L.E. & May, G.L. (2002). Collaborative Peer Evaluation: Best Practices for Group Member Assessments. *Business Communication Quarterly*, *9*, 9-20.

Guion (1998). *Assessment, Measurement, and Prediction for Personnel Decisions*. NJ: Lawrence Erlbaum.

Mcllveen, H., Greenan, K. & Humphreys, P. (1997). Involving students in teaching and learning: a necessary evil? *Quaterly Assurance in Education*, *4*, 231-238.

Kappe F.R. (2005). *Assessment centers binnen de opleiding HRM*. Evaluatie-onderzoek, INHOLLAND.

Kappe, F.R. & Scholten-Linde, J.J. (2005). Het assessment center beoordeeld: face- en consequentiële validiteit. *Tijdschrift voor Hoger Onderwijs*, *3*, 170-181.

Kolk, N.J. (2001). *Understanding and improving construct-related validity*. Dissertatie. Vrije Universiteit, Amsterdam.

Landis, J.R. & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174.

Lievens, F. (2001). Assessor training strategies and their effect on accuracy, inter-rater reliability, and discriminant validity. *Journal of Applied Psychology*, *86*, 255-264.

Marcoulides, G.A. (1989). The application of generalizability analysis to observational studies. *Quality and Quantity*, *23* 115-127.

McDowell, L. (1995). The impact of innovative assessment on student learning. *Innovations in Education and Training International*, *34*, 3022-313.

Ones, D.S. & Viswesvaran, C. (1996). Bandwidth-fidelity dilemma in personality measurement for personnel selection. *Journal of Organizational Behavior*, *17*, 609-626.

Orsmond, P. (1996). The importance of marking criteria in the use of peer-assessment. *Assessment & Evaluation in Higher Education*, *231*, 239-250.

Pond, K. (1995). Peer Review: A Percursor to Peer-assessment. *Innovation in Education and Training International*, *32* (4), 314.

Reichter, M. & Rotteveel, P. (2004). *Competentiegericht toetsen: Best Practises*. Digitale Universiteit.

Sherrard, W.R. & Raafat, F. (1994). An empirical study of peer bias in evaluations: Students rating students. *Journal of Education for Business*, *71*, 403-48.

Sluijsmans, D.M.A. (2002). *Student involvement in assessment: The training of peer-assessment skills*. Dissertatie. Open Universiteit.

Strachan, I.B. & Wilcox, S. (1996). Peer and self assessment of group work: Developing an effective response to increased enrollment in a third year course in microclimatology. *Journal of Geography in Higher Education*, *20*, 343-353.

Volder, M. de & Kappe, F.R. (2007). *Espace: a new web tool for peer-assessment with in-built feedback quality system*. Paper gepresenteerd op de 'College and Teaching Conference' in Honolulu, Hawaï, 2-5 januari.